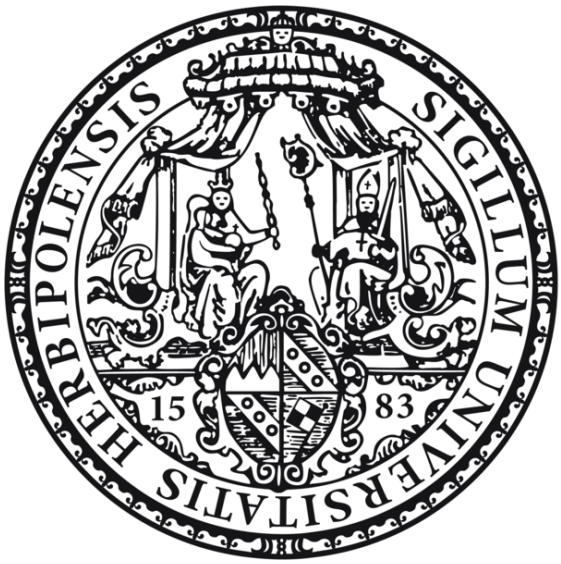


Identification and structural analysis of the *Schizosaccharomyces pombe* SMN complex

Identifizierung und Strukturanalyse des *Schizosaccharomyces pombe* SMN-Komplex

Doctoral thesis for a doctoral degree
at the Graduate School of Life Sciences,
Julius-Maximilians-Universität Würzburg,
Section: Biomedicine

submitted by
Jyotishman Veepaschit

from
Tahansir, India

Würzburg, 2020

Submitted on:

Members of the Thesis committee:

Chairperson:

Primary Supervisor:

Professor Dr. Utz Fischer
Department of Biochemistry
Biocenter, University of Würzburg
Am Hubland
97074 Würzburg
e-mail: utz.fischer@biozentrum.uni-wuerzburg.de

Supervisor (Second):

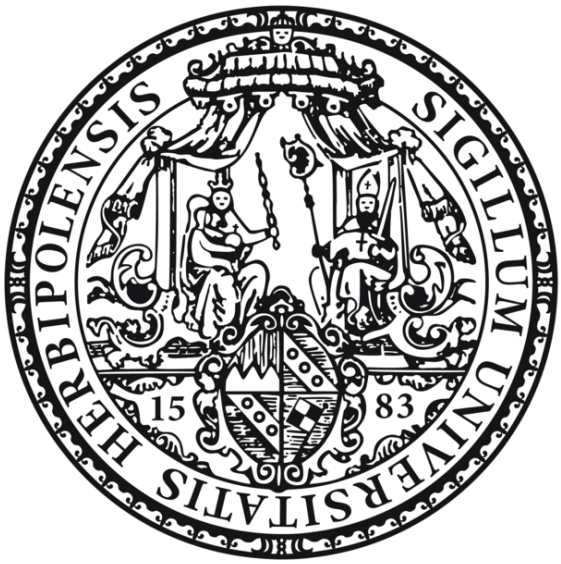
Dr. Clemens Grimm
Department of Biochemistry
Biocenter, University of Würzburg
Am Hubland
97074 Würzburg
e-mail: clemens.grimm@biozentrum.uni-wuerzburg.de

Supervisor (Third):

Professor Dr. Thomas Müller
Professor for Molecular Plant-Genetics
Department for Molecular Plant-Physiology and Biophysics - Botany I
Biocenter, University of Würzburg
Am Hubland
97082 Würzburg
e-mail: mueller@botanik.uni-wuerzburg.de

Date of Public Defence:

Date of Receipt of Certificates:



Identification and structural analysis of the *Schizosaccharomyces pombe* SMN complex

Identifizierung und Strukturanalyse des *Schizosaccharomyces pombe* SMN-Komplex

Dissertation zur Erlangung des naturwissenschaftlichen Doktorgrades
der Graduate School of Life Sciences,
Julius-Maximilians-Universität Würzburg,
Klasse Biomedizin

Vorgelegt von
Jyotishman Veepaschit

aus
Tahansir, Indien

Würzburg, 2020

Eingereicht am:

Mitglieder des Promotionskomitees:

Vorsitzende/r:

1. Betreuer:
Professor Dr. Utz Fishcer
Lehrstuhl für Biochemie
Biozentrum, Universität Würzburg
Am Hubland
97074, Wuerzburg
e-mail: utz.fischer@biozentrum.uni-wuerzburg.de

2. Betreuer:
Dr. Clemens Grimm
Lehrstuhl für Biochemie
Biozentrum, Universität Würzburg
Am Hubland
97074, Wuerzburg
mail: clemens.grimm@biozentrum.uni-wuerzburg.de

3. Betreuer:
Professor Dr. Thomas Müller
Professur für Molekulare Pflanzengenetik
Lehrstuhl für Molekulare Pflanzenphysiologie und Biophysik - Botanik I
Biozentrum, Universität Würzburg
Am Hubland
97082 Würzburg
e-mail: mueller@botanik.uni-wuerzburg.de

Tag des Promotionskolloquiums:

Doktorurkunden ausgehändigt am:

Affidavit

I hereby confirm that my thesis entitled '**Identification and structural analysis of the *Schizosaccharomyces pombe* SMN complex**' is the result of my own work. I did not receive any help or support from commercial consultants. All sources and/or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Place, Date:
Signature:

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, die Dissertation '**Identifizierung und Strukturanalyse des *Schizosaccharomyces pombe* SMN-Komplex**' eigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Ort, Datum:
Unterschrift:

Dedicated to my parents Binaya and Jyotirmayee

Summary

The biogenesis of spliceosomal UsnRNPs is a highly elaborate cellular process that occurs both in the nucleus and the cytoplasm. A major part of the process is the assembly of the Sm-core particle, which consists of a ring shaped heptameric unit of seven Sm proteins (SmD1•D2•F•E•G•D3•B) wrapped around a single stranded RNA motif (termed Sm-site) of spliceosomal UsnRNAs. This process occurs mainly in the cytoplasm by the sequential action of two biogenesis factors united in PRMT5- and SMN-complexes, respectively. The PRMT5-complex composed of the three proteins PRMT5, WD45 and pICln is responsible for the symmetric dimethylation of designated arginine residues in the C-terminal tails of some Sm proteins. The action of the PRMT5-complex results in the formation of assembly incompetent Sm-protein intermediates sequestered by the assembly chaperone pICln (SmD1•D2•F•E•G•pICln and pICln•D3•B). Due to the action of pICln, the Sm proteins in these complexes fail to interact with UsnRNAs to form the mature Sm-core. This kinetic trap is relieved by the action of the SMN-complex, which removes the pICln subunit and facilitates the binding of the Sm-core intermediates to the UsnRNA, thus forming the mature Sm-core particle. The human SMN complex consists of 9 subunits termed SMN, Gemin2-8 and Unrip. So far, there are no available atomic structures of the whole SMN-complex, but structures of isolated domains and subunits of the complex have been reported by several laboratories in the past years. The lack of structural information about the entire SMN complex most likely lies in the biophysical properties of the SMN complex, which possesses an oligomeric SMN core, and many unstructured and flexible regions. These were the biggest roadblocks for its structural elucidation using traditional methods such as X-ray crystallography, NMR or CryoEM. To circumvent these obstacles and to obtain structural insight into the SMN-complex, the *Schizosaccharomyces pombe* SMN complex was used as a model system in this work. In a collaboration with the laboratory of Dr. Remy Bordonne (IGMM, CNRS, France), we could show that the SpSMN complex is minimalistic in its composition, consisting only of SpSMN, SpGemin2, SpGemin8, SpGemin7 and SpGemin6. Using biochemical experiments, an interaction map of the SpSMN complex was established which was found to be highly similar to the reported map of the human SMN complex. The results of this study clearly show that SpSMN is the oligomeric core of the complex and provides the binding sites for the rest of the subunits. Through biochemical and X-ray scattering experiments, the properties of the SpSMN subunit such as oligomerization

and intrinsic disorder, were shown to determine the overall biophysical characteristics of the whole complex. The structural basis of SpSMN oligomerization is presented in atomic detail which establishes a dimeric SpSMN as the fundamental unit of higher order SpSMN oligomers. In addition to oligomerization, the YG-box domain of SpSMN serves as the binding site for SpGemin8. The unstructured region of SpSMN imparts an unusual large hydrodynamic size, intrinsic disorder, and flexibility to the whole complex. Interestingly, these biophysical properties are partially mitigated by the presence of SpGemin8•SpGemin7•SpGemin6 subunits. These results classify the SpSMN complex as a multidomain entity connected with flexible linkers and characterize the SpSMN subunit to be the central oligomeric structural organizer of the whole complex.

Zusammenfassung

Die Biogenese von spliceosomalen UsnRNPs ist ein hochkomplexer zellulärer Prozess, der sowohl im Zellkern als auch im Zytoplasma stattfindet. Ein Hauptteil dieses Prozesses ist der Aufbau des Sm-Kernpartikels, der aus einem ringförmigen Heptamer aus sieben Sm-Proteinen (SmD1 · D2 · F · E · G · D3 · B) besteht, die um ein einzelsträngiges RNA-Motiv (das auch als Sm-Stelle bezeichnet wird) der spliceosomalen U snRNAs gewickelt ist. Dieser Prozess findet hauptsächlich im Zytoplasma durch die sequenzielle Wirkung von zwei Biogenesefaktoren statt, den PRMT5 und den SMN-Komplexen. Der PRMT5-Komplex besteht aus den drei Proteinen PRMT5, WD45 und pICln und ist für die symmetrische Dimethylierung bestimmter Argininreste in den C-terminalen Schwänzen einiger Sm-Proteine verantwortlich. Die Wirkung des PRMT5-Komplexes führt zur Bildung von inkompetenten Sm-Protein-Intermediaten, die durch das Assemblierungs-Chaperon pICln (SmD1 · D2 · F · E · G · pICln und pICln · D3 · B) sequestriert werden. Aufgrund der Wirkung von pICln interagieren die Sm-Proteine in diesen Komplexen nicht mit den U snRNAs, um den reifen Sm-Kern zu bilden. Diese kinetische Falle wird durch die Wirkung des SMN-Komplexes aufgelöst, der die pICln-Untereinheit entfernt und die Bindung der Sm-Core-Zwischenprodukte an die U snRNA erleichtert, wodurch der reife Sm-Core-Partikel gebildet wird. Der menschliche SMN-Komplex besteht aus 9 Untereinheiten, die als SMN, Gemin2-8 und Unrip bezeichnet werden. Bisher sind keine atomaren Strukturen des gesamten SMN-Komplexes verfügbar, aber Strukturen isolierter Domänen und Untereinheiten des Komplexes wurden in den letzten Jahren von mehreren Laboratorien beschrieben. Der Mangel an strukturellen Informationen über den gesamten SMN-Komplex liegt höchstwahrscheinlich in den biophysikalischen Eigenschaften des SMN-Komplexes, der einen oligomeren SMN-Kern und viele unstrukturierte und flexible Regionen besitzt. Dies waren die größten Hindernisse für die Strukturaufklärung mit traditionellen Methoden wie Röntgenkristallographie, NMR oder CryoEM. Um diese Hindernisse zu umgehen und strukturelle Einblicke in den SMN-Komplex zu erhalten, wurde in dieser Arbeit der SMN-Komplex von Schizosaccharomyces pombe als Modellsystem verwendet.

In Zusammenarbeit mit dem Labor von Dr. Remy Bordonne (IGMM, CNRS, Frankreich) konnten wir zeigen, dass der SpSMN-Komplex in seiner Zusammensetzung minimalistisch ist und nur aus SpSMN, SpGemin2, SpGemin8, SpGemin7 und SpGemin6 besteht. Mit biochemischer Experimenten wurde eine

Interaktionskarte des SpSMN-Komplexes erstellt, die der bekannten Karte des menschlichen SMN-Komplexes sehr ähnlich war. Die Ergebnisse dieser Studie zeigen deutlich, dass SpSMN der oligomere Kern des Komplexes ist und die Bindungsstellen für den Rest der Untereinheiten bereitstellt. Durch biochemische und Röntgenstreuungsexperimente wurde gezeigt, dass die Eigenschaften der SpSMN-Untereinheit wie Oligomerisierung und intrinsische Störung die gesamten biophysikalischen Eigenschaften des gesamten Komplexes bestimmen. Die strukturelle Basis der SpSMN-Oligomerisierung wird atomar detailliert dargestellt, wodurch ein dimeres SpSMN als zentrale Grundeinheit der SpSMN-Oligomere höherer Ordnung festgelegt wird. Zusätzlich zur Oligomerisierung dient die YG-Box-Domäne von SpSMN als Bindungsstelle für SpGemin8. Die unstrukturierte Region von SpSMN verleiht dem gesamten Komplex eine ungewöhnlich große hydrodynamische Größe, intrinsische Unordnung und Flexibilität. Interessanterweise werden diese biophysikalischen Eigenschaften teilweise durch das Vorhandensein von SpGemin8 • SpGemin7 • SpGemin6-Untereinheiten gemindert. Diese Ergebnisse klassifizieren den SpSMN-Komplex als eine mit flexiblen Wechselwirkungen verbundene Multidomäneneinheit und charakterisieren die SpSMN-Untereinheit als den zentralen oligomeren Strukturorganisator des gesamten Komplexes.

TABLE OF CONTENTS

I. List of Figures

II. List of Tables

1. INTRODUCTION

The genetic information of the cell is stored in the form of Deoxyribonucleic acid (DNA) in the nucleus. During gene expression, this information must first be converted into Ribonucleic acid (RNA) which is then translated into proteins. The simplistic view of information transfer from continuous genetic codes to proteins was reformed with the discovery of split genes (Berget et al. 1977a-b; Broker et al. 1977; Chow et al. 1977a-b). It was found that most eukaryotic protein coding genes are in fact discontinuous in nature where the coding regions (termed exons) are interrupted or "split" by non-coding regions (termed introns). During gene expression, all protein coding genes are first transcribed by the RNA polymerase II through a complex transcription process within the nucleus. This results in the production of a primary messenger RNA transcripts (pre-mRNAs) containing both exons and introns. Before nuclear export and eventual translation into polypeptides by the translation machinery in the cytoplasm, pre-mRNAs must first be converted into mature mRNAs. This occurs through a series of post transcriptional modifications (5' end capping, splicing and polyadenylation) that ensure stability, error-free flow of genetic information and correct cytoplasmic fate (Moore and Proudfoot, 2009). Early during transcription, a 7-methylguanosine base is enzymatically added to the 5' end of nascent pre-mRNA transcripts, thus forming the m^7G cap. At the end of transcription process, the 3' end of pre-mRNAs is processed by the polyadenylation complex which cleaves the 3'-most part of the transcript downstream of the polyadenylation signal and catalyzes addition of several hundred Adenosine bases termed poly-A tail. Perhaps most importantly, through a process called splicing, the introns are excised from pre-mRNA transcripts and the exons ligated to produce the mature mRNA, thus generating an open reading frame for the translation into a given protein. This process is facilitated by the action of the splicing machinery termed spliceosome, which recognizes specific *cis*-acting elements at the exon-intron boundaries as well as within the intronic and exonic regions (Wahl et al. 2009).

1.1 *cis*-acting splicing signals on eukaryotic pre-mRNAs

Each intron of a pre-mRNA is marked by characteristic sequence elements, that define the borders to the adjacent exons (Figure 1.1). For 99.5% of all pre-mRNA transcripts (U2 type), the typical consensus motif at the 5' exon-intron boundary is a di-nucleotide GU- termed 5' splice site (ss), whereas at the 3' intron-exon boundary is defined by an

almost invariant AG dinucleotide (3' splice site). 20-50 nucleotide upstream of the 3' ss, an adenosine base forms the branch point (BP), which is followed by a stretch of pyrimidine bases (Py-tract, Figure 1.3 A) (Feltz et al. 2012). The remaining 0.5% intron containing pre-mRNA transcripts (U12 type) differ slightly in their consensus *cis*-acting recognition elements (Burge et al. 1999). In addition to these, other *cis*-acting elements that also play crucial roles in splicing are the Exonic Splicing Enhancers (ESE) and the Exonic Splicing Silencers (ESS) (Feltz et al. 2012; Will et al. 2011). Of note, the vast majority of human protein coding genes undergo alternative splicing where some exons are excluded to generate a variety of protein isoforms from the same gene, which forms the basis for proteome diversity within the cell.

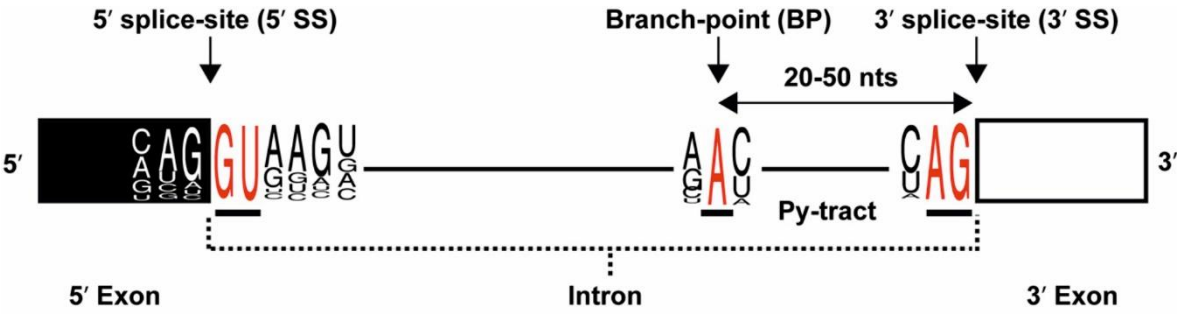


Figure 1.1: Structure of an U2 type eukaryotic pre-mRNA

The U2 type eukaryotic pre-mRNAs contain mainly four *cis*-acting splicing signals: the 5' splice site (ss) with a typical GU di-nucleotide, the 3' ss with an almost invariant AG di-nucleotide, the branch point (BP) adenosine 20-50 nucleotides upstream of the 3' ss, and a poly-pyrimidine stretch between the BP and 3' ss. (Image taken from Feltz et al. 2012)

1.2 Structure of the spliceosome

The spliceosome is a multi-megadalton (4-5 MDa) nuclear ensemble of compositionally and conformationally dynamic RNA-protein complexes (ribonucleoproteins or RNPs) called **u**ridine rich **s**mall **n**uclear **r**ibo**n**ucleo**p**roteins (UsnRNPs) and additional protein factors (reviewed by Feltz et al. 2012; Wahl et al. 2009; Will et al. 2011). Depending on the type of *cis*-acting splicing signals on introns, splicing is catalyzed by one of two spliceosomes present in higher eukaryotes: the major and the minor spliceosome. The major spliceosome consists of five UsnRNPs termed U1, U2, U5, U4 and U6 (Figure 1.2, UsnRNP), and is responsible for the splicing of the vast majority of pre-mRNA transcripts (99.5%) which contain the U2-type introns. The U4- U6- and U5-snRNPs combine to form the U4/U6•U5 tri-snRNP. The minor spliceosome on the other hand, consists of U11, U12, U5, U4atac, and U6atac snRNPs, and is responsible for the splicing of the remaining 0.5% rare introns

of the U12-type. The spliceosomal UsnRNPs as a major building blocks of the spliceosome have been biochemically and functionally characterized in detail. Each individual UsnRNP consists of a name giving **u**ridine rich **s**mall **n**uclear **r**ibonucleic **a**cid (UsnRNA) unit and various proteins (Figure 1.2, UsnRNA). The protein composition between UsnRNPs varies greatly among the different particles apart from a common heptameric core domain termed the Sm-core (Figure 1.2, proteins, Sm-core). The Sm-core of all UsnRNPs (except that of U6 and U6atac which contain a variant so-called Lsm-cores), is composed of seven Sm proteins namely B/B', D1, D2, D3, F, E and G, which assemble into a toroidal heptameric ring around a single stranded region on UsnRNA termed Sm-site (Figure 1.2, Sm-site), which conforms to a $AU_{4\text{-}6}G$ consensus sequence. The Lsm-cores of U6 and U6atac snRNPs are also toroidal ring like structures that assemble around a single stranded Lsm-site of U6/U6atac snRNA but are composed of seven Lsm (like Sm) proteins (Lsm2-8) instead of Sm proteins. In the following paragraphs, the review of literature will mostly focus on the major spliceosomal components and the Sm-core.

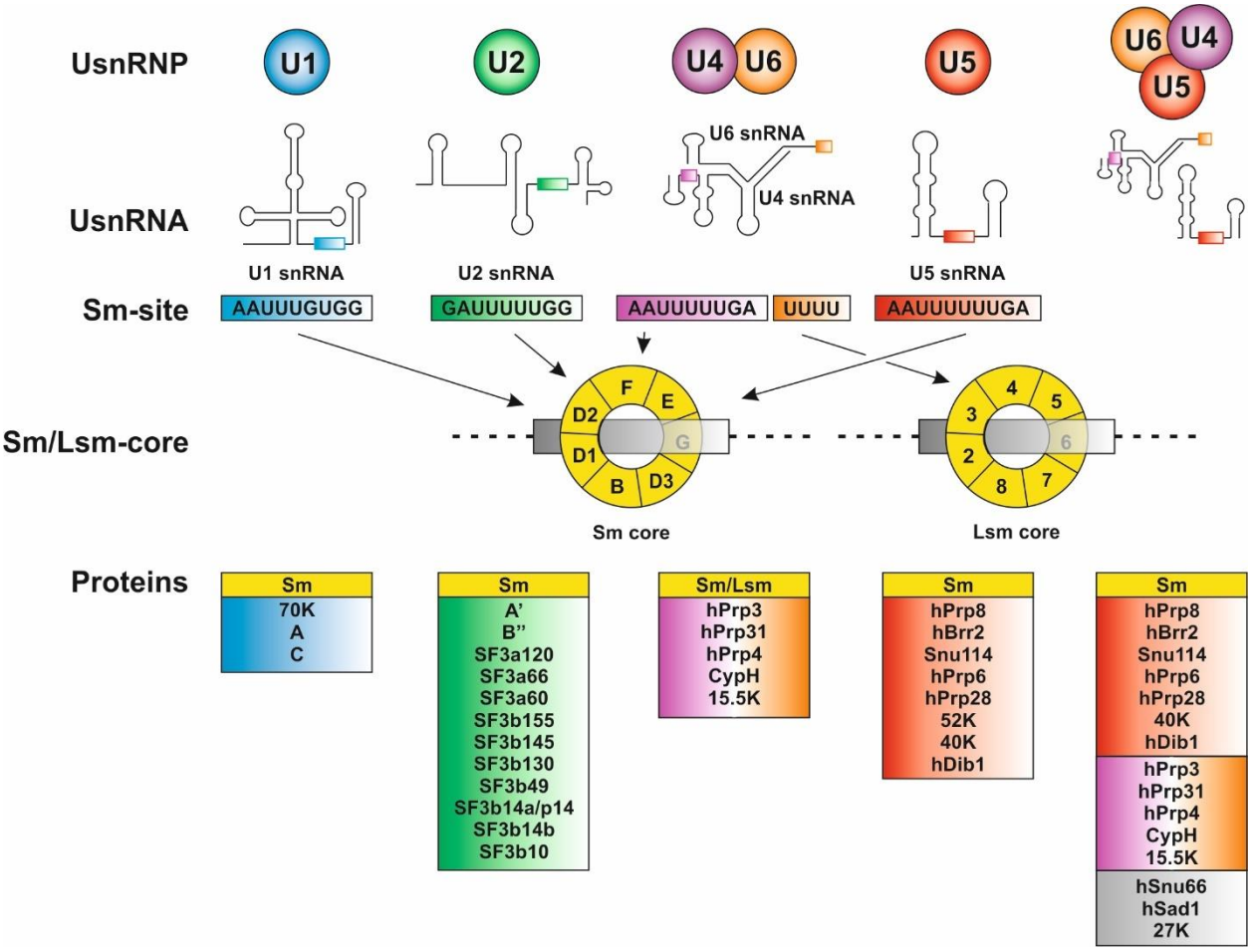


Figure 1.2: Composition of the major spliceosomal UsnRNPs
The major spliceosome is a compositionally dynamic ensemble whose major constituents are the U1, U2, U4/U6, and U5 snRNPs. Of note, U4/U6 and U5 snRNP join to form a functional unit termed tri-snRNP. Each UsnRNP consists of a name giving RNA unit, which contains a consensus uridine rich single stranded sequence motif termed Sm-site. Only U6 snRNA is different and instead contains an Lsm binding site. Common to all UsnRNPs are seven Sm proteins (SmB/B', SmD1, SmD2, SmD3, SmE, SmF and SmG) that form a toroidal ring-shaped heptameric complex termed Sm-core around the Sm-site of the RNA. Additionally, each UsnRNP contains specific proteins indicated in the tables below. Image adapted from Will and Lührmann 2011 (with permission from Cold Spring Harbor Laboratory Press)

1.3 Mechanism of pre-mRNA splicing

Fundamentally, splicing is a simple two step transesterification reaction. First, the 2' OH of the ribose of the branch point adenosine engages in a nucleophilic attack on the phosphodiester bond of the 5' splice site. This results in a lariat like structure and a free 3' OH of the upstream exon that can act in the second step as a nucleophile. This group then engages in a nucleophilic attack on the phosphodiester bond at the 3' splice site. The final products of this process are an excised intron lariat and ligated (or "spliced") exons with a continuous open reading frame (Figure 1.3 A) (Will et al. 2011). To accomplish the splicing reaction, an exceptional RNA remodeling through a dynamic network of RNA-protein and RNA-RNA interactions within the spliceosomal complexes must occur (reviewed in detail by Will et al. 2011; Yan et al. 2019; Feltz et al. 2012). During the initial phase of the splicing process, individual UsnRNPs are assembled co-transcriptionally and in a stepwise manner onto the pre-mRNA. This occurs through base-pairings between UsnRNAs and the *cis*-acting splicing elements as well as protein-RNA interactions. In the subsequent phases of the splicing reaction extensive spatial rearrangements of the snRNAs, caused by the action of diverse ATP-dependent RNA helicases form the catalytic core of the spliceosome in which the reactive groups (the 5' ss, the BP adenosine and the 3' ss) are in close proximity to facilitate splicing. A step-by-step consideration of these events as determined by genetic, biochemical, and structural studies is depicted in Figure 1.3 B. In the early spliceosomal complex (E-complex), U1snRNP defines the 5' ss through base pairing between the U1snRNA and the 5' ss of the pre-mRNA. Within this complex, additional contacts between the splicing factor SF1 and the BP region and U2AF the Py-tract are established. This allows the subsequent formation of the A-complex where SF1 is displaced by U2snRNP, which recognizes the sequence around the BP adenosine and establishes interactions between U1snRNP as well as U2AF. Base pairing of U2snRNA to the BP sequence causes the BP adenosine to bulge out, which is crucial

for the first step of splicing. The transition from A-complex to a subsequent B-complex (or pre-spliceosomal complex) is characterized by the recruitment of U4/U6•U5 tri-snRNP where an early B-complex is thought to retain the U1snRNP. After significant structural rearrangements, base pairing between U4/U6 snRNAs is disrupted which facilitates release of U4snRNP, recognition of 5’ ss by U6snRNA and displacement of U1snRNP. This is also accompanied by novel base pairing between U2 and U6 snRNAs. These extensive rearrangements result in the B*-complex primed for catalysis of the first step. After the first transesterification reaction, the resulting C-complex undergoes further structural rearrangements and the second transesterification reaction occurs. The resulting post-spliceosomal complex contains the intron lariat and the spliced exons (mature mRNA).

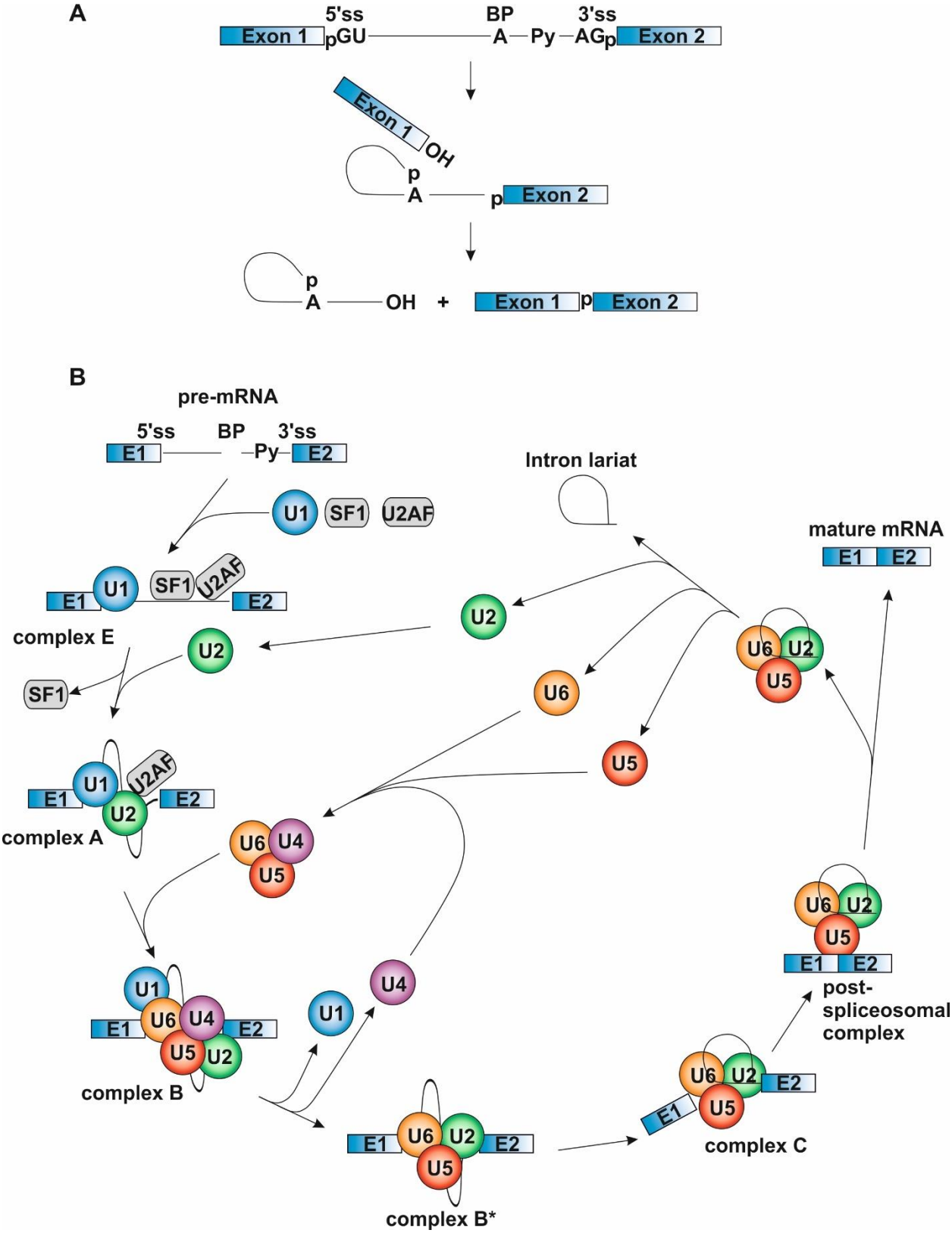


Figure 1.3: Pre-mRNA splicing by eukaryotic major spliceosome

A. In a eukaryotic pre-mRNA, the U2 type introns contain specialized *cis*-acting splicing signals within their sequences. A consensus GU dinucleotide forms the 5' splice site (ss) at the 5' exon-intron boundary, whereas an AG dinucleotide forms the 3' ss at the 3' intron-exon boundary. 20-50 nucleotide upstream of 3' ss an adenosine nucleotide forms the branch point and is usually followed by a poly-pyrimidine tract (Py-tract). Fundamentally, pre-mRNA splicing consists of two sequential transesterification reactions. First the 2' OH of the branch point adenosine engages in a nucleophilic with the phosphodiester bond at the 5' ss resulting in a lariat structure and free 5' exon. Then, the 3' OH

of the 5’ exon engages in a nucleophilic attack with the 3’ ss phosphodiester bond which results in the final products namely the fused exons and an intron lariat. B. Nuclear pre-mRNA splicing begins by the recognition of 5’ ss, branch point (BP) adenosine and the Py-tract by U1snRNP, SF1 and U2AF respectively, which gives rise to the early spliceosomal complex E. Recruitment of U2snRNP to BP results in the dissociation of SF1 and establishment of new contacts between U1, U2 snRNP and U2AF resulting in pre-spliceosomal complex A. This is followed by the recruitment of U4/U6-U5 tri-snRNP resulting in the spliceosomal B complex. Recognition of 5’ ss by U6 and establishment of novel interactions between U6 and U2 results in the dissociation of U1 and U4, respectively, resulting in a catalytically primed B* complex. This complex facilitates the first transesterification reaction, resulting in the C complex which in turn facilitates the second transesterification reaction resulting in the post-spliceosomal complex. This is followed by the release of intron lariat and the fused exons, and recycling of the UsnRNPs for further cycles of splicing reactions. Image adapted from Will and Lührmann 2011 (with permission from Cold Spring Harbor Laboratory Press)

Further structural rearrangements are required for the release of spliced exons, intron lariat and U2, U6 and U5 snRNPs for recycling into subsequent splicing cycles. During the entire process of splicing, i.e. the transition from A-complex to B-complex, from B-complex to B*-complex, from C-complex to post-spliceosomal complex and subsequent release of splicing products and RNP rearrangements for UsnRNP recycling, involve the actions of at least eight DExD/H-box RNA helicase. These enzymes use ATP hydrolysis for the modulation of RNA-protein, RNA-RNA interactions and RNA structural remodeling and thus drive the splicing cycle (reviewed by Cordin et al. 2013).

1.4 Overview of spliceosomal Sm-core biogenesis

As the vast majority of mRNAs need to be spliced in higher eukaryotes, it is not surprising that U snRNPs are very abundant entities in each and every cell nucleus. Indeed, it has been estimated that U snRNP concentration approach 10 µM in human cells which is equivalent to 10^6 particles (Montzka and Steitz 1988). The cell thus has to employ mechanisms and measures to ensure the efficient and proper formation of spliceosomal particles. Initial studies using isolated components (i.e. U snRNA and snRNP proteins) had indicated that the assembly of U snRNPs can form spontaneously in vitro. This has been most convincingly shown for the Sm core domain, which assembles from heterooligomeric complexes composed of SmD1D2, SmD3B and SmEFG in a two-step manner. First, SmD1/D2 and SmEFG bind to the Sm site to form the Sm subcore domain, which is then completed by the addition of SmD3B. As this reaction occurs at 4°C and in the complete absence of additional factors and metabolic energy, this procedure follows a typical self-assembly pathway where the information for Sm core formation resides within the individual components (Figure 1.4 A). Experiments in Xenopus laevis oocytes and egg extracts from the same organism,

however, revealed that assembly of U snRNPs in vivo is an ATP dependent and factor mediated process (Fischer et al., 1997, Meister et al. 2001, Chari et al. 2008, Pellizzoni et al. 2002). Biogenesis of snRNPs starts in the nucleus with the transcription of the UsnRNAs U1, U2, U5, U4, U11, U12, U4atac by polymerase II (Pol II). snRNAs are generated as precursors that acquire an m^7G cap co-transcriptionally and an extended 3' end (Figure 1.4 B (1)). These snRNA transcripts then traffic through nuclear Cajal Bodies (CBs) and associate with factors of the CRM1 (exportin 1 or chromosome region maintenance 1) dependent nuclear export pathway (Ohno et al. 2000). During export complex formation, UsnRNA transcripts bind to the cap binding complex (CBC) and ARS2 (arsenite resistance protein 2) through their 5' m^7G cap structure. This is followed by the recruitment of PHAX (hyper-phosphorylated adaptor of RNA export protein) and CRM1, which servs as the nuclear export receptor (2). In the cytoplasm, the export factors are dissociated by the dephosphorylation of PHAX and the UsnRNAs undergo further maturation steps such as 3' end trimming and hypermethylation of the m^7G cap into m^3G cap (3). Assembly of the snRNA with Sm proteins occurs exclusively in the cytoplasm in higher eukaryotes. Even though this process can occur spontaneously in vitro (see above), it requires the assistance of specialized assembly factors united in the PRMT5 (protein arginine methyl transferase 5) and the SMN (survival motor neuron) complexes. The PRMT5 complex consists of the name giving PRMT5, pICln, and WD45 (also called MEP50) proteins. The SMN complex is a multiprotein complex consisting of at least 9 subunits namely the SMN (Survival Motor Neuron) protein, Gemins 2-8 and Unrip (UNR interacting protein) (Figure 1.4 A).

Both assembly complexes act sequentially in the snRNP assembly pathway. The early phase is dominated by the PRMT5 complex. The pICln subunit binds newly synthesized Sm proteins arising at the exit tunnel of the ribosome and ties them to the PRMT5 complex. In a step-wise manner oligomeric complexes composed of pICln/D1/D2, pICln/D1/D2/ E/F/G and pICln/D3/B assemble at the PRMT5 complex and become symmetrically dimethylated (sDMA) through the methyltransferase activity of PRMT5. While pICln/D3/B is believed to remain attached to the PRMT5 complex, pICln/D1/D2, together with SmEFG, assembles into a toroidal ring-shaped assembly intermediate termed the 6S complex, which dissociates from the PRMT5 complex. Within the 6S complex, the Sm proteins are topologically pre-organized (pICln/D1/D2/F/E/G) for the final assembly steps later in the late assembly phase. pICln

bound Sm proteins are kinetically trapped in an assembly incompetent state and hence fail to interact with the UsnRNA (Figure 1.4 B: steps (a), (b) and (c)). In the late assembly phase, the SMN complex takes over the assembly intermediates from the PRMT5 complex (Figure 1.4 B: step (d)), relieves the pICln induced kinetic trap and facilitates the loading of the Sm protein heterooligomers onto the UsnRNA, forming the mature Sm-core particle (step (5)).

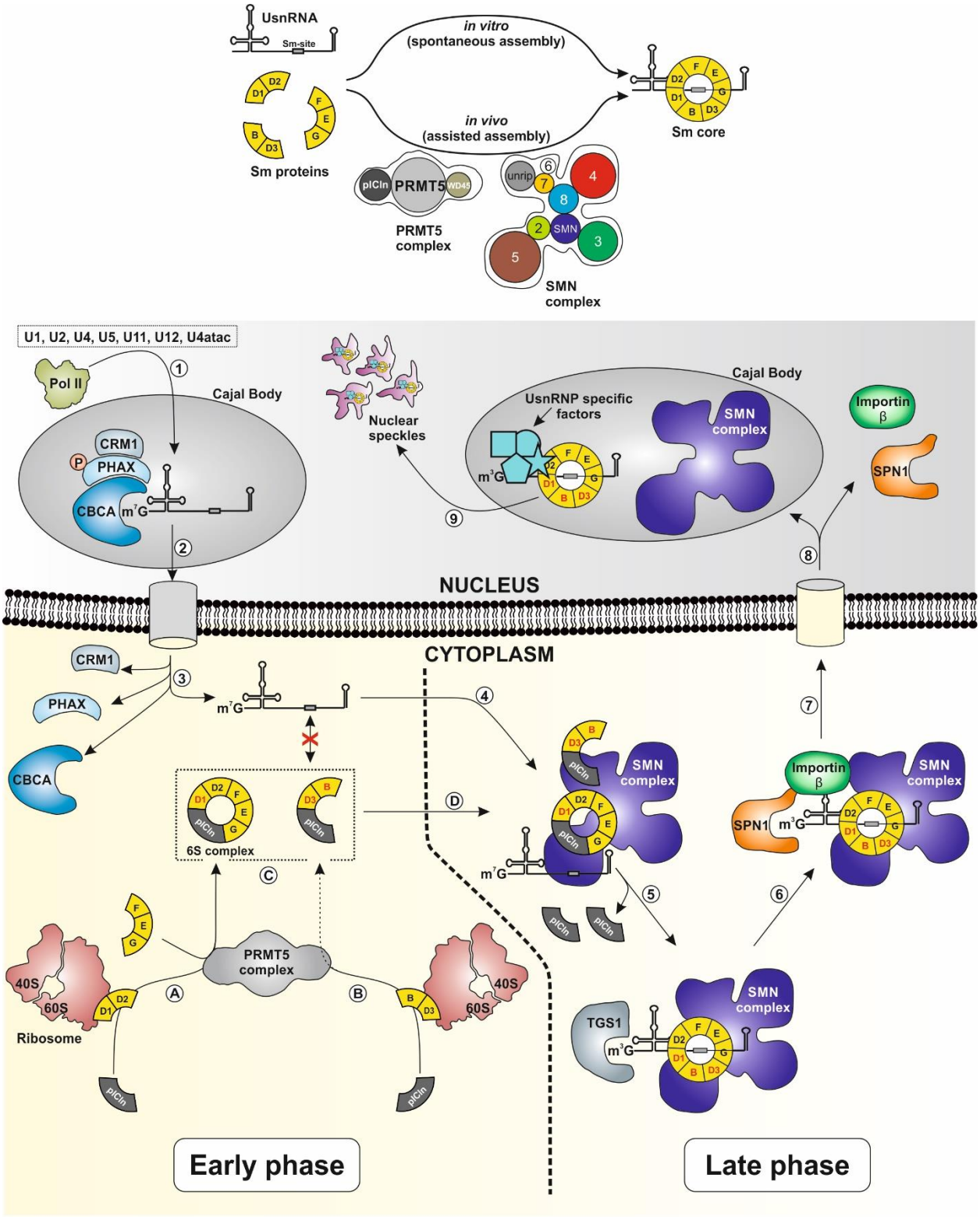


Figure 1.4: Spliceosomal Sm-core biogenesis

Upper panel: schematic of *in* vitro and *in* vivo assembly of UsnRNP Sm-core. Lower panel: cellular Sm-core biogenesis. The UsnRNA genes are transcribed by Pol II followed by 5' capping (m^7G) (1). Then, the cap binding complex CBCA recognizes and binds to the 5' cap structure. This is followed by the recruitment of PHAX and CRM1 which initiates the nuclear export (2). Once in the cytoplasm, the dephosphorylation of PHAX triggers the dissociation of the export complex from the UsnRNA (3). The cytoplasmic Sm-core assembly process is divided in two phases, early and late phase. In the early phase, the newly translated Sm proteins are picked up from the ribosome by the assembly chaperone pICln and are recruited to the PRMT5 complex (A and B). Then, the Sm proteins D1, D3 and B are symmetrically dimethylated by PRMT5 in specific arginine residues in their C-terminal tails. After

dimethylation, pICln/D1/D2 leaves the PRMT5 complex in the form of a ring-shaped assembly intermediate termed the 6S complex (C). However, it is still unclear how pICln/D3/B is handed over to the SMN complex. Both the 6S and the pICln/D3/B assembly intermediates are assembly incompetent (kinetically trapped) and do not interact with the target UsnRNA. In the late assembly phase, the SMN complex receives the pICln bound Sm intermediates as well as the UsnRNA and facilitates the formation of mature Sm-core by relieving the pICln induced kinetic trap (D, 4). The mature Sm-core serves as the recruitment signal for trimethyl guanosine synthase 1 (TGS1) which hypermethylates the m^7G cap structure to 2,2,7-trimethyl guanosine (m^3G) cap (5). The m3G cap together with the Sm-core provides the binding platform for the snRNP specific import factor snurportin 1 (SNP1) as well as importin β (6). After nuclear import, the import complexes dissociate and the Sm-core is targeted to cajal bodies where it is released from the SMN complex (7, 8). Within the cajal bodies, the Sm-core undergoes further maturation steps by the addition of UsnRNP specific proteins. Finally, mature UsnRNPs are targeted to nuclear speckles for storage before they participate in pre-mRNA splicing (9).

The SMN complex not only facilitates the assembly of the Sm core domain but also serves as an integrator of additional activities related to snRNP biogenesis. This includes the recruitment of the trimethylguanosine synthase 1 (TGS1) for hypermethylation of the m^7G cap into 2,2,7-trimethylguanosine (m_3G) cap which serves as part of a bipartite nuclear import signal for the assembled U snRNPs. The assembled Sm core domain together with the m3G-cap enables the binding of the import complex consisting of snurportin 1 (SNP1) and importin β (step (6)). After nuclear import, Sm-core dissociates from the SMN complex and undergoes further maturation in CBs where additional UsnRNP specific proteins are added (steps (7) and (8)). Finally, mature UsnRNPs accumulate in nuclear speckles, which are believed to be storage sites from which they can be recruited to nascent pre-mRNA to form the spliceosome (step (9)) (UsnRNP biogenesis has been reviewed in detail by Fischer et al. 2011, Matera et al. 2014, and Gruss et al. 2017).

1.5 Structural biology of the human SMN complex

The human SMN complex consists of SMN, Gemins 2-8 and Unrip (Kroiss et al. 2008), which together form a macromolecular machine acting in snRNP biogenesis. An inter-subunit interaction map and the secondary structure elements and domains of each subunit are shown in Figure 1.5 A (Otter et al. 2007). The Gemin2 subunit serves as the 6S recruiting module of the SMN complex. A crystal structure of the SMN_N-term•Gemin2•6S complex (Grimm et al. 2013) reveals that the N-terminal 6S-binding arm of Gemin2 makes extensive contacts to the 6S complex, and the SMN_N-terminus tethers the Gemin2 subunit to the rest of the SMN complex (Figure 1.5 C3). Apart from its Gemin2 interaction module, the SMN subunit contains 3 additional structural features: a central Tudor domain, a proline-rich (P-rich) region, and a C-terminal YG-box. The Tudor domain has been shown to bind symmetrically dimethylated arginine

(Figure 1.5 C4), which may increase the affinity of symmetrically dimethylated Sm proteins D1, D3 and B/B', for the SMN complex (Selenko et al. 2001; Tripsianes et al. 2011; Gonsalvez et al. 2008). The C-terminal YG-box domain is essential for the oligomerization of SMN and has been shown to form dimers, tetramers, and octamers *in vitro*. A structure of the MBP-tagged YG-box revealed that it forms helical glycine zipper dimers (Figure 1.5 C5) (Martin et al. 2012). There are no known structures or functions for Gemin8 and Gemin4. Although the specific functions are unknown, both Gemins 6 and 7 exhibit Sm-fold-like structures (Figure 1.5 C1) (Ma et al. 2005), which suggests that they may interact at some stages in the assembly pathway with Sm proteins. The Gemin5 subunit interacts directly with the Sm-site of UsnRNAs and may serve a crucial role in the recruitment of UsnRNAs to the SMN complex. A structure of the WD40-1 and WD40-2 domains of Gemin5 bound to the Sm-site of U4snRNA is shown in Figure 1.5 C2 (Jin et al. 2016; Xu et al. 2016; Wahl et al. 2016). Although no specific functions have been assigned to the Gemin3 subunit, the N-terminal helicase domain suggests a role in RNA rearrangements and/or RNP remodeling. So far, only a structure of the N-terminal DEAD domain of Gemin3 has been reported (Figure 1.5 C6) (Schutz et al. 2010). Gemin3 belongs to the DExD/H-box family of proteins which play crucial roles in RNA processing and function as RNA helicases, foldases and/or RNP modelers. This suggests an RNA handling activity for Gemin3 within the SMN complex (Schutz et al. 2010). Notwithstanding, the lack of structural information for the inter subunit interactions among SMN↔Gemin8↔Gemin7, Gemin8↔Gemin4↔Gemin3↔SMN and Gemin2↔Gemin5 subunits, remains a roadblock for the generation of a model of the SMN complex. Moreover, the oligomeric properties of the central SMN subunit (↔SMN↔SMN↔SMN↔) further complicates attempts to understand the overall structural features of the SMN complex.

1.6 The *Schizosaccharomyces pombe* (Sp) UsnRNP assembly machinery

As nearly as 50% of *S. pombe* genes contain introns (Kupfer et al. 2004; Zhu et al. 2013). The *S.* pombe genome is also intron-rich (0.9 introns per gene) compared to *S. cerevisiae*. Additionally, *cis*-acting intronic signals in *S. pombe* resemble more closely to those of humans compared to *S. cerevisiae* (Fair et al. 2017). These factors necessitate a crucial role of the spliceosome, and hence the SpSMN complex in yeast gene expression. Although the yeast spliceosome has been shown to possess many similarities with the human spliceosome (Fair et al. 2017), the current knowledge of the

UsnRNP biogenesis factors are limited to a very minimalistic SpSMN complex consisting of only the orthologs of SMN and Gemin2 (Hannus et al. 2000; Owen et al. 2000; Paushkin et al 2000). The domain architecture and secondary structural elements of the two proteins have been shown in Figure 1.5 B. It is currently not known whether *S. pombe* contains orthologs of the remaining Gemins. The homologous N-terminal are (nearly 80 residues long) of the SpGemin2 subunit suggests a substrate recruitment mechanism that is structurally similar to the human SMN complex (Grimm et al. 2013; Zhang et al. 2011). Interestingly, the SpSMN subunit shows structural homology only at the N- and the C-terminal extremes (the SpGemin2 binding domain and the YG-box, respectively. Figure 1.5 B). The SpSMN lacks any structural features such as the Tudor domain, poly-proline region, as in the case of the human SMN. The only reported structure from the SpSMN complex is a dimeric YG-box fused to an MBP tag (Figure 1.7 D, Gupta et al. 2015) which shows a glycine-zipper YG-box dimer identical to the human YG-box (Figure 1.7 C, Martin et al. 2012). As the presence of many homologous splicing factors has drawn considerable interest in the recent years to adopt *S. pombe* as model system to study splicing (Fair et al. 2017; Yan et al. 2015; Nguyen et al. 2016), further investigation into the structure and composition of the SpSMN complex appears plausible. In this work, additional Gemins in *S. pombe* are identified and characterized.

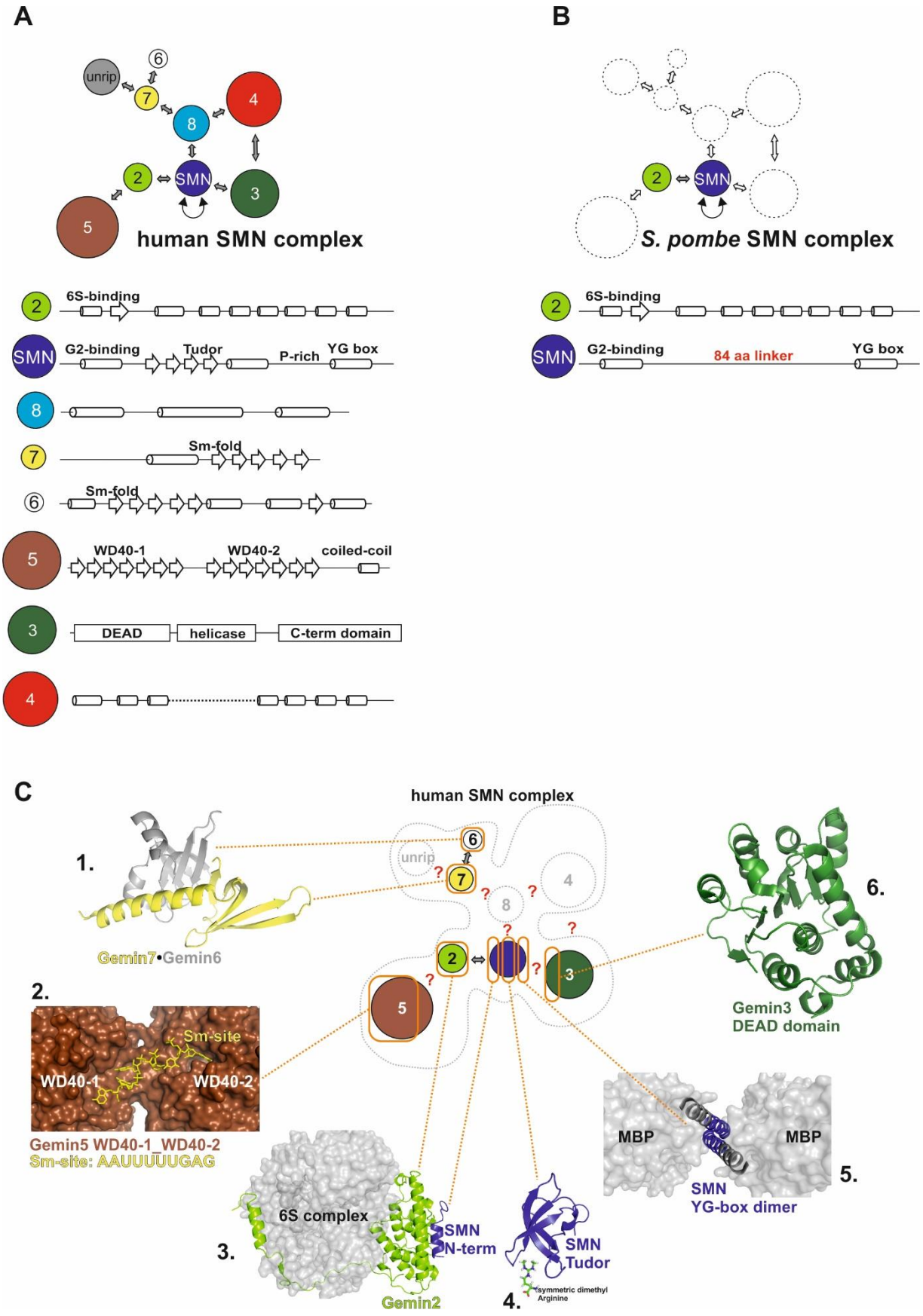


Figure 1.5: Structural biology of the SMN complex

A: Interaction map of the human SMN complex which consists of SMN, Gemins 2-8 and unrip. Each Gemin is denoted only by its number. The interaction between subunits is indicated by double headed arrows. The self-interaction/oligomeric property of SMN is denoted by a curved arrow. Below the

interaction map, the secondary structural elements and/or domain composition of each subunit is shown (except that of unrip). Helical regions are shown as cylinders, and beta strands are shown as arrows. B: *S. pombe* SMN complex interaction map, secondary structural elements and/or domain compositions of the subunits. C: Solved atomic structures from the human SMN complex: 1. Gemin7•Gemin6 dimer (Sm-folds, PDB 1Y96), 2. Gemin5 WD-40 domains with U4snRNA Sm-site (5GXI), 3. SMN_N-term•Gemin2•6S complex (PDB 4V98, here the Gemin2 subunit and the SMN_N-term is derived from *Drosophila melanogaster*), 4. Tudor domain of SMN with symmetric dimethyl arginine (PDB 4A4E, NMR structure), 5. Glycine-zipper homodimer of SMN YG-box with MBP-tag (PDB 4GLI), 6. DEAD domain of Gemin3 (PDB 3B7G). Uncharacterized atomic interactions between subunits are denoted by (?).

1.7 Spinal Muscular Atrophy (SMA) and the YG-box

SMA is a debilitating neuromuscular disorder that affects about 1 in 10,000 infants and is the primary genetic cause of infant mortality. SMA is characterized by degeneration of anterior horn cells (α-motor neurons) and muscular atrophy, which results in weakness. The disease is progressive and generally divided into types 0-4 depending on the time of onset and severity. Loss of functional SMN protein has been shown to be the primary genetic cause associated with Spinal Muscular Atrophy (SMA). There are 2 copies of the *SMN* gene in healthy individuals, the telomeric *SMN1*, and the centromeric *SMN2* (chromosome 5q13). Both copies of the *SMN* gene are almost identical except for a C>T conversion in the exon 7 of *SMN2*. While the *SMN1* gene allows expression of the full-length SMN protein, *SMN2* undergoes exon 7 skipping due to the C>T conversion which results in mostly non-functional truncated version of SMN (SMNΔ7), which fails to oligomerize. While the majority of functional SMN in the cell comes mainly from *SMN1*, approx. 15% of *SMN2* products is full-length SMN protein due to inefficient exon skipping (Figure 1.6 A) (reviewed in detail by Kolb et al. 2015 and Burghes et al. 2009).

The underlying genetic anomaly in 95% of SMA cases was found to be homozygous deletions or mutations in the *SMN1* gene. This results in non-functional SMN protein, and the residual *SMN2* product is insufficient to compensate the required SMN protein levels within the cell. Nearly 50% of all SMA causing mutations are located at the C-terminus of SMN encompassing the YG-box region (reviewed by Jędrzejowska et al. 2014). The YG-box is essential for the oligomerization of SMN, and many of these mutations have been shown to prevent oligomerization. Specifically, the mutations Y272C, G275S, G279C and G279V, have been shown to cause oligomerization defect, resulting in monomeric SMN (Figure 1.6 B) (Martin et al. 2012). The structure of the YG-box dimer revealed that these mutations are located at the glycine-zipper dimeric interface (Figure 1.6 C-D). The mutations S266P and T274I, also lead to monomeric forms of YG-box although they are situated away from the glycine-zipper dimeric

interface (Figure 1.6 C-D). All these observations suggest a crucial role of YG-box oligomerization defect in SMA pathology. Since the structure of only dimeric form of the YG-box is available (Martin et al. 2012; Gupta et al. 2015), the structural basis of higher order oligomers of YG-box seems to be fundamental to better understand function of SMN.

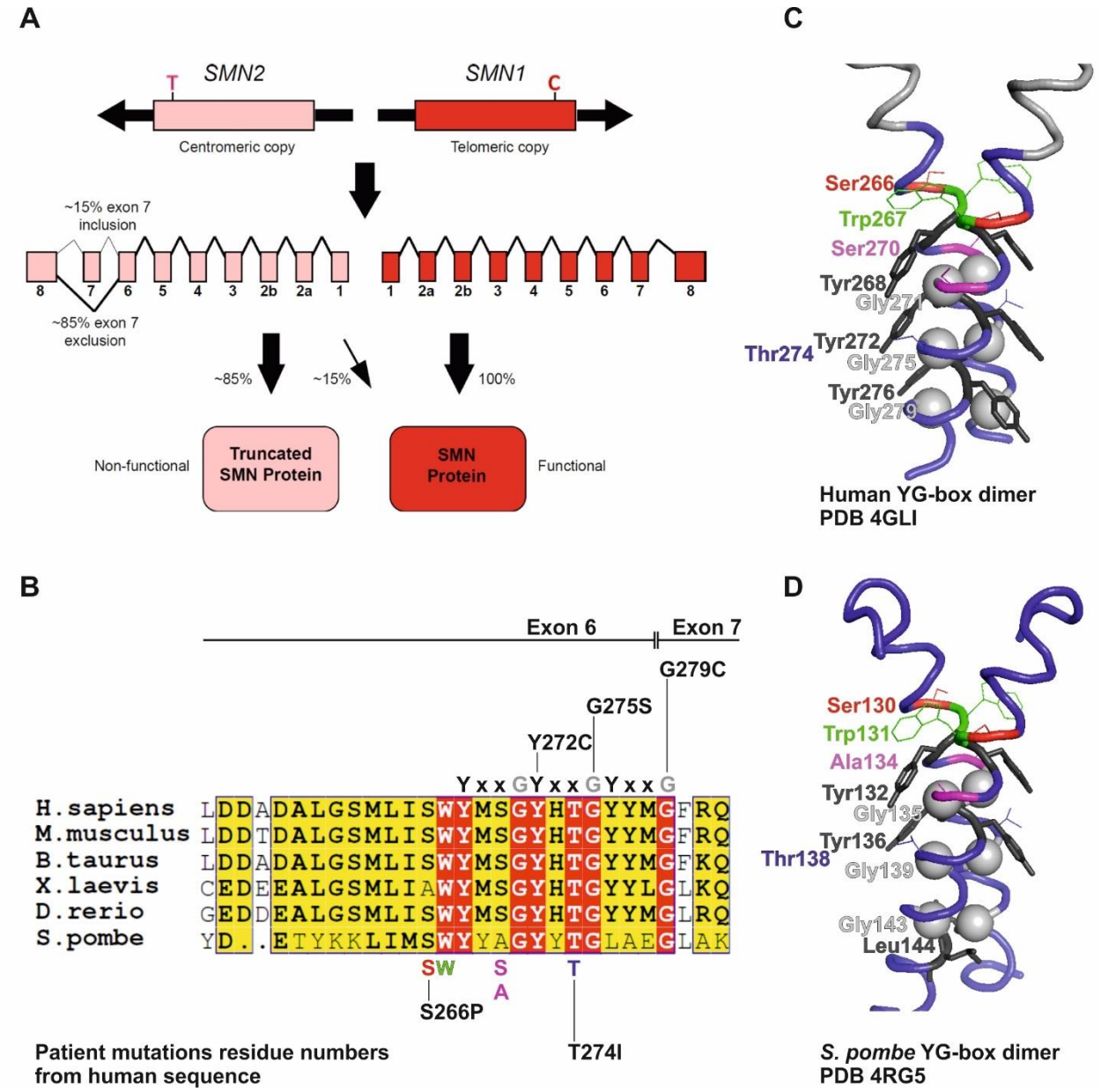


Figure 1.6: Spinal Muscular Atrophy and the YG-box

A: Healthy individuals contain 2 copies of the *SMN* gene, the telomeric *SMN1* and the centromeric *SMN2,* which differ in a single nucleotide within exon 7. While all *SMN1* gene products are full-length, functional SMN protein, *SMN2* undergoes alternative splicing resulting in 85% of its products being non-functional SMNΔ7 (Image taken from Kolb et al. 2015). B: Sequence homology of the YG-box region of SMN from different organisms. The YG-box motif is a triple repeat of YxxG. Exception: the third repeat in *S. pombe* is LxxG. SMA causing mutations within this motif are shown above the alignment, and SMA causing mutations outside this motif are shown below the alignment. C & D: Glycine-zipper homodimers of the human and the *S. pombe* YG-box, respectively. Residues directly involved in dimerization (3(YxxG) motif) are shown in black and grey. Residues not involved in dimerization are shown in colors.

2. AIM OF THE STUDY

In the past year, key structures of SMN complex components and functional subcomplexes have been solved. These structural data in combination with biochemical work allowed already detailed insight into the functional architecture of the assembly machinery. Structural insight into the architecture of the entire SMN complex, however are still limited. This is mostly due to the fact that the SMN complex displays a large degree of heterogeneity in vivo, which prevents its isolation and structural determination by either X-ray or electron cryomicroscopy (cryo-EM).

The aim of this thesis is to investigate the structure of the SMN complex using *Schizosaccharomyces pombe* as a model organism. This organism expresses a minimalistic SMN complex consisting of SpGemin2, SpSMN, SpGemin8, SpGemin7 and SpGemin6 only. It was therefore a possibility that this would be an appropriate model system to study the oligomeric and overall biophysical properties of the SMN complex.

The specific objectives of this thesis are to

- Establish a recombinant expression system for the *S. pombe* SMN complex components.
- Reconstitute the SpSMN complex *in vitro.*
- Determine the structural basis of SpSMN oligomerization by X-ray crystallography
- Investigate the oligomeric and overall biophysical properties of the SpSMN complex by small angle X-ray scattering.
- Develop a structural model of the SpSMN complex.

3 MATERIALS AND METHODS

3.1 Materials

3.1.1

Buffers and solutions	**Composition**
Coomassie staining solution	0.15 % (w/v) Serva Blue R 25 % Isopropanol 10 % Acetic acid
Destaining solution	7.5 % Ethanol 7.5 % Acetic acid
Dialysis buffer	150 mM NaCl 20 mM Hepes, pH 7.4 2 mM DTT
Final culture medium	TB medium 1X TB buffer 1X antibiotics 1 mM $MgCl_2$ 1 drop PPG 2000 per 800 ml medium
Gelfiltration buffer	150 mM NaCl 20 mM Hepes, pH 7.4 2 mM DTT
Lämmli stacking gel (5 %)	For 4 gels 1.7 ml Rotiphorese Gel 30 A/B 37.5:1 2.5 ml 1 M Tris-HCl, pH 6.8 5.7 ml ddH_2O 100 µl 10 % SDS 50 µl 10 % APS 30 µl TEMED
Lämmli resolving gel (12 %)	For 4 gels 9.6 ml Rotiphorese Gel 30 A/B 37.5:1 9.0 ml 1 M Tris-HCl, pH 8.8 5.0 ml ddH_2O 121.6 µl 10 % SDS 121.6 µl 10 % APS 121.6 µl TEMED
Luria Bertani (LB) medium	1 % (w/v) Bacto Tryptone 0.4 % (w/v) Yeast Extract 1 % (w/v) NaCl
LB agar	LB medium 1.5 % Agarose
List of 1000X Antibiotics	25 mg/ml Kanamycin 100 mg/ml Ampicillin 30 mg/ml Chloramphenicol
List of 1000X protease inhibitor stocks	PMSF (200 mM in Ethanol) AEBSF (100 mM in water) Aprotinin (1 mM in water) Leupeptin/Pepstatin A (1 mM in DMSO)

List of chromatography regeneration buffers	0.25 M EDTA 1 M NaOH 6 M Guanidine hydrochloride 0.25 M $NiCl_2$ 1 M Imidazole 20 % Ethanol
List of crystallization screens (All conditions prepared by Emilia Gärtner)	NPPT1 (Grimm et al, 2010) NPPT2 (Grimm et al, 2010) NPPT3 (Grimm et al, 2010) MIDAS (Molecular Dimensions) Wizard Classic 1 & 2 (Rigaku) Hampton Crystal Screen 1 & 2 (Hampton) JCSG+ (QIAGEN) Natrix-HT (Hampton) JJS (In-house composition) Hampton Additive Screen (Hampton)
Ni-NTA lysis buffer	150 mM NaCl 50 mM Hepes, pH 7.4 20 mM Imidazole 2 mM β-Mercaptoethanol 10 % Glycerol
Ni-NTA elution buffer	150 mM NaCl 50 mM Hepes, pH 7.4 250 mM Imidazole 2 mM β-Mercaptoethanol 10 % Glycerol
Terrific Broth (TB) medium	1.2 % (w/v) Bacto Tryptone 2.4 % (w/v) Yeast Extract 4 % (v/v) Glycerol
TFB-I	30 mM KAc, 100 mM RbCl, 10 mM $CaCl_2$, 50 mM $MnCl_2$, 15 % glycerol, pH 5.8
TFB-II	10 mM RbCl, 75 mM $CaCl_2$, 15 % Glycerol, 10 mM MOPS, pH 6.5
Tris-Tricine stacking gel (3 %)	<u>For 4 gels</u> 1 ml Rotiphorese Gel 30 A/B 29:1 2.5 ml Gel buffer 6.4 ml ddH_2O 50 µl 10 % SDS 30 µl 10 % APS 100 µl TEMED
Tris-Tricine resolving gel (15 % / 13 %)	<u>For 4 gels</u> 12/10.4 ml Rotiphorese Gel 30 A/B 29:1 8 ml Gel buffer 3.8/5.4 ml ddH_2O 240 µl 10 % SDS 100 µl 10 % APS 30 µl TEMED

6X DNA loading dye	10 mM Tris-HCl, pH 7.6, 60 mM EDTA, pH 8.0, 0.03 % (w/v) Bromophenol blue, 0.03 % (w/v) Xylene Cyanol F
3X Gel buffer	3 M Tris base, pH 8.45 0.3 % (w/v) SDS
10X Lämmli running buffer	0.25 M Tris base 1.92 M Glycine 1 % (w/v) SDS
5X SDS loading buffer	0.25 M Tris-HCl, pH 6.8 100 mM DTT 10 % (w/v) SDS 50 % (v/v) Glycerol 0.04 % (w/v) Bromophenol blue
50X TAE (Tris-acetate EDTA) buffer	2 M Tris base 1 M Acetate 50 mM Na_2EDTA
10X TB buffer	0.17 M KH_2PO_4 0.72 M K_2HPO_4
10X Tris-Tricine Anode buffer	2 M Tris base, pH 8.9
10X Tris-Tricine Cathode buffer	1 M Tris base, pH 8.25 1 M Tricine 1 % (w/v) SDS

3.1.2

Chromatography Materials and Systems	**Features**	**Vendors**
Superdex 75 26/60	320 ml, max 5 ml sample, max Flow rate 2.5 ml/min	GE Life Sciences
Superdex 75 10/300	24 ml, max 0.5 ml sample, max Flow rate 1 ml/min	GE Life Sciences
Superdex 200 26/60	320 ml, max 5 ml sample, max Flow rate 2.5 ml/min	GE Life Sciences
Superdex 200 10/300	24 ml, max 0.5 ml sample, max Flow rate 0.75 ml/min	GE Life Sciences
Superose 6 10/300	24 ml, max 0.5 ml sample, max Flow rate 0.5 ml/min	GE Life Sciences
Ni-NTA agarose	Gravity flow, binding capacity 50 mg/ml	QIAGEN
Liquid Chromatography Columns	Gravity flow, 8- and 98- ml variants	Sigma-Aldrich
ÄKTA Explorer 100	Upto 100 ml/min flow	GE Life Sciences
ÄKTA Purifier 10	Upto 10 ml/min flow	GE Life Sciences

3.1.3

Crystallization Consumables and Equipment	**Vendors**
CrystalQuick LP 96 well plate, 609171	Greiner Bio-One
CrystalQuick LP 96 well plate, 609101	Greiner Bio-One
Pregreased 24 well plate	Crystalgen
96 well Masterblock, 2 ml, 780270	Greiner Bio-One
22 mm circular cover slips-plain, CSL-104	Jena Bioscience
22 mm circular cover slips-siliconized, CSL-107	Jena Bioscience
Mosquito Crystallization Robot	TTP Labtech
Mounted Litho Loops	Molecular Dimensions

3.1.4

Kits and General Lab Consumables	**Vendors**
NucleoSpin Plasmid Mini Kit	Machery-Nagel
NucleoSpin Gel and PCR clean up Kit	Machery-Nagel
KAPA HiFi DNA Polymerase	KAPA Biosystems
Gel Filtration HMW Markers (29-700 kDa)	Sigma-Aldrich
Bio-Rad Protein Assay Dye, #5000006	Bio-Rad
Restriction Enzymes and T4 DNA ligase	Thermo Fischer Scientific
PageRuler Unstained Protein ladder (200-10 kDa)	Thermo Fischer Scientific
PageRuler Prestained Protein ladder (170-10 kDa)	Thermo Fischer Scientific
TEV enzyme in 50% Glycerol, 4.5 mg/ml	In-house preparation
15- and 50-ml falcon tubes	Greiner Bio-One
0.2, 1.5- and 2.0-ml reaction tubes	Eppendorf
1.5-ml reaction tubes, without lid	Hartenstein
Petridishes	Greiner Bio-One
Sterile Filters (0.22 and 0.45 µm)	Merc Milipore
ZelluTrans Dialysis Tubes, MWCO 3.5 kDa	Carl Roth

3.1.5

***E. coli* strain**	**Genotype**	**Resistance**	**Product Nr.**
DH5α (T1, T5 phage resistant)	*fhuA2Δ(argF-lacZ)U169 phoA glnV44ϕ80Δ(lacZ)M15gy rA96 recA1 relA1 endA1 thi-1 hsdR17*	None	NEB C2987I
BL21 (DE3) (T1, T5 phage resistant)	*fhuA2 [lon] ompT gal (λ DE3) [dcm] ΔhsdS λ DE3= λsBamHIoΔEcoRI-B int::(lacI::PlacUV5::T7 gene1) i21 Δnin5*	None	NEB C2527I

3.1.6

Plasmid Vectors	Features
pETM-11 (EMBL)	Cloning performed between *NcoI-XhoI*, N-terminal 6xHis-tag, TEV cleavage site, Kanamycin resistance
pETM-13 (EMBL)	Cloning performed between *NcoI-XhoI*, No-tags, Kanamycin resistance
pET28M-SUMO (Prof. Alexander Buchberger)	Cloning performed between *BamHI-XhoI*, N-terminal 6xHis-SUMO-tag, SENP2 cleavage site, Kanamycin resistance

3.1.7

Designed Monocistronic Inserts	Vector, Affinity tag, Sites used
NcoI --- SpG2 --- *NheI-XhoI*	pETM-11, 6xHis, *NcoI-XhoI*
NcoI --- SpG2 --- *NheI-XhoI*	pETM-13, untagged, *NcoI-XhoI*
NcoI --- SpG2$^{\Delta arm}$ --- *NheI-XhoI*	pETM-11, 6xHis, *NcoI-XhoI*
NcoI --- SpSMN --- *NheI-XhoI*	pETM-11, 6xHis, *NcoI-XhoI*
NcoI --- SpSMN --- *NheI-XhoI*	pETM-13, untagged, *NcoI-XhoI*
NcoI --- SpSMN$^{\Delta YG}$ --- *NheI-XhoI*	pETM-13, untagged, *NcoI-XhoI*
NcoI --- SpSMN$^{\Delta 36\text{-}119}$ --- *NheI-XhoI*	pETM-13, untagged, *NcoI-XhoI*
NcoI --- SpSMN$^{\Delta 36\text{-}117}$ --- *NheI-XhoI*	pETM-13, untagged, *NcoI-XhoI*
NcoI --- SpSMN$^{\Delta 36\text{-}111}$ --- *NheI-XhoI*	pETM-13, untagged, *NcoI-XhoI*
NcoI --- SpSMN$^{\Delta 36\text{-}109}$ --- *NheI-XhoI*	pETM-13, untagged, *NcoI-XhoI*
NcoI --- SpG8 --- *NheI-XhoI* (His-SpG8)	pETM-11, 6xHis, *NcoI-XhoI*
NcoI --- SpG8 --- *NheI-XhoI*	pETM-13, untagged, *NcoI-XhoI*
NcoI --- SpG8$^{\Delta loop}$ --- *NheI-XhoI*	pETM-11, 6xHis, *NcoI-XhoI*
NcoI --- SpG8$^{\Delta N58}$ --- *NheI-XhoI*	pETM-11, 6xHis, *NcoI-XhoI*
BamHI --- SpG8$^{3\text{-}34}$ --- *XhoI*	pET28M, 6xHis-SUMO, *BamHI-XhoI*
NcoI --- SpG7 --- *NheI-XhoI*	pETM-11, 6xHis, *NcoI-XhoI*
NcoI --- SpG7 --- *NheI-XhoI*	pETM-13, untagged, *NcoI-XhoI*
NcoI --- SpG6 --- *NheI-XhoI*	pETM-13, untagged, *NcoI-XhoI*

**Δarm refers to the first 80 residues of SpG2. ΔYG refers to residues 131-54 of SpSMN. Δloop refers to residues 35-58 of SpG8. ΔN58 refers to first 58 residues of SpG8.

Designing the polycistronic constructs: First the 3' ends of monocistronic constructs were cut open with NheI-XhoI digestion, which served as the site for cloning the second gene. The second gene was cut out from its monocistronic construct with XbaI-XhoI digestion. This insert contained the ribosome binding site (Shine Dalgarno sequence) from the vector backbone. Then, the insert was ligated to the NheI-XhoI site resulting in a di-cistronic construct (NheI is compatible with XbaI and ligation results in no site). For cloning a third gene, first the di-cistronic construct was similarly cut open with NheI-XhoI, the third insert was obtained by XbaI-XhoI digestion from its

monocistronic construct, and the same ligation procedure was followed as before resulting in a tri-cistronic construct.

3.1.8

Designed Polycistronic Constructs	**Vector backbone, resistance**
His-SpG2•SpSMN	pETM-11, Kanamycin
SpG2•His-SpSMN•SpG8	pETM-13, Kanamycin
His-SpG8•SpG7•SpG6	pETM-11, Kanamycin
His-SpG7•SpG6	pETM-11, Kanamycin
His-SpG2$^{\Delta arm}$•SpSMN	pETM-11, Kanamycin
His-SpG2$^{\Delta arm}$•SpSMN$^{\Delta YG}$	pETM-11, Kanamycin
His-SpG2•SpSMN$^{\Delta YG}$•SpG8	pETM-11, Kanamycin
His-SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36\text{-}119}$	pETM-11, Kanamycin
His-SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36\text{-}117}$	pETM-11, Kanamycin
His-SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36\text{-}111}$	pETM-11, Kanamycin
His-SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36\text{-}109}$	pETM-11, Kanamycin
His-SpG8$^{\Delta N58}$•SpG7•SpG6	pETM-11, Kanamycin
His-SpG8$^{\Delta loop}$•SpG7•SpG6	pETM-11, Kanamycin

3.1.9

Web tools and Software
1. Expasy webtools
2. EBI webtools
3. PRALINE multiple sequence alignment, secondary structure prediction
4. ESPript multiple sequence alignment
5. UniProtKB
6. NCBI
7. PyMol visualization software
8. Chimera visualization software
9. PHENIX
10. PHASER
11. COOT
12. ATSAS package 3.0

3.2 Methods

3.2.1 Molecular biological methods

3.2.1.1 Transformation of chemically competent *E. coli* cells

The RbCl method was employed to prepare competent *E. coli* cells (Hanahan, 1983). Diluted pre-culture was grown until OD600 0.6. Then the culture was cooled down on ice-water bath for 15 min and harvested by centrifugation. Pellet was resuspended in cold TFB-I buffer and afterwards in cold TFB-II buffer. Competent cells were aliquoted, and flash frozen in liquid nitrogen and stored at -80° C. For transformation, 100 µl of competent cells were first thawed on ice for 10 min. Afterwards, 100 ng of plasmid was added and incubated on ice for 20-30 min. Then, the cells were subjected to heat shock at 42 °C for 1 min, followed by 2 min incubation on ice. 1 ml LB medium was added to the sample and left to grow for 1 h at 37 °C and 1300 rpm. Then the cells were pelleted at 1500 g for 4 min and excess liquid was discarded. The pellet was resuspended in the remaining 100 µl medium and plated on LB-Agar plates containing appropriate antibiotics. Then the plates were incubated overnight at 37 °C and colonies were picked and selected as described in 3.1.3.

3.2.1.2 Agarose gel electrophoresis (analytical and preparative)

For the visualization of nucleic acids, 0.8-1.0% agarose gels containing ethidium bromide were prepared in 1X TAE buffer. The DNA samples were mixed with 6X loading dye and the run was performed in 1X TAE buffer at 100 V. After the run, the DNA fragments were visualized using an UV transilluminator. Gene Ruler DNA ladder Mix (Thermo Scientific) was employed as molecular weight marker.

3.2.1.3 Polymerase Chain Reaction (PCR)

Individual SpSMN constructs were PCR amplified using KAPA HiFi PCR kit (KAPA Biosystems) according to the manufacturer’s protocol. Amplified fragments were purified using NucleoSpin Gel and PCR clean up kit (Machery-Nagel). For introduction of single point mutations, the Quick-Change site-directed mutagenesis system (Stratagene, USA) was used.

3.2.1.4 Molecular cloning

All steps of molecular cloning were performed following Sambrook and Russel (2001). PCR products and empty vectors were digested using restriction enzymes from Fermentas according to manufacturer's specifications. Ligation of insert and vectors were performed using T4 DNA ligase kit (Thermo Scientific) overnight at 16 °C. Following this, the ligation reaction was transformed into DH5α competent cells (see 3.1.1). Then, constructs were purified from several colonies using NucleoSpin Plasmid Mini Kit (Machery-Nagel, REF 740588) and checked for the presence of correct inserts by restriction digestion. Concentrations of DNA samples were determined by NanoVue spectrometer (GE Life Sciences) at 260 nm. Positive clones were further analysed by sanger sequencing (Eurofins Genomics sequencing services). Sequencing results were aligned with expected nucleotide sequence using Clustal Omega web tool.

3.2.2 Biochemical methods

3.2.2.1 Heterologous protein expression in *E. coli*

Designed plasmid with mono- di- or tri-cistronic construct of the SpSMN complex components was transformed into BL21 (DE3) competent *E. coli* cells. 4.5 L of TB medium and 500 ml of 10X TB salts were prewarmed at 37 °C overnight. The following day, final culture medium was prepared accordingly (see section 3.1.1). Colonies from the transformation plates were washed into the final medium and the medium was aliquoted into 6X 5 L baffled flasks. The inoculated culture was then grown to an OD600 of 1.0 at 37 °C and 215 rpm. Hereafter, protein expression was induced by adding 0.5 mM IPTG, and the cultures were left to grow for 20 h at 16 °C. Cells were harvested by centrifugation at 5000 g for 25 min at 16 °C and resuspended in chilled Ni-NTA lysis buffer. Resuspended cells were flash frozen in liquid nitrogen and stored in -20 °C until further use.

3.2.2.2 Ni-NTA affinity purification

Frozen cell suspensions were quickly thawed using lukewarm water and protease inhibitors were added to 1X concentration (see section 3.1.1). Hereafter, all the steps of purification were performed at 4 °C or on ice using cold buffers and equipment. The cells were then placed on an ice-water bath and lysed by sonication until homogeneity (Branson Sonifier 250, duty cycle 50%, output control 10, 8X 1 min pulse with 2 min breaks in between). Lysed cell suspension was centrifuged at 45,000 rpm for 1 h at 4

°C (Beckman coulter, 45 Ti). After this, Ni-NTA agarose beads pre-equilibrated with Ni-NTA lysis buffer was added to the supernatant and incubated for 2 h on a rotor at 4 °C. Using a gravity flow flex column (Sigma-Aldrich), the beads were then collected and washed with 20-40 bed volume (BV) of Ni-NTA lysis buffer. Bound His-tagged proteins were then eluted in fractions using 6-10 BV of Ni-NTA elution buffer. Protein concentration were measured by Bradford or UV280 (see section). Fractions were then analysed by SDS-PAGE to verify the presence of target proteins.

3.2.2.3 Lämmli SDS-PAGE and Tris-Tricine SDS-PAGE

Proteins samples were separated by denaturing polyacrylamide gel-electrophoresis. In this work 12% Lämmli SDS-PAGE gels were employed; for most of the experiments, 13 or 15% Tris-Tricine SDS-PAGE were used. Prior to loading on the gel, the samples were mixed with 1X SDS-buffer and incubated at 95 °C for 5 min. The electrophoretic run was performed at a constant power of 15 Watt. To detect proteins, the gels were stained with Coomassie staining solution for 1 h at RT. Afterwards, the Coomassie solution was washed away with water and the gels were incubated with destaining solution for approximately 1 h at RT until protein bands were visible.

3.2.2.4 Dialysis and removal of His-tag by TEV protease

After the Ni-NTA purification, fractions containing the target proteins were combined and dialysed overnight against dialysis buffer at 4 °C and gentle stirring. To cleave the His-tag, 1-2% (w/w) of TEV protease was added to the samples during dialysis. The following day, completion of His-tag cleavage was checked by comparing uncleaved and cleaved samples on SDS-PAGE. For further purification steps or storage, the dialysed samples were concentrated to appropriate volumes using Vivaspin centricons (Sartorius).

3.2.2.5 Calibration of gelfiltration columns

In order to estimate the hydrodynamic sizes of protein complexes, the gelfiltration columns were calibrated by known globular standards (Sigma GE28-4038-42). The retention volumes were fitted to obtain a linear calibration curve in the log MW vs log Elution volume plot.

3.2.2.6 Preparative and analytical gelfiltration chromatography

To further purify/characterize the protein complexes, various gelfiltration chromatography columns were used with either ÄKTA Explorer 100 or ÄKTA purifier 10 systems. Prior to application, all protein samples were centrifuged at 10,000 g for 10 min at 4 °C to remove aggregates. Protein samples with amounts >20 mg were applied onto preparative gelfiltration columns such as Superdex 75 26/60 and Superdex 200 26/60 (GE Life Sciences). The samples were loaded in a maximum volume of 2.5 ml and eluted at a flow rate of 2 ml/min. 4 ml fractions were collected for approximately 1 column volume (CV). Eluted fractions were then analysed by SDS-PAGE. For accurate characterization of the hydrodynamic sizes of the purified complexes, analytical gelfiltration columns such as Superdex 75 10/300, Superdex 200 10/300 and Superose 6 10/300 (GE Life Sciences) were used. For this, 0.5-1.0 mg protein samples were loaded in a maximum volume of 0.5 ml and eluted at a flow rate of 0.3-0.5 ml/min. 0.4 ml fractions were collected for approximately 1 CV. Eluted fractions were then analysed by SDS-PAGE.

3.2.2.7 Complexation assay using analytical gelfiltration chromatography

Interaction between SpSMN sub-complexes were assessed using gelfiltration chromatography. For this, equimolar amounts of sub-complexes were first mixed and placed on ice for 15 min followed by 5 min incubation at 37 °C. The mixture was placed once again on ice for a final 15 min and then centrifuged at 10,000 g for 10 min at 4 °C. Afterwards, the sample was applied on an analytical gelfiltration column and the fractions analysed by SDS-PAGE.

3.2.2.8 Measurement of protein concentrations by UV absorbance at 280 nm

Protein concentrations were measured using Beer-Lambert law which is given by $A=\varepsilon.c.L$, where A is absorbance at 280 nm, ε is the co-efficient of extinction of the protein/protein complex, and L is the path length of the UV spectrophotometer. ε for proteins or protein complexes were obtained from their linear amino acid sequences using ProtParam wed tool of Expasy. L for the spectrophotometer was 1 cm (Eppendorf BioSpectrometer).

3.2.2.9 Regeneration of Ni-NTA agarose beads

All regeneration steps were performed at RT using gravity flow flex columns. Used Ni-NTA beads were first washed with 1M imidazole to remove bound proteins, followed by extensive wash with water. The bound Ni ions were stripped from the agarose material using 0.5M EDTA, followed by extensive wash with water. The beads were then washed with 1M NaOH solution to remove hydrophobically bound proteins, followed by extensive wash with water. Afterwards, 6M guanidine hydrochloride solution was added to the beads and incubated for few hours to overnight, followed by extensive wash with water. The cleaned beads were then incubated with 0.2M $NiCl_2$ solution overnight on a rotor to replenish the Ni ions, followed by extensive wash with water. Unbound Ni ions were removed by washing the freshly recharged beads with at least 10 BV of 1M imidazole, followed by extensive wash with water. Finally, the water was replaced by 20% ethanol, and the beads were stored at 4 °C until further use.

3.2.2.10 Cleaning of gelfiltration columns

In order to get rid of hydrophobically bound protein contaminant from the Superdex material, all gelfiltration columns were cleaned by reverse flow. The first wash was performed by 0.5 CV 1M NaOH followed by 0.5 of water. Afterwards, the column was washed with 0.5 CV 6M guanidine hydrochloride, followed by 0.5 CV water. At the end, the column was supplied with 2 CV 20% ethanol for storage.

3.2.3 Structural biological methods

3.2.3.1 Crystallization of $SpSMN^{\Delta 36\text{-}119}$

$SpG2^{\Delta arm}$•$SpSMN^{\Delta 36\text{-}119}$ complex at a concentration of 19.7 mg/ml was screened for crystallization using 9 different (commercial and in-house) crystallization screens. An initial hit (needle shaped crystals) was obtained in a condition Natrix-H5 (80 mM KCl, 40 mM Hepes 7.0, 60% 2-methylpentanediol 12 mM spermine tetrahydrochloride). The crystals could be obtained again in the same condition without spermine tetrahydrochloride which ruled out the possibility of spermine crystals. Final crystals of approximately 0.6 mm sizes were obtained in an optimized condition containing 80 mM KCl, 40 mM Hepes at 3 different pH (6.8, 6.9, 7.2) and 65% 2-methylpentanediol, by hanging-drop vapour diffusion method in 24 well plates (2 µl + 2 µl).

3.2.3.2 Structure determination by molecular replacement

The X-ray diffraction data were collected at the ID-14 beamline at the European Synchrotron Radiation Facility (ESRF, Grenoble). The data were then processed by XDS (Kabsch 2010). Initial attempts to solve the structure by molecular replacement using homologous Gemin2 structure (PDB ID 4V98) were unsuccessful. Molecular replacement using the already available SpYG-box structure (PDB ID 4RG5) was successful. Molecular replacement was done using PHASER (McCoy et al, 2007). Structure refinement was performed by PHENIX (Adams et al., 2010). Displayed structural models were generated using PyMOL (Schrödinger, LLC).

3.2.3.3 Small angle X-ray scattering of SpSMN complexes

3.2.3.3.1 Introductory notes

Macromolecular samples of proteins or protein-nucleic-acid complexes, that are classified as difficult or impossible to analyze by traditional structural biological methods such as X-ray crystallography or electron cryo-microscopy, exhibit two major biophysical properties. One, conformational polydispersity arising from inherently flexible (intrinsically disordered) regions that can adopt exceptionally large numbers of conformations in solution, and two, polydispersity arising from co-existence of multiple oligomeric states within the sample. These two situations give rise to far from ideal samples and cannot be analyzed due to technical limitations of these traditional high-resolution methods. **Small angle X-ray scattering (SAXS)** is a structural biological technique where macromolecular samples are studied in solution. Although all sample types can be studied by SAXS, through decades of technical improvements SAXS has been shown to successfully help study the afore mentioned non-ideal samples (Kikhney et al. 2015; Bernado et al. 2012). Articles for SAXS data collection and interpretations have been listed in the references section.

3.2.3.3.2 SAXS data collection

In a SAXS experiment, a sample of macromolecular solution in a quartz capillary is exposed to collimated monochromatic beam of X-rays (ESRF BM29 beamline, wavelength λ=1.54 Å, beam geometry 0.7 mm x 0.7 mm) and the intensity of the scattered beam is recorded by a 2D X-ray detector (PILATUS 1M) as a function of the scattering angle 2θ. The **scattering intensity I(s)** is represented as a function of the **scattering vector s=(4πsinθ)/λ**. Data was collected from s=0.032 nm^{-1} upto s=4.994

nm^{-1}). Since the macromolecules and the solvent particles in the sample are present in all possible orientations, the resulting scattering intensity is isotropic and is radially averaged yielding a 1D scattering curve. Scattering of the pure solvent is also recorded and subtracted from the sample scattering which yields the final 1D scattering curve of macromolecules inside the solution.

3.2.3.3.3 SAXS data interpretation

1D SAXS curve: ***lnI(s) vs s*** *or* ***lnI(s) vs ln(s)***

The 1D scattering curve of macromolecules (usually represented as semi-log plot lnI(s) vs s or as double logarithmic plot lnI(s) vs ln(s), is not informative in itself but rather must be transformed into several different forms to extract the shape and size information of the macromolecules. The most important transformations, their principles, and the information they provide are described below. The SAXS data was processed using the ATSAS 3.0 package (Franke et al. 2017).

(*Data quality assessment and data representations in the results section were followed closely as the prescribed guidelines (Jacques et al. 2010; Trewhella et al. 2017; Chaudhuri et al. 2017))*

Radius of gyration (Rg) from ***Guinier plot: lnI(s) vs s^2***

The Guinier plot was derived from the Guinier approximation which states that scattering close to the origin is related to the radius of gyration (Rg) of the macromolecule and is given by I(s) = I(0) e^(− $s^2Rg^2/3$), where I(0) is the theoretical 0^{th} angle scattering. In the log scale, lnI(s) = lnI(0) − ($s^2Rg^2/3$), a plot of lnI(s) vs s^2 follows a straight line within a limited range (π/Dmax < sRg < 1.3), where Dmax is the maximum dimension of the particle. From the slope (−$Rg^2/3$) of this straight line the Rg of the macromolecule were calculated. While comparing data from two different macromolecules of similar sizes, linearity up to a smaller angular range indicates relative flexibility compared to globular and compact macromolecules which exhibit linearity for a longer angular range. (Receveur-Bréchot et al. 2012)

Quantitative flexibility from ***Dimensionless Kratky plot: $(sRg)^2I(s)/I(0)$ vs sRg***

The SAXS data transformed as traditional Kratky plot (s^2I(s) vs s) reliably provides qualitative information about the macromolecule's nature of flexibility/compactness. The scattering intensity I(s) of rigid globular proteins with well-defined electron density

contrast exhibit decay as s^{-4} resulting in a bell-shaped curve with a distinct maximum followed by gradual convergence of signal to 0 at higher s. The I(s) of a fully unfolded protein (electron density contrast not well-defined) on the other hand decays as s^{-2} and therefore results in a plateau followed by constant increase at higher s without a distinct maximum. Globular proteins with long unstructured tails or multidomain proteins with flexible linkers exhibit an intermediate behavior. Kratky plots, however, are scaled by the particles' volume and concentration, and are therefore not quantitative. For a quantitative view of the degree of flexibility/compactness between proteins of different sizes, molecular weights and concentrations, a Dimensionless Kratky plot was generated by normalizing the individual curves for respective Rg (related to size) and I(0) (related to molecular weight and concentration). Here, globular proteins show an initial bell-shaped region with a maximum of 1.104 at sRg=√3 followed by constant decrease in signal at higher s. Fully unfolded proteins do not exhibit a maximum and show a constant increase in signal at higher s. Multidomain proteins with flexible linkers display intermediate behavior between these two. In such cases, after the initial bell-shaped peak at sRg=√3 (corresponding to all globular domains), the relative trajectories of signals was used to quantitatively assess the degree of flexibility between different samples. (Receveur-Bréchot et al. 2012)

Particle dimension (Dmax) from ***pairwise distance distribution function P(r)***

The P(r) function is obtained by taking indirect fourier transform of the scattering curve. It is a histogram of the distribution of interatomic distances between all pairs of atoms within the macromolecule and can reveal the domain architecture of a protein. This allows one to determine the maximum dimension of the macromolecule (Dmax) as well as radius of gyration (Rg). The P(r) curve of an ideally globular protein shows a gaussian curve with a single peak whereas that of protein with two domains connected with a linker shows a bimodal curve. In such cases, change in the peak pairwise distance can indicate compaction or elongation respectively, which will be accompanied by change in Dmax. Of note, matching Rg values obtained from Guinier approximation and P(r) function for the same scattering curve was used as an indicator of the internal consistency (and hence the quality) of the scattering data (Receveur-Bréchot et al. 2012).

Porod volume (Vp) and molecular weight (MW) from ***Porod Invariant Qp***

As explained earlier, for a globular protein with sharp electron density contrast between the solute and the solvent, transformation of its scattering data as $s^2I(s)$ vs s (Kratky plot) results in a bell-shaped curve. The area under the curve is well-defined and is given by the Porod Invariant termed Qp which is scaled by the solvent excluded volume of the particle (Vp) and its concentration (related to I(0)),

$$Q = \int_0^\infty I(s)s^2 ds = 2\pi I_0/V_p$$

Assuming a mass density of 1.37 g cm^{-3} for globular proteins, the molecular weight (in Daltons) is related to Vp by the following equation,

$$\mathrm{MW} = \frac{V_p}{1.66}$$

Hence, the accuracy of MW determination with this method greatly depends on the accuracy of determination of the Porod Invariant Qp (area under the curve in the Kratky plot). While for compact, rigid and globular proteins the area under the curve is well define owing to the intensity decay as s^{-4}, for flexible and partially folded proteins the decay is much slower resulting in large (and inaccurate) area under the curve which is prone to buffer subtraction errors at higher s. (Mylonas et al. 2007; Rambo et al. 2013)

Quantitative flexibility from ***Porod-Debye fourth power law: $s^4I(s)$ vs s^4***

This is also an indicator of macromolecular flexibility. Scattering of well-folded globular proteins decays as s^{-4}. Hence, transforming the data as $s^4I(s)$ vs s^4 shows a plateau within the low-resolution region of the data ($0 < s < 3.2 nm^{-1}$) and is indicative of sharp electron density contrast between the particles and the solvent. Intrinsically disordered proteins or highly flexible proteins, however, do not show a plateau in this region owing to relatively diffuse electron density contrast between particles and the solvent and slower decay of intensity (Rambo et al. 2011).

3.2.3.3.4 SAXS based structural modeling

As discussed in the previous section, important biophysical parameters such as MW, Rg, Dmax, Vp and P(r) were obtained from SAXS data. But in addition to this, SAXS data (angular range of up to 2-3 nm^{-1}) was used to generate low-resolution structural

models (bead models) of the macromolecule *ab initio* using the program DAMMIF (part of ATSAS 3.0). Typically, starting with a search volume of diameter Dmax of the particle, dummy residues/beads are assigned until optimal values of parameters of the model are achieved. These values must give rise to a computationally generated scattering curve that fits the experimental data. The process is done iteratively until a minimum value for a discrepancy term called X^2 between the experimental scattering intensity $I_{exp}(s)$ and computationally calculated scattering intensity $I_{calc}(s)$ is achieved. X^2 is given by,

$$\chi^2 = \frac{1}{N-1}\sum_{j=1}^{N}\left[\frac{I_{\exp}(s_j) - cI_{calc}(s_j)}{\sigma(s_j)}\right]^2$$

Where, N is the number of angular points, c is a scaling factor, and σ represents experimental errors. A X^2 value of around 1.0 is considered an excellent fit of the generated model to the experimental data. Generation of structural models directly from SAXS scattering curves in this method is straightforward for rigid and non-flexible macromolecules that are conformationally monodisperse.

3.2.3.3.5 SAXS based structural modeling: Ensemble Optimization Method (EOM)

Due to conformational polydispersity, scattering from highly flexible macromolecules (intrinsically disordered proteins, multidomain proteins with flexible linkers, etc.) is the average scattering of all possible conformations and/or oligomeric states present in the sample. If there are K conformations in the sample, the resulting SAXS scattering pattern I(s) is the sum of individual scattering from all K conformers,

$$I(s) = \sum_{k=1}^{K} \nu_k I_k(s),$$

Where, v_k and $I_k(s)$ are volume fraction and scattering intensity of the k^{th} component. Depending on the degree of flexibility of the macromolecule, the number of conformations can be an astronomically high number (K>>1). While it is impossible to generate a single or even a few models directly from the experimental SAXS curve in such cases, it is possible, using the Ensemble Optimization Method (EOM), to reliably derive the distribution of biophysical parameters (Rg and Dmax) of the flexible macromolecule within the sample, that adequately describe the experimental

scattering curve. The basic principle of EOM is simple: 1. Computational generation of a pool of large number of theoretical conformers of the flexible polypeptide chain, 2. Selection of a sub-population of conformers whose computed theoretical average scattering curve fits the experimental scattering curve with minimum X^2 discrepancy. In an EOM analysis, first a pool M of 10,000 theoretical conformers of an unstructured polypeptide is generated. In case of multidomain proteins with long unstructured linkers, computational models of the flexible linker are generated with the folded domains attached, covering almost all possible relative spatial positions of the domains with respect to each other. Theoretical scattering profiles from each conformer is computed. Then, C number of ensembles (usually 50) are randomly selected, each containing N distinct conformers (usually 50). Computed scattering profile of each ensemble is the average of the sum of the computed scattering profile of each conformer within the ensemble,

$$I(s) = \frac{1}{N}\sum_{n=1}^{N} I_n(s)$$

Each ensemble is subjected to two genetic operations, mutations, and crossings. During mutations, conformers from each ensemble is exchanged with conformers from the original pool M or from ensembles of the same generation. During crossing, sets of conformers from 2 randomly selected ensembles are exchanged. Both these processes maintain the size of each ensemble, i.e., 50. At the end of the two genetic operations 3C number of ensembles are obtained with a total of 3CxN conformers. Then, C number of ensembles, each yielding the best fitting scattering profile to the experimental data (lowest X^2 discrepancy) are chosen which undergo a further round of mutations and crossings. This process of mutations and crossings is performed for at least 500-1000 generations yielding an optimized final single ensemble (with N conformers) with the lowest X^2 discrepancy. The whole EOM run is performed R times to obtain R number of best fitting ensembles. All conformers of the final set of ensembles (a pool of RxN conformers) are analyzed and the distributions of Rg and Dmax are plotted as a function of frequency within this final optimized pool. Comparing this distribution with the theoretical distribution of the original pool M of 10,000 conformations may reveal conformational preference of the particles within the solution. This is especially interesting to study changes in the size upon binding or dissociation of substrate or subunit from a macromolecular complex. From the ATSAS

data analysis software package, RANCH is used for generation of theoretical models and GAJOE is used for ensemble optimization. (Bernado et al. 2007; Bernado et al. 2012; Tria et al. 2015)

3.2.3.3.6 SEC-SAXS strategy

SEC-SAXS is a small angle x-ray scattering data collection strategy where the scattering profile is collected as the protein sample elutes from a size exclusion chromatography (SEC) column (Superdex 200 10/300, 1 ml/min, 20 °C) and passes through a capillary exposed to x-ray beam. The data is collected as thousands of frames (1800 frames, 1 sec exposure per frame) over the elution profile where each frame in the data represents the scattering of the protein solution in the capillary at that instant. Buffer frames before or after the protein peak are used for background subtraction. For each frame within the peak, the radius of gyration (Rg) and molecular weight of the eluting particles is calculated. This reveals the distribution of particle sizes within the peak(s) (monodisperse or polydisperse). For a monodisperse protein sample, all frames corresponding to the protein peak can be combined by scaling and averaging for further processing. For a polydisperse sample, frames of constant Rg and MW can be combined for further processing.

4 RESULTS

4.1 Immunoprecipitation of endogenous SpSMN complex and identification by mass spectrometry

Orthologs of SMN and Gemin2 in *S. pombe* (previously termed Yab8 and Yip1 respectively) have long been identified and characterized (Hannus et al. 2000; Owen et al. 2000; Paushkin et al. 2000). So far, orthologs of the remaining Gemins have not been identified. Using computational methods and yeast two-hybrid assays, orthologs of Gemin8, Gemin7 and Gemin6 have been identified in the laboratory of Dr. Remy Bordonne and were shown to be essential for viability (unpublished data). For UniProt accession codes, sequences, and multiple sequence alignments of the proteins, refer to annexure (section 8). Using GFP-SpG6 as the sole source of SpG6, the endogenous SpSMN complex as well as the Sm proteins could be purified from *S. pombe* whole cell extracts (Figure 4.1) and identified by mass spectrometry (Table 4.1).

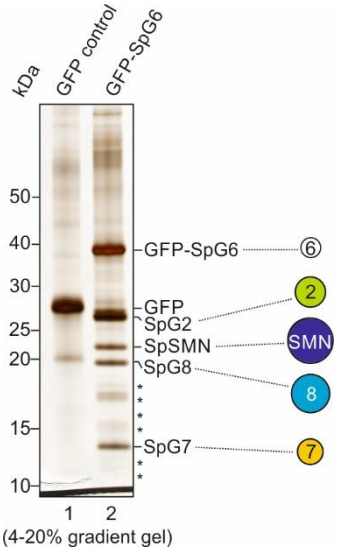


Table 4.1: Mass spectrometry analysis of GFP-SpG6 IP

Protein name	MW (kDa)	PomBase ID	Nr. of Peptides identified		Coverage (%)	
			IP	Control	IP	Control
SpGemin7 (SpG7)	10.336	SPBC32F12.16	10	2	87	25
SpGemin8 (SpG8)	19.685	SPBC16H5.15	28	-	80	-
SpSMN	17.385	SPAC2G11.08c	13	-	74	-
SpGemin6 (SpG6)	10.623	SPAC4D7.15	3	-	74	-
SpGemin2 (SpG2)	26.994	SPAC19B12.12c	15	-	46	-
SmD2	13.095	SPAC2C4.03c	6	2	41	15
SmB	15.476	SPAC26A3.08	7	-	37	-
SmD1	13.089	SPAC27D7.07c	3	-	30	-
SmD3	11.033	SPBC19C2.14	3	1	25	7
SmG	8.604	SPBC4B4.05	2	-	22	-
SmF	8.660	SPBC3E7.14	1	-	13	-

Figure 4.1: Endogenous SpSMN complex
Immunoprecipitation of endogenous SpSMN complex using GFP-SpG6 from *S. pombe* whole cell extract (lane 2, performed by Dr. Remy Bordonne, IGMM, CNRS). GFP control IP is shown in lane 1. All five proteins are indicated by their representative cartoons. *Possible positions of Sm proteins.

4.2 Co-expression and purification of SpSMN complex components

4.2.1 Introductory notes

A comprehensive interaction map of the human SMN complex components has been determined by Otter et al. (2007). Although the SpSMN complex consists of only the core subunits SpG2, SpSMN, SpG8, SpG7 and SpG6, it is possible that they exhibit a similar interaction pattern as their human counterparts due to structural homology of the subunits (Figure 1.5 B). For the recombinant expression of SpSMN complex components, a co-expression strategy based on the interaction pattern of the human SMN complex was devised by designing di- and tri-cistronic vectors (see sections 3.1.7 and 3.1.8) with only one subunit containing an affinity tag serving as bait. The designed constructs were,

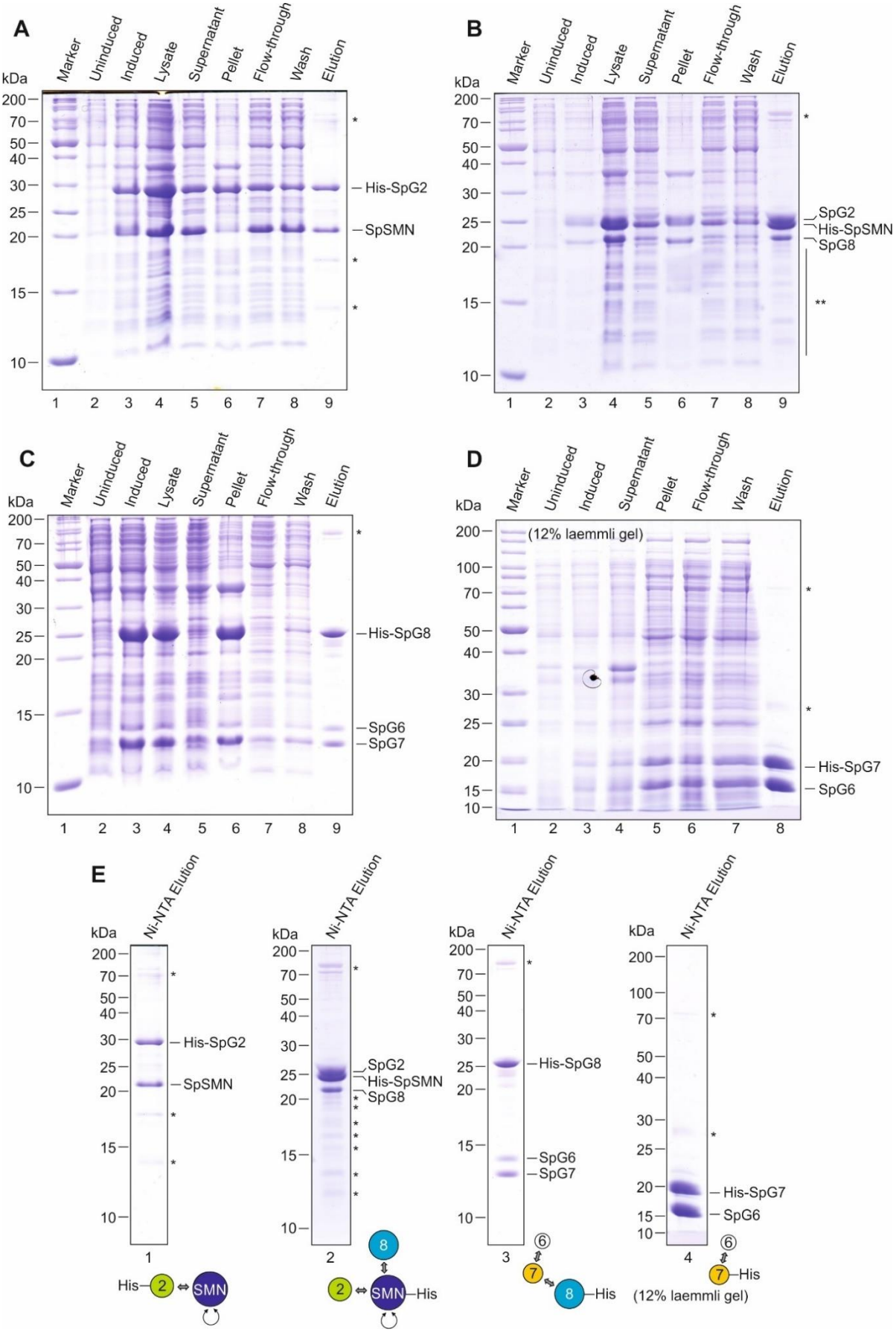
A. His-SpG2•SpSMN
B. SpG2•His-SpSMN•SpG8
C. His-SpG8•SpG7•SpG6
D. His-SpG7•SpG6

4.2.2 Prokaryotic protein expression and Ni-NTA affinity purification

In order to test the interaction patterns within the SpSMN complex subunits, the afore mentioned constructs were expressed in *E. coli* and purified by Ni-NTA chromatography. The samples were analyzed on 15% Tris-Tricine SDS-PAGE gels unless otherwise mentioned. As shown in Figure 4.2 A-D, all sub-complexes were co-expressed in *E. coli* and could be purified with this strategy (Elution lanes). Within constructs A and D, the individual subunits are highly stable and do not show any signs of degradation. In construct B, the SpSMN subunit was His-tagged instead of the SpG2 subunit to allow separation of the SpSMN and SpG8 bands on SDS-PAGE. Multiple degradation bands (**) could be seen below the band corresponding to SpG8 indicating that the co-purified SpG8 subunit might be unstable.

Figure 4.2: Ni-NTA purifications of recombinantly expressed SpSMN complex components

Heterooligomeric SpSMN sub-complexes were co-expressed in E. coli and purified by Ni-NTA chromatography via a single His-tag on one subunit. A-D: SDS-PAGE (15% Tris-tricine unless otherwise specified) of the purification of His-SpG2•SpSMN (A), SpG2•His-SpSMN•SpG8 (B), His-SpG8•SpG7•SpG6 (C) and His-SpG7•SpG6 (D). E (lanes 1-4): overview of the purified complexes. Under each lane a representative cartoon of the complex is depicted showing the location of His-tag.


A
Marker
Uninduced
Induced
Lysate
Supernatant
Pellet
Flow-through
Wash
Elution
kDa
200
70
50
40
30
25
20
15
10
His-SpG2
SpSMN
1 2 3 4 5 6 7 8 9
B
SpG2
His-SpSMN
SpG8
C
His-SpG8
SpG6
SpG7
D
(12% laemmli gel)
His-SpG7
SpG6
1 2 3 4 5 6 7 8
E
Ni-NTA Elution
His-SpG2
SpSMN
SpG2
His-SpSMN
SpG8
His-SpG8
SpG6
SpG7
His-SpG7
SpG6
His
SMN
(12% laemmli gel)

Within construct C, SpG7 and SpG6 subunits appear underrepresented compared to His-SpG8. This might be indicative of a non-stoichiometric complex or differential expression of subunits from the tri-cistronic construct. Nevertheless, the co-purification of untagged subunits in each of these cases, confirms their direct interaction with the His-tagged subunit and shows that SpSMN complex components exhibit a similar interaction pattern as their human counterparts. An overview of the purified complexes is shown in panel E.

4.2.3 Gelfiltration chromatography

To investigate whether the purified complexes were defined entities (rather than aggregates or heterogenous populations), they were analyzed by gelfiltration chromatography. The Ni-NTA elutions were subjected to dialysis (+TEV protease, except constructs B and D) and subsequently gelfiltration chromatography using various analytical columns. The gelfiltration chromatograms and SDS-PAGE of fractions are shown in Figure 4.3 columns 1 and 2, respectively. Both SpG2•SpSMN (calculated monomeric MW ~44 kDa) and the SpG2•His-SpSMN•SpG8 (calculated monomeric MW ~64 kDa) complexes exhibit unusually large and very similar hydrodynamic properties and elute from gelfiltration column near the 669 kDa MW marker (Figure 4.3 A-B, peak I and peak II respectively). Such hydrodynamic behavior necessitates further investigation of the oligomeric and conformational properties of the SpSMN's YG-box and the unstructured region (see sections 4.4 and 4.5). SpG8•SpG7•SpG6 (calculated monomeric MW ~41 kDa) eluted in the low molecular weight region between 44 kDa and 68 kDa MW markers indicating a trimeric complex of globular shape (Figure 4.3 C, peak I). His-SpG7•SpG6 (calculated monomeric MW 21 kDa) eluted between 66 kDa and 150 kDa MW markers (Figure 4.3 D, peak I) indicating higher order oligomers of the complex. Further assessment of the exact oligomeric states and stoichiometry of these complexes is not possible due to limitations of the technique (see section 4.8). In conclusion, the co-elution from gelfiltration of individual subunits within each complex reiterates their direct interactions and forms the basis for the reconstitution of the SpSMN pentameric complex.

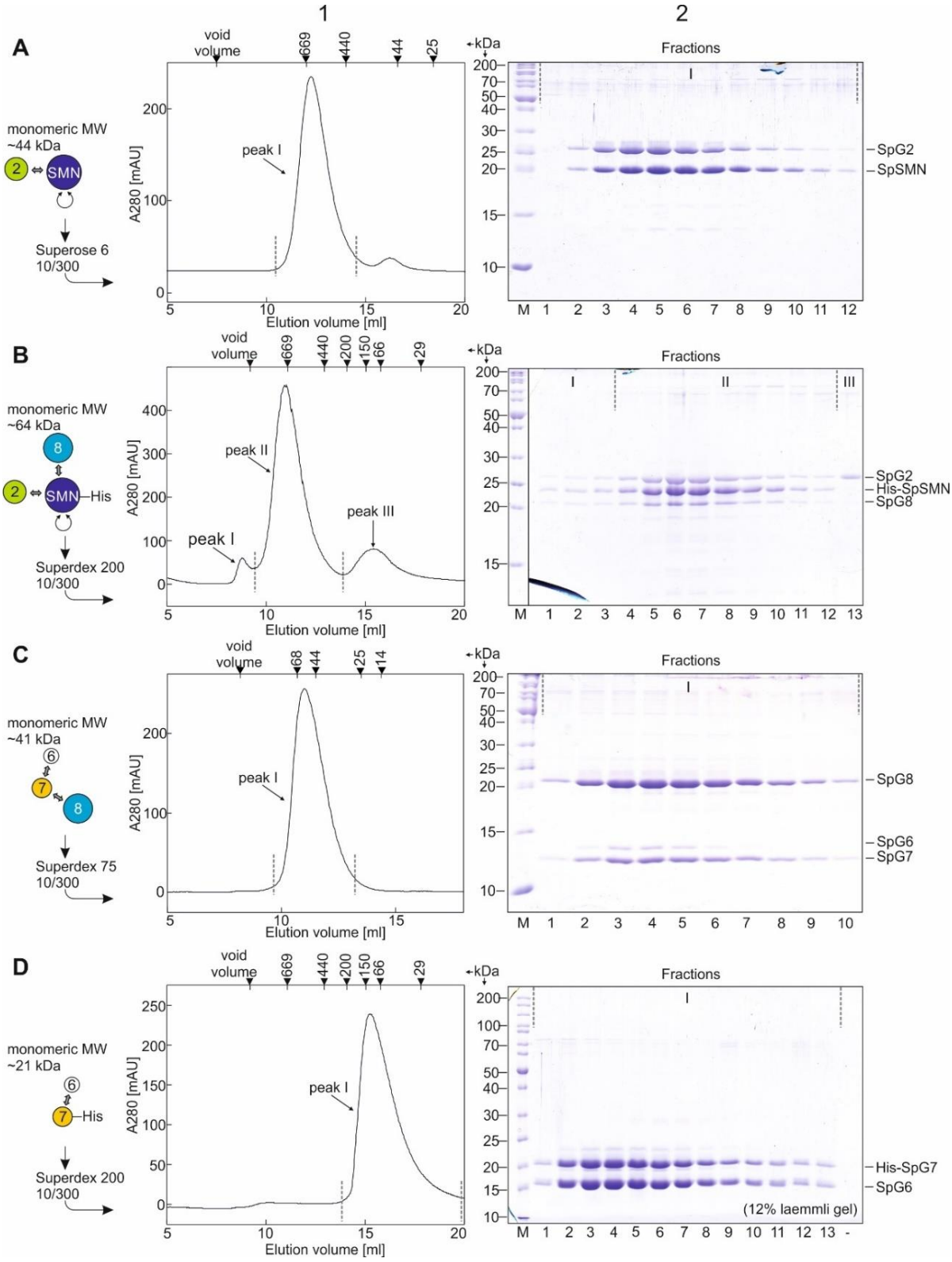


Figure 4.3: Gelfiltration chromatography of recombinantly purified SpSMN sub-complexes

SpSMN sub-complexes purified by Ni-NTA were analyzed by gelfiltration chromatography (after overnight dialysis with TEV protease except B and D) using Superdex 200, Superdex 75 and Superose 6 10/300 analytical columns. Elution profiles with positions of standard MW markers are shown for each chromatogram (panels 1). Monomeric MWs of each sub-complex is indicated next to the representative

cartoons. Fractions under the peak (region within grey dashed lines) were analyzed by SDS-PAGE (15% Tris-tricine unless specified) (panels 2). M= unstained marker.

4.3 In vitro reconstitution of SpSMN pentameric complex

4.3.1 SpG8 forms the link between SpG2•SpSMN and SpG7•SpG6

As shown in Figure 4.2 E, SpSMN sub-complexes were successfully purified as heterodimers (SpG2•SpSMN and SpG7•SpG6) and as heterotrimers (SpG2•SpSMN•SpG8 and SpG8•SpG7•SpG6), showing that the SpSMN complex components exhibit a similar interaction pattern as their human counterparts. This lays the foundation for the reconstitution of the pentameric SpSMN complex where the SpG8 serves as the link between SpG2•SpSMN and SpG7•SpG6. Interaction and complex formation were monitored by gelfiltration chromatography by following the shift and co-elution of smaller Gemins (SpG8, SpG7 and SpG6) with SpG2•SpSMN. Equimolar mixtures of sub-complexes (SpG2•SpSMN+SpG7•SpG6 and (SpG2•SpSMN+SpG8•SpG7•SpG6) were prepared and analyzed by gelfiltration chromatography using Superose 6 10/300 analytical column. The eluted fractions were then analyzed by 15% Tris-Tricine SDS-PAGE. The gelfiltration chromatogram and SDS-PAGE of fractions are shown in Figure 4.4 columns 1 and 2, respectively. In the absence of SpG8 (Figure 4.4 A), the complex mixture is resolved into two distinct peaks and analysis of the fractions by SDS-PAGE does not show co-elution of SpG7•SpG6 with SpG2•SpSMN in peak I. SpG8•SpG7•SpG6, however, co-elutes with SpG2•SpSMN (Figure 4.4 B) in peak I, which illustrates the formation of the pentameric SpSMN complex. Almost identical elution volumes of SpG2•SpSMN and SpSMN pentamer indicate that SpG8•SpG7•SpG6 has minimal contribution towards the unusually large hydrodynamic sizes of SpSMN containing complexes, which might be attributed to SpSMN's YG-box and its unstructured region (see sections 4.4 and 4.5).

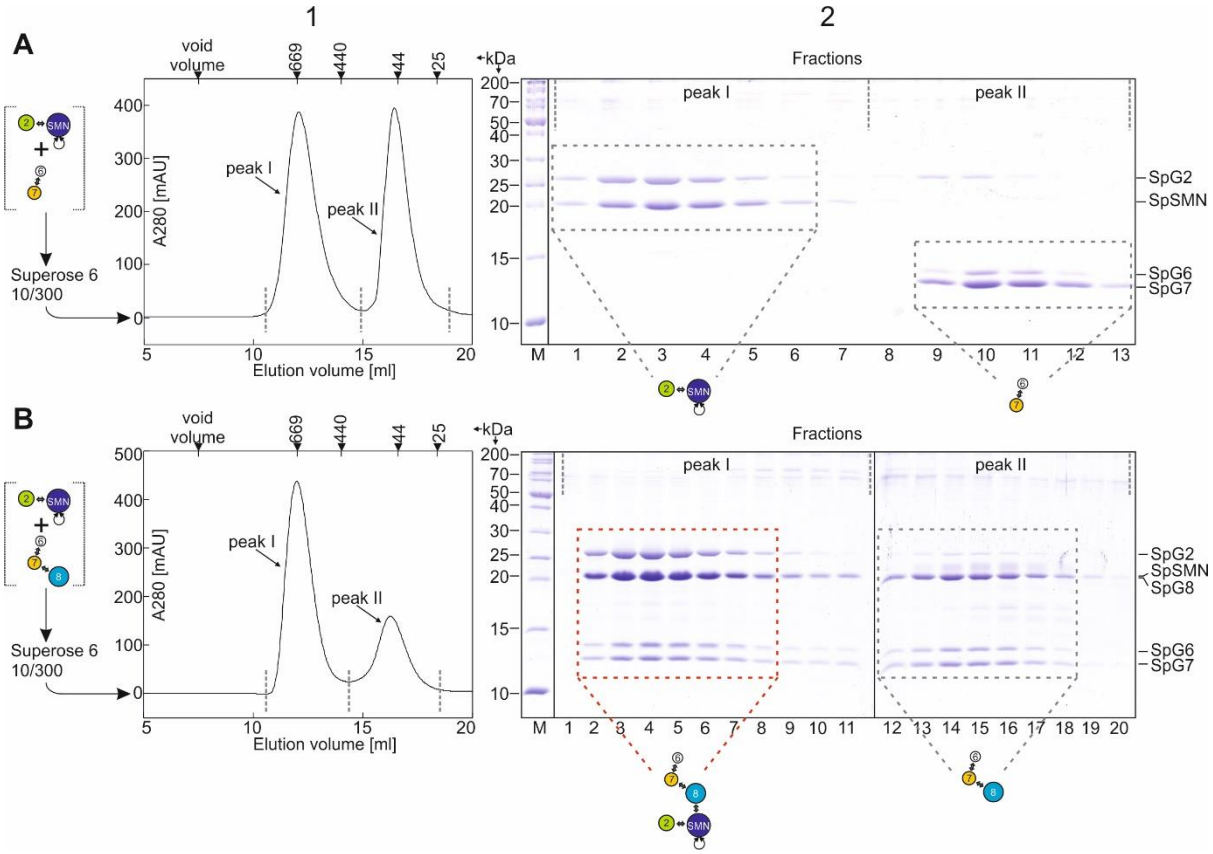


Figure 4.4: In vitro reconstitution of pentameric SpSMN complex from purified sub-complexes

SpG2•SpSMN was mixed in equimolar amounts with SpG7•SpG6 (A) and SpG8•SpG7•SpG6 (B). Both protein mixes were applied on Superose 6 10/300 gelfiltration column to monitor complex formation. Elution profiles with MW standards are shown for each chromatogram (panels 1). Fractions under eluting peaks (region within grey dashed lines) were analyzed by SDS-PAGE (panels 2). Complexes identified in the gels are indicated by grey boxes and a corresponding cartoon. The reconstituted pentameric complex is indicated by a red box. M= unstained marker.

4.4 Roles of the YG-box and the unstructured region of SpSMN

4.4.1 SpSMN$^{\Delta YG}$ and SpSMN$^{\Delta 36\text{-}119}$ cause remarkable loss in hydrodynamic size

SpG2•SpSMN, SpG2•SpSMN•SpG8 (Figure 4.3 A-B), and SpG2•SpSMN•SpG8•SpG7•SpG6 (Figure 4.4 B) exhibit unusually large and very similar hydrodynamic properties, which might be attributed to SpSMN's YG-box and its unstructured region (residues 36-119). The oligomeric properties of the YG-box domain of SpSMN as well as its human ortholog has been extensively studied as MBP fusion proteins. These studies have revealed that the YG-box forms at least a dimer through self-association. But so far, the individual contributions of the YG-box domain and the 84 residues long unstructured region (residues 36-119) on the size and oligomeric states of SpG2•SpSMN heterodimer has not been analyzed. To investigate this, variants of SpG2•SpSMN construct were designed lacking either SpSMN's C-terminal YG-box domain or the unstructured linker. The properties of the purified constructs were studied by gelfiltration. In addition, a truncation mutant of the similarly sized N-terminal substrate binding arm (80 residues) of SpG2, which also adopts an extended conformation, was included in the study. SpG2•SpSMN, SpG2$^{\Delta arm}$•SpSMN, SpG2$^{\Delta arm}$•SpSMN$^{\Delta YG}$ and SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36\text{-}119}$ were analyzed by gelfiltration chromatography using a Superdex 200 10/300 column. Truncation of the N-terminal 80 residues long substrate binding arm of SpG2 did not have any effect on the apparent molecular size of SpG2•SpSMN complex and eluted near 669 kDa marker (Figure 4.5 A-B, black dashed and dotted). On the other hand, removing the unstructured region of SpSMN of similar residue length (84 aa) resulted in a complex (SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36\text{-}119}$) nearly 10 times smaller in size, eluting near the 66 kDa marker (grey line). Secondly, truncation of the SpSMN's YG-box oligomerization domain (SpG2$^{\Delta arm}$•SpSMN$^{\Delta YG}$) also resulted in a nearly 10 times smaller complex eluting near the 66 kDa marker (red line). These observations suggest a combined role of SpSMN's YG-box domain and the unstructured region on the observed large hydrodynamic properties of SpSMN containing complexes.

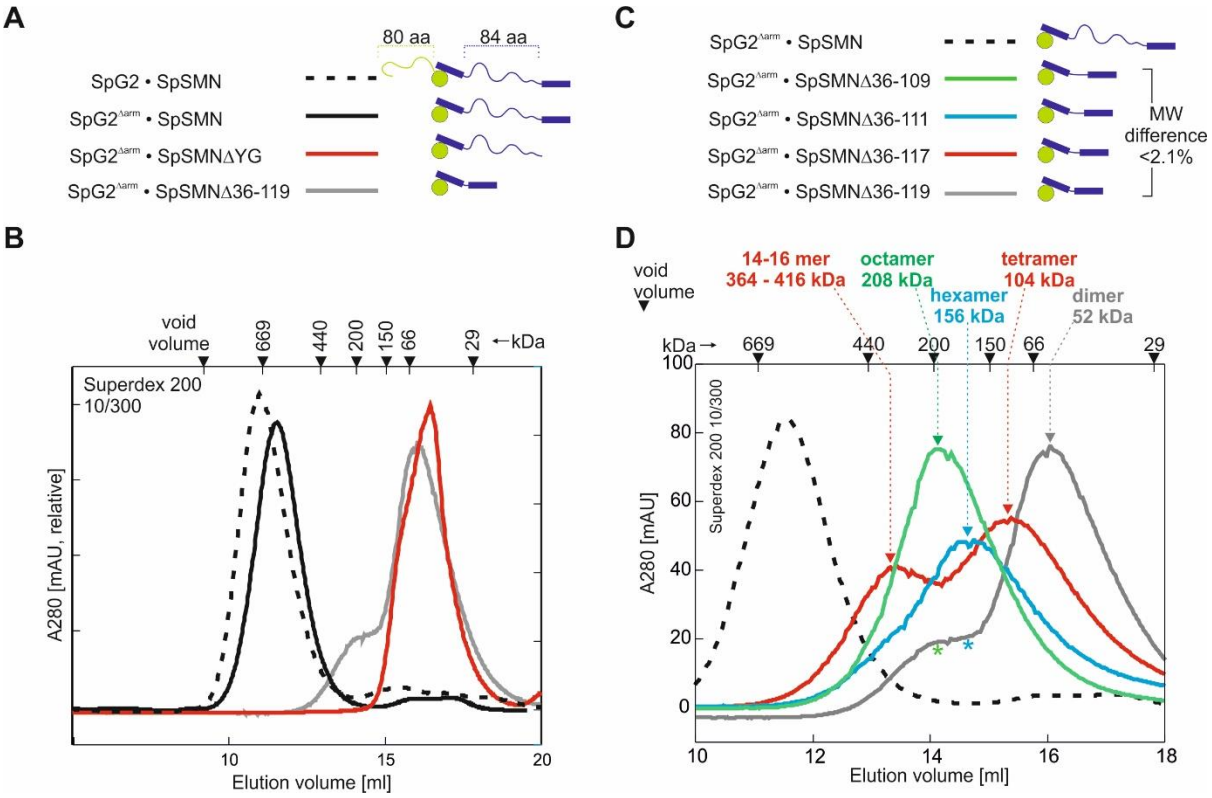


Figure 4.5: Role of SpSMN's YG-box domain and unstructured region on the hydrodynamic properties of SpSMN complexes

A-B: Gelfiltration studies of mutant constructs lacking SpG2's N-terminal 80 residues long arm (solid lines), SpSMN' 84 residues long unstructured region (grey line) and the YG-box oligomerization domain (red line). Wild type SpSMN•SpG2 is shown in dashed line. C-D: Determination of the oligomeric states of SpSMN by gelfiltration studies using unstructured region deleted constructs (Δ36-109, Δ36-111, Δ36-117, and Δ36-119). These gradual deletions were designed to eliminate potential steric clash by the SpG2 domain in the Δ36-119 construct which could preclude detection of otherwise possible oligomeric states. The complex with full-length SpSMN is shown in dashed line as control.

4.5 Oligomeric states of SpSMN

4.5.1 SpSMN can form Dimers, Tetramers, Hexamers and Octamers *in vitro*

Macromolecules emerge from a gelfiltration column in the decreasing order of their individual hydrodynamic radii. The retention volume of a globular protein correlates well with its actual molecular weight. Thus, using calibration of standard proteins (compact, globular) of known molecular weights, the size and/or oligomeric state of an unknown protein can be calculated. However, accuracy of the calculated molecular weight depends largely on the degree of compactness (globular/flexible) of the unknown protein relative to those of the standard proteins. Due to the long unstructured region of SpSMN, gelfiltration might not reveal accurate oligomeric states of SpSMN complexes. To accurately assess the oligomeric states of SpSMN through its YG-box, variants of SpG2•SpSMN construct with deletions of SpSMN's unstructured region were designed. The molecular weights and oligomeric states of the resulting relatively globular complexes were studied by gelfiltration. As shown in Figure 4.5 C, in addition to SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36\text{-}119}$**,** a series of deletion mutants with slightly longer unstructured region preceding the YG-box (SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36\text{-}117}$, SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36\text{-}111}$ and SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36\text{-}109}$) were designed. This was done in order to eliminate any effect of potential steric hindrance by SpG2, which could preclude detection of otherwise possible oligomeric states. Since the difference in molecular weights between the constructs is ≤ 2.1 %, oligomeric states resulting solely due to SpSMN oligomerization can be reliably identified by gelfiltration as a result of significant difference in apparent size. As shown in the chromatogram (Figure 4.5 D), oligomers of SpSMN ranging from dimers to octamers are clearly identified with this strategy. Interestingly, degree of oligomerization increased with increasing length of unstructured linker (Δ36-109 > Δ36-111 > Δ36-117 > Δ36-119). Although SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36\text{-}119}$ and SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36\text{-}117}$ predominantly resulted in dimers and tetramers, respectively, additional peaks corresponding to octamers and >octamers, respectively, were also found. These results illustrate the oligomeric properties of SpSMN and provide a basis for the observed large hydrodynamic properties of SpSMN containing complexes.

4.6 Crystal structure of SpSMN$^{\Delta 36\text{-}119}$ reveals YG-box dimers as the fundamental unit of higher order oligomers

4.6.1 Introductory notes

The first major feature of SpSMN which contributes to the large hydrodynamic properties of SpSMN complexes is the C-terminal YG-box oligomerization domain. As established in the previous sections, the SpSMN subunit is the central oligomeric core of SpSMN complex, suggesting that the mode of oligomerization might play a crucial role in determining the spatial positions of the peripheral subunits, and consequently the overall architecture of SpSMN complex. Understanding the structure of the oligomer may also explain the large molecular sizes seen in gelfiltration studies. Available structures of MBP-tagged YG-box (*S. pombe*: PDB 4RG5, human PDB 4GLI) illustrate a glycine-zipper dimerization of YG-box via the YxxGYxxG(Y/L)xxG motif, but they fail to explain how higher order oligomers of SMN are formed. Using analytical centrifugation and small angle X-ray scattering experiments, Gupta *et. al.* (2015) have shown that the fundamental unit of YG-box oligomers is a dimer, but atomic details of a potential oligomeric interface have so far been elusive. In this section, the crystal structure of SpSMN$^{\Delta 36\text{-}119}$ helix alone is presented at a resolution of 2.16 Å. For validation metrics and crystallographic data, see annexure (section 8). In addition to the glycine-zipper dimeric interface, this structure reveals atomic details of a previously unseen interface between dimers, entirely distinct from the dimeric interface, and is shown subsequently to be the oligomeric interface in solution by mutational analysis.

4.6.2 The YG-box glycine-zipper dimeric interface

The YG-box domain of SMN contains a consensus 3x(YxxG) motif (Figure 4.6 A). In the case of SpSMN, the third repeat is a LxxG instead of YxxG. As expected, SpSMN$^{\Delta 36\text{-}119}$ forms helical homodimers at the YG-box region (Figure 4.6 B). The individual helices are parallel to each other in that the N- and the C-termini of the monomers point to the same direction. Upstream of the YG-box, an intra-helix salt bridge is seen between Asp121 and Lys125, which may contribute towards the stability of each helical monomer (Figure 4.6 C). This was previously not seen in the MBP-SpYG-box dimer structure (Gupta et al. 2015) due to the strongly distorted conformation adopted by residues Tyr120 to Thr123 as a result of N-terminal MBP

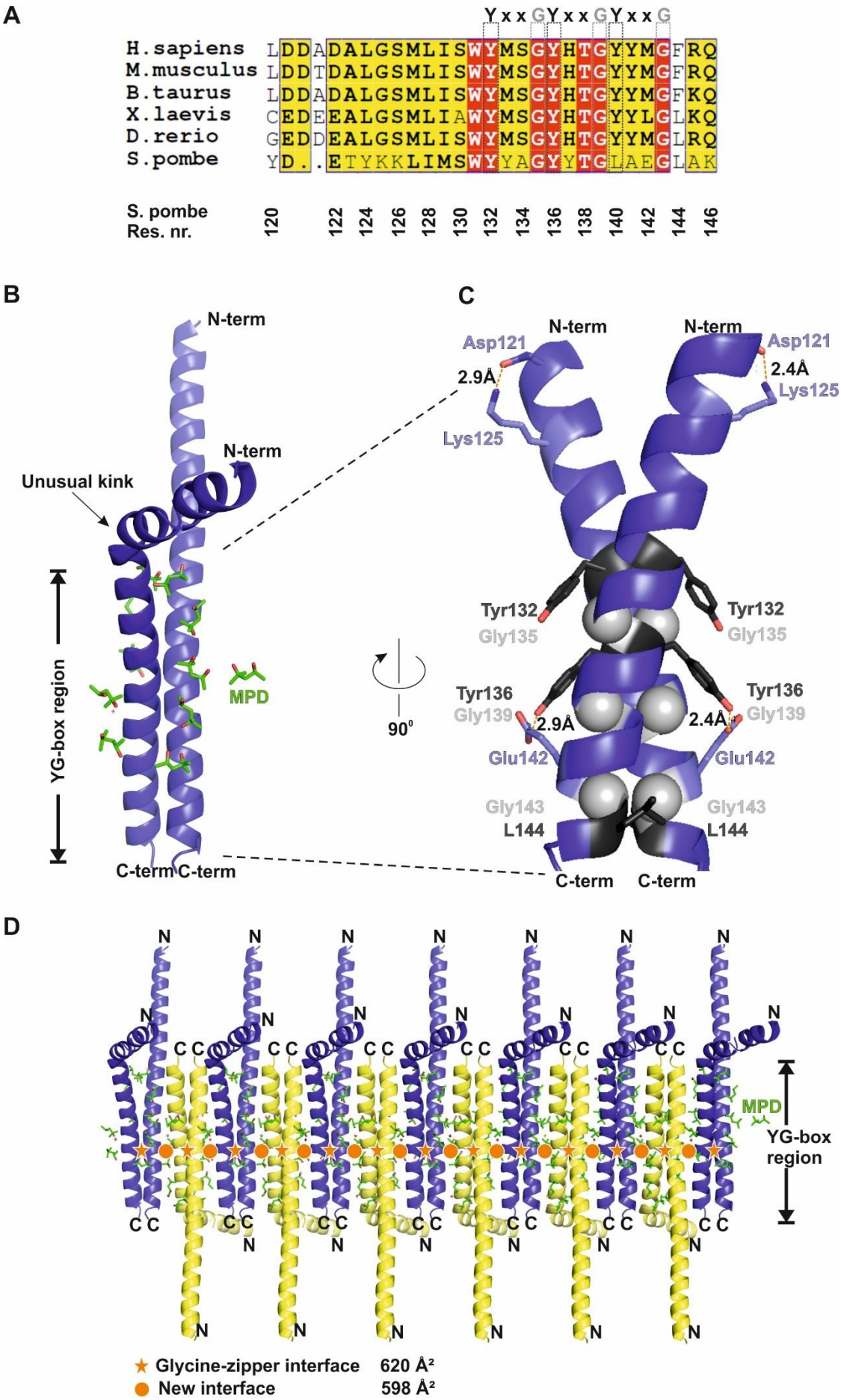

A
Y x x G Y x x G Y x x G
H.sapiens LDDADALGSMLISWYMSGYHTGYYMGFRQ
M.musculus LDDTDALGSMLISWYMSGYHTGYYMGFRQ
B.taurus LDDADALGSMLISWYMSGYHTGYYMGFKQ
X.laevis CEDEEALGSMLIAWYMSGYHTGYYLGLKQ
D.rerio GEDDEALGSMLISWYMSGYHTGYYMGLRQ
S.pombe YD..ETYKKLIMSWYYAGYYTGLAEGLAK
S. pombe Res. nr.
120 122 124 126 128 130 132 134 136 138 140 142 144 146
B
N-term
N-term
Unusual kink
MPD
YG-box region
90⁰
C-term C-term
C
N-term
N-term
Asp121
2.9Å
Lys125
Asp121
2.4Å
Lys125
Tyr132
Gly135
Tyr132
Gly135
Tyr136
Gly139
2.9Å
2.4Å
Tyr136
Gly139
Glu142
Glu142
Gly143
L144
Gly143
L144
C-term C-term
D
MPD
YG-box region
★ Glycine-zipper interface 620 Å²
● New interface 598 Å²

Figure 4.6: Structure of SpSMN$^{\Delta 36\text{-}119}$ and crystallographic packing
A: Alignment of YG-box sequences from different organisms showing the triple YxxG motif. B: Crystal structure of SpSMN$^{\Delta 36\text{-}119}$ dimer at 2.16 Å resolution, showing an unusual kink in one of the monomers and MPD molecules (=2-methyl-2,4-pentanediol) from crystallization condition (shown in green). C: Specific interactions within the YG-box region depicting residues involved in the glycine-zipper dimeric interface. Gly135, Gly139 and Gly143 pack tightly against the corresponding residues of the interfacing parallel monomer. Tyr132, Tyr136 and Leu144 pack tightly against the three Glycine residues of the interfacing monomer. Tyr136 is stabilized by Glu142 by H-bonding. A previously uncharacterized intra-helix salt bridge between Asp121 and Lys125 is depicted. D: Crystal-packing showing the crystallographic interface (New interface) between SpSMN$^{\Delta 36\text{-}119}$ anti-parallel glycine-zipper dimers at the YG-box region. Anti-parallel dimers are colored blue and yellow for clarity. Sequence alignment done using ESPript.

fusion. Within the YG-box, Gly135, Gly139 and Gly143 of one monomer pack tightly against the corresponding glycine residues of the interfacing parallel monomer. This forms a glycine-zipper homodimer (Figure 4.6 C). Aromatic sidechains of Tyr132, Tyr136 pack against Gly135 and Gly139, respectively. Each Tyr136 sidechain is stabilized by Glu142 through weak hydrogen bonding. Within the third YG-box repeat (LxxG in case of SpSMN), Gly143 does not make any close contacts with Leu140, instead, it is tightly packed against Leu144 sidechain which lies outside of the YG-box consensus motif (Figure 4.6 A & C). These observations further extend the YG-box helical region up to Tyr120, characterize a previously unseen salt bridge upstream of the YG-box between Asp121 and Lys125, and highlight a crucial role of Glu142 in stabilizing Tyr136. For the crystallographic parameters of the structure, refer to annexure Table 8.2.

4.6.3 Interaction surfaces engaged by SpSMN$^{\Delta 36\text{-}119}$ helix within the crystal

Analysis of the crystal structure by PISA server lists 6 major interaction surfaces engaged by SpSMN$^{\Delta 36\text{-}119}$ helix (Table 4.2). Interaction surface 1 is the well characterized glycine-zipper dimeric interface. Interaction surfaces 3-4 are between tightly packed N-terminal region of the helix (away from the YG-box region) and most likely to be bona fide crystal contacts only (not shown in Figures), as the N-terminus of SMN does not self-associate. The interaction surface 2 (termed New interface) is between dimers at the YG-box region and has a surface area of interaction comparable (598 Å^2) to that of the glycine-zipper dimeric interface (620 Å^2). In addition, the calculated solvation free energy gain (ΔG) is identical to that of the physiological glycine-zipper interface (-13 kcal/mol) and is significantly higher than that of all other interfaces. Further, the lower than 0.5 ΔG P-value (0.438) and remarkably similar complexation significance score (CSS, 0.130) to that of the glycine-zipper interface

(0.124), makes a strong case for further investigation of the physiological relevance of the New interface.

Table 4.2: PISA analysis of SpSMN$^{\Delta 36\text{-}119}$ crystal structure

#	Interaction surface	Area (Å^2)	ΔG (kcal/mol)	ΔG (P-value)	CSS
1	Glycine-zipper interface	620	-13.0	0.403	0.124
2	New interface	598	-13.0	0.438	0.130
3	Contact between N-termini	326	-2.9	0.706	0.000
4	Contact between N-termini	229	-5.8	0.291	0.005
5	Contact between N-termini	217	-5.9	0.374	0.000
6	Contact between N-termini	142	-1.4	0.725	0.000

4.6.4 Mutational analysis establishes New Interface as the oligomeric interface

Visualization of the crystal packing shows that, as a consequence of the New interface, glycine-zipper homodimers are stacked upon each other in an anti-parallel fashion precisely at the YG-box region (Figure 4.6 D). At the point of closest contact in this interface (Figure 4.7 B-C, orange box), Ser130 and Ala134 from one monomer pack tightly against Ala134 and Ser130 of the interfacing monomer, respectively. As shown in the sequence alignment (Figure 4.7 A), these two positions are almost always occupied by either Serine or Alanine, but never a bulkier residue. In order to test whether the New interface could be the physiological oligomeric interface, each of these residues were individually mutated to a residue with a bulkier sidechain (S130D and A134E) to introduce steric hindrance and prevent formation of oligomers larger than dimers (since these residues are situated away from the physiological glycine zipper dimeric interface, it is assumed that these mutations will not prevent YG-box dimerization). The molecular size and oligomeric states of the resulting mutant constructs SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36\text{-}119}$S130D and SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36\text{-}119}$A134E were compared with the wildtype construct by SEC-SAXS analysis (see methods section 3.2.3.3.6). As shown in Figure 4.7 D-E (black chromatograms and scatter plots), at relatively low concentrations, the wildtype construct exists as mixture of dimers and higher order oligomers. At high concentrations, it exists predominantly as higher oligomers and the dimeric peak completely disappears (although SEC-SAXS frames corresponding to dimeric molecular weight could be found at the tail of the peak). Both mutants (Figure 4.7 D-E, red and blue chromatograms, and scatter plots), however, -

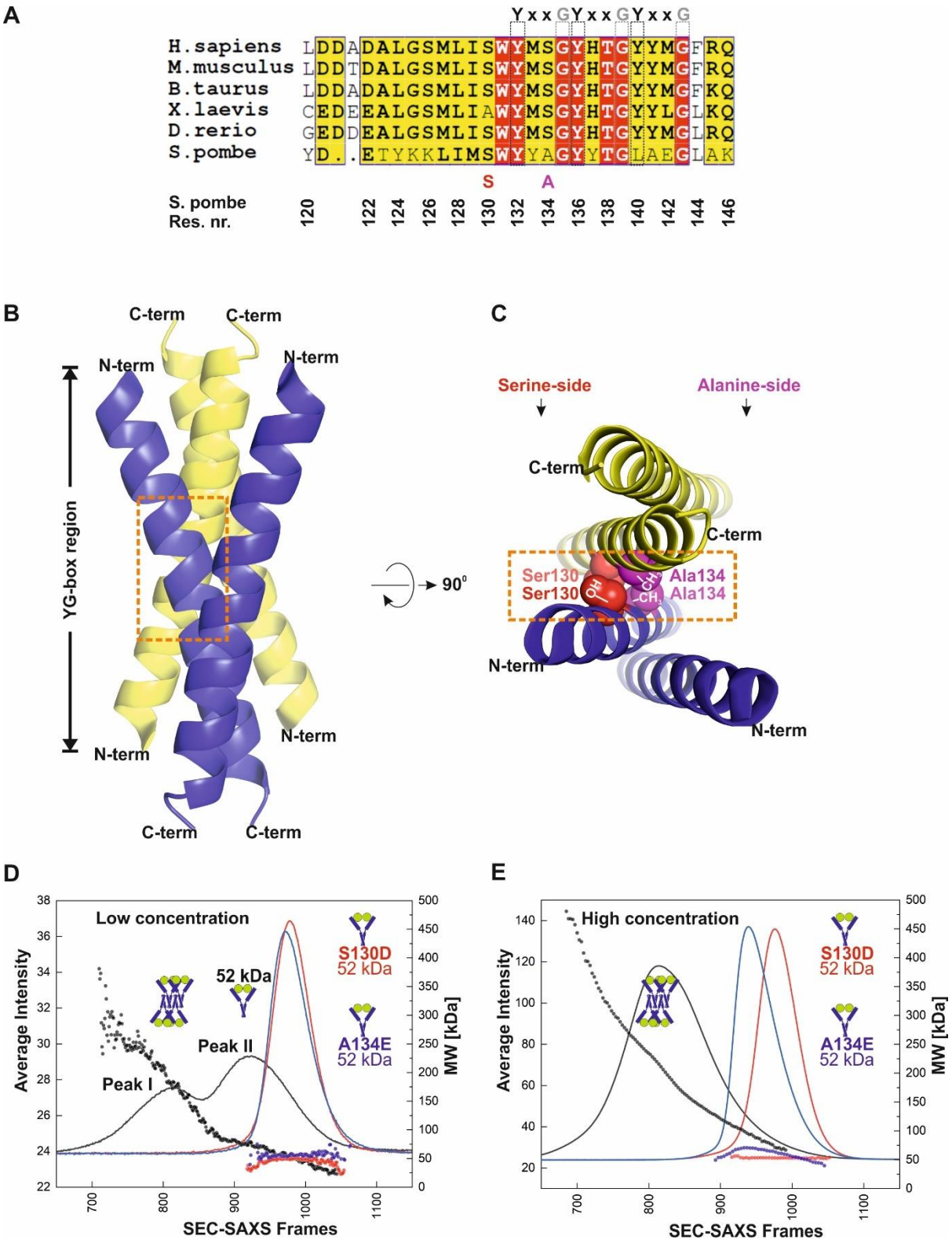


Figure 4.7: Mutational analysis of the crystallographic New interface

A: Alignment of YG-box sequences from different organisms showing the triple YxxG motif as well as Ser130 and Ala134 found in the New interface. B-C: Ser130 and Ala134 pack tightly against each other at the point of closest contact (orange box) in the crystallographic New interface. Both interfacing Serines are located on one side of the YG-box stack (Serine-side), and both interfacing Alanines on the other (Alanine-side). D-E: SEC-SAXS chromatograms of SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36\text{-}119}$ (black), SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36\text{-}119}$S130D (red) and SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36\text{-}119}$A134E (blue). The molecular weight distribution over the chromatogram is shown as scatter plot.

exist entirely as dimers at low as well as at high concentrations, showing no propensity to oligomerize. In addition, the calculated MWs from SEC-SAXS frames are remarkably accurate to the expected MWs of the dimeric forms of the mutants (52.5 kDa) This clearly demonstrates the effectiveness of these mutations to prevent formation of oligomers larger than dimers in solution. This also shows that the mutations do not interfere in YG-box dimerization. Together, these experiments establish the New interface as the previously unseen oligomeric interface for SpSMN's YG-box dimers in solution.

4.6.5 The YG-box anti-parallel oligomeric interface

As demonstrated in the previous section, the fundamental unit of SpSMN oligomers are parallel glycine zipper dimers of YG-box, which stack upon each other in an anti-parallel fashion to form higher order oligomers. The specific interactions of the oligomeric interface are entirely different from those of the dimeric interface. The tight packing of Ser130 and Ala134 against Ala134 and Ser130, respectively, of the interfacing anti-parallel monomer occurs through mainchain atoms such that the sidechains (-CH_2-OH and -CH_3) are oriented away from the oligomeric interface pointing towards opposite faces of the oligomeric stack (Figure 4.7 C). These interactions between reciprocal residues of the interfacing monomers are such that the reciprocal Serines are situated on one side of the stack (Serine-side) and the reciprocal Alanines on the other (Alanine-side) (Figure 4.7 C). On the Alanine-side, each methyl group is oriented precisely below the benzene ring of Trp131 of the interfacing monomer (Figure 4.8 B). The distance from the methyl group carbon atom to the benzene ring of Trp131 is between 4-5 Å, strongly suggesting that such arrangements of sidechains is to facilitate CH-π interactions within each Trp131:Ala134 pair. Furthermore, this specific conformation of Trp131 sidechains is stabilized by hydrogen bonding to the Thr138 sidechains of the reciprocal monomer. As a consequence of these interactions, both Ala134 residues are completely buried (100%) within the oligomeric interface (Figure 4.8 B). The Serine-side, however, exhibits no residue specific interactions and the Ser130 residues are only partially buried (70%) within the oligomeric interface (Figure 4.8 C).

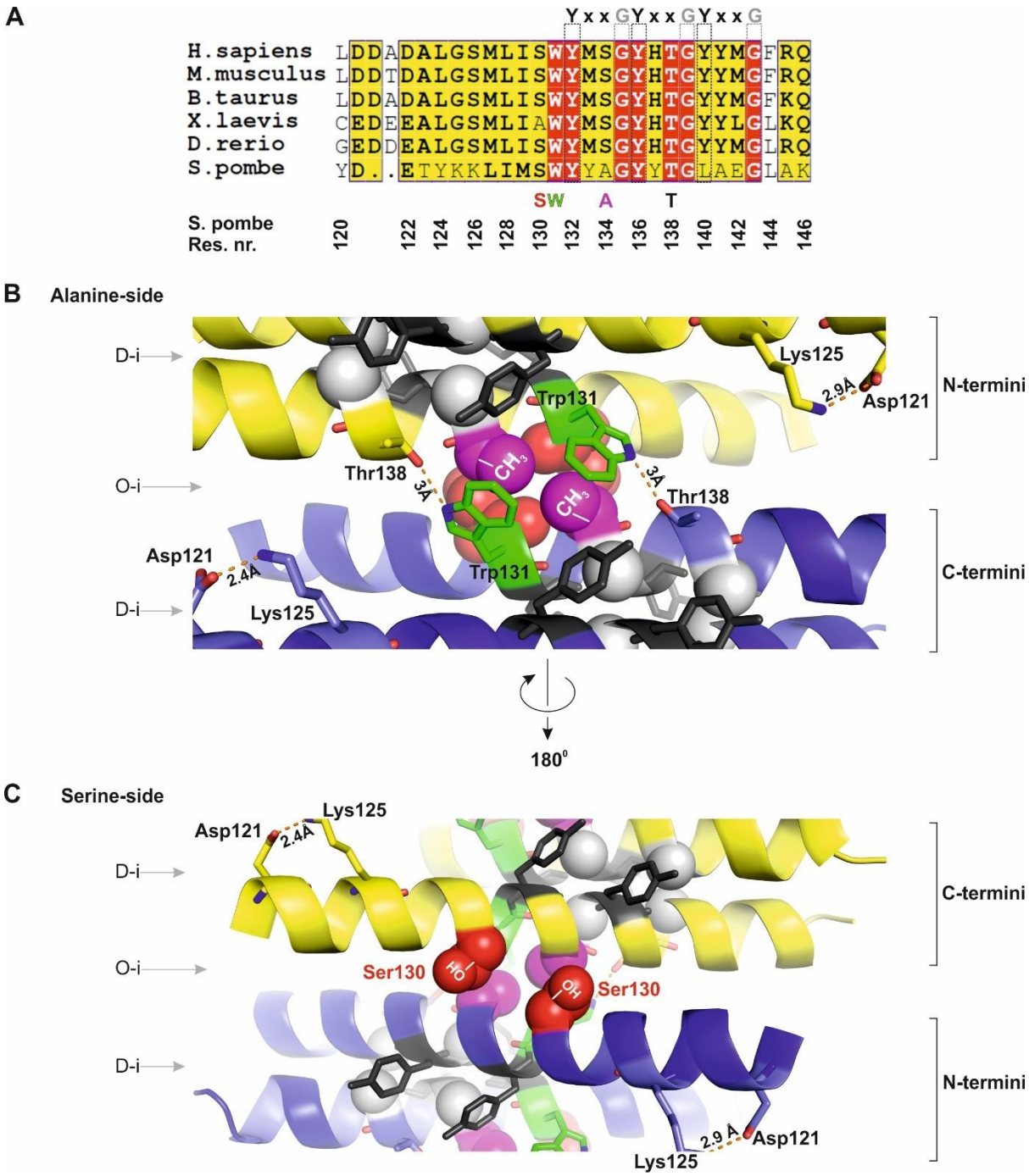


Figure 4.8: Characteristics of the YG-box oligomeric interface

A: Alignment of YG-box sequences from different organisms showing the triple YxxG motif, Ser130, Trp131, Ala134 and Thr138. B: The Alanine-side of the oligomeric interface showing residue specific interactions. Ser130 and Ala134 of one monomer packs tightly against Ala134 and Ser130 of the interfacing monomer, respectively. Trp131 is stabilized by interfacing Thr138 through H-bonding. C: The Serine-side of the oligomeric interface shows no residue specific interactions. D-i refers to dimeric interface and O-i to oligomeric interface.

Finally, since the dimers are stacked in an anti-parallel fashion, the Serine- and Alanine-sides alternate between consecutive oligomeric interfaces. Together, these observations provide a comprehensive understanding of the hitherto unseen oligomeric interface of SpSMN and highlight the crucial roles of Ser130, Trp131, Ala134 and Thr138 residues of the C-terminal YG-box domain. An interaction scheme of an octameric YG-box stack is shown in Figure 4.9.

Figure 4.9: Interaction scheme of an octameric SpSMN YG-box stack.
YG-box monomers within the glycine-zipper dimers (yellow or blue dimer) are parallel. At the dimeric interface (D-i), the tight packing between Glycine residues (grey) are shown by ↔ and the packing of Tyrosines/Leucines sidechains against Glycines is denoted by −•. In an oligomer, dimers are anti-parallel with respect to each other. At the oligomeric interface (O-i), tight packing of Serines against Alanines are denoted by ↔ and the packing of Tryptophan sidechains against Alanines is denoted by −•. Hydrogen bonds between Tryptophan sidechains and Threonines is denoted by orange dashed lines.

4.7 SpSMN's C-terminal YG-box domain serves as an interaction platform for SpG8's N-terminus

4.7.1 Introductory notes

Within the human SMN complex, through comprehensive subunit interaction mapping, it has been shown that G8 exclusively interacts with SMN on one side, and G7 on the other. Accordingly, as demonstrated in section 4.3, the SpG8 forms the link between SpG2•SpSMN and SpG7•SpG6. To map the interacting regions between SpSMN and SpG8, mutants of SpSMN and SpG8 were designed and the interaction was assessed through expression tests and gelfiltration.

4.7.2 YG-box domain is essential for soluble expression of SpG8

His-SpG8 alone was expressed in *E. coli* but formed inclusion bodies (misfolded protein, Figure 4.10 A, pellet). The soluble fraction was highly unstable (elution) as indicated by degradation products. Expression test of His-SpG2•SpSMN$^{\Delta YG}$•SpG8 also revealed that SpG8 is highly expressed as insoluble protein (Figure 4.10 B, pellet). However, soluble SpG8 could be co-purified in complex with SpG2•SpSMN (full-length), indicating that YG-box may serve as the interaction platform for SpG8 and facilitate correct folding.

4.7.3 SpG8ΔN58 failed to interact with SpG2•SpSMN

The predicted secondary structure composition of SpG8 shows presence of three helical regions, the N-terminus, the central helical region, and the C-terminus. A tricistronic construct lacking the N-terminal helix of SpG8 (SpG8ΔN58•His-SpG7•SpG6) was expressed in *E. coli* and purified by Ni-NTA chromatography (data not shown). The purified trimeric complex and SpG2•SpSMN were mixed in equimolar amounts and applied onto a Superose 6 10/300 analytical gelfiltration column. As shown in Figure 4.10 D, the sub-complexes did not interact and separated into two distinct peaks, showing that the interaction depends on the SpG8's N-terminal helical region. To test this possibility, SpG8 residues 3-34 was cloned as an N-terminal His-SUMO-tagged construct, expressed in *E. coli,* and purified by Ni-NTA. SUMO-SpG8$^{3-34}$ was mixed in 5-fold molar excess with SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36-119}$ which contains only the SpG2 binding domain and the YG-box. The complex formation was then analyzed by gelfiltration.

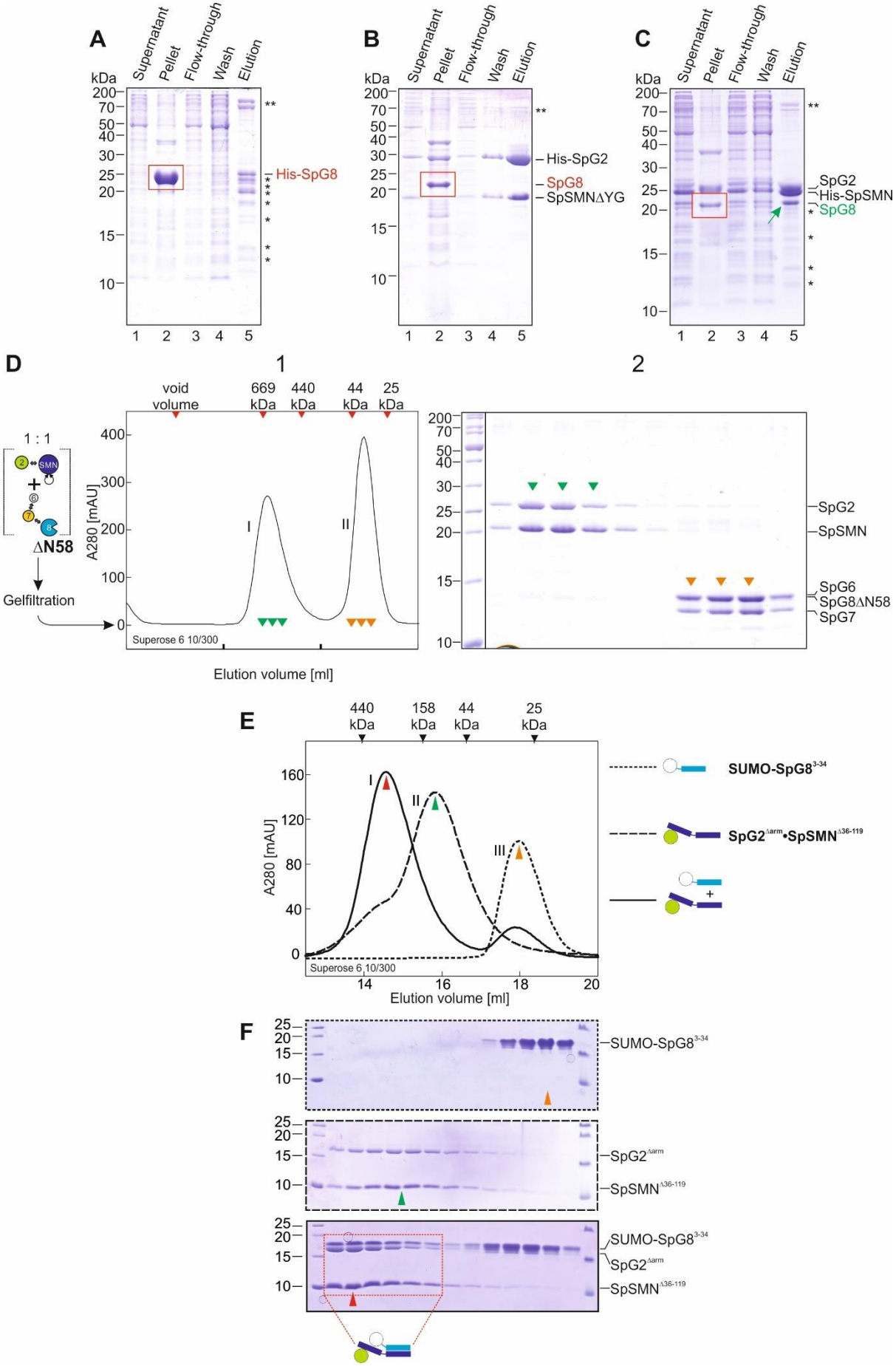

A
B
C
Supernatant
Pellet
Flow-through
Wash
Elution
kDa
His-SpG8
His-SpG2
SpG8
SpSMNΔYG
SpG2
His-SpSMN
SpG8
D
1
2
void volume
669 kDa
440 kDa
44 kDa
25 kDa
1 : 1
SMN
ΔN58
Gelfiltration
A280 [mAU]
I
II
Superose 6 10/300
Elution volume [ml]
SpG2
SpSMN
SpG6
SpG8ΔN58
SpG7
E
440 kDa
158 kDa
44 kDa
25 kDa
III
SUMO-SpG8^3-34
SpG2^Δarm•SpSMN^Δ36-119
F
SpG2^Δarm
SpSMN^Δ36-119

Figure 4.10: Interaction between YG-box and SpG8 N-terminus
Expression tests of His-SpG8 (A), His-SpG2•SpSMN$^{\Delta YG}$•SpG8 (B) and His-SpG2•HisSpSMN•SpG8 (C). Insoluble fraction of SpG8 is indicated by a red box. * correspond to possible degradation products from soluble His-SpG8 in the elution lane (A and C). Co-purified soluble SpG8 is shown with green arrow in C. Gelfiltration chromatogram of equimolar mix of (SpG2•SpSMN + SpG8ΔN58•SpG7•SpG6) from Superose 6 10/300 column (D, panel 1), and analysis of fractions by SDS-PAGE (panel 2). Three central fractions within each peak is show by green and orange arrowheads. Gelfiltration chromatograms of SUMO-SpG8$^{3-34}$ (E, dotted line), SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36-119}$ (E, dashed line) and mix of (5x SUMO-SpG8$^{3-34}$ + SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36-119}$) (E, solid line) from Superose 6 10/300 column. Analysis of fractions by SDS-PAGE is shown in F. Positions of peak fractions are indicated by orange, green and red arrowheads.

The individual proteins SUMO-SpG8$^{3-34}$ and SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36-119}$ elute near 25 kDa and 158 kDa markers (Figure 4.10 E-F, dotted and dashed chromatograms, respectively). The mixture, however, exhibits significantly shifted elution profile (solid chromatogram) and the trimeric complex SUMO-SpG8$^{3-34}$•SpG2$^{\Delta arm}$•SpSMN$^{\Delta 36-119}$ could be easily identified by SDS-PAGE within the first peak. The peak fractions in the SDS-PAGE are indicated by arrows. Since the SMN$^{\Delta 36-119}$ construct contains only the N-terminal SpG2 binding domain and the YG-box, these experiments show that SpG8 N-terminus (residues 3-34) directly bind to the YG-box of SpSMN.

4.8 Characterization of SpSMN linker (residues 36-119) by small angle X-ray scattering

4.8.1 Introductory notes

In the previous sections, the oligomeric states of SpSMN were determined and a crystal structure was presented describing the structural basis of YG-box oligomerization. These results provide a basis for the observed large hydrodynamic behaviors of SpSMN containing complexes. The second major feature of SpSMN that contributes to the large hydrodynamic property of SpSMN complexes is the central unstructured linker spanning from residue 36 to residue 119. To study its properties and its contribution to the overall size, shape, and flexibility of SpSMN containing complexes, we prepared samples with or without the linker, with or without SpG8•SpG7•SpG6, and collected small angle X-ray solution scattering data. The N-terminal extended arm of SpG2 (residues 1-80, arm) and an internal unstructured loop within SpG8 (residues 35-58, loop) were excluded from the analyzed complexes in order to focus exclusively on the unstructured linker of SpSMN.

Standard constructs:

1. SpG2$^{\Delta arm}$•**SpSMN**$^{\Delta 36\text{-}119}$A134E (monomeric MW 26.25 kDa)
2. SpG2$^{\Delta arm}$•**SpSMN**$^{\Delta 36\text{-}119}$S130D (monomeric MW 26.25 kDa)

SpSMN$^{\Delta 36\text{-}119}$ complexes:

3. SpG2$^{\Delta arm}$•**SpSMN**$^{\Delta 36\text{-}119}$ (monomeric MW 26.25 kDa)
4. SpG2$^{\Delta arm}$•**SpSMN**$^{\Delta 36\text{-}119}$•SpG8$^{\Delta loop}$•SpG7•SpG6 (monomeric MW 64.16 kDa)

SpSMN-FL complexes:

5. SpG2$^{\Delta arm}$•**SpSMN-FL** (monomeric MW 35.47 kDa)
6. SpG2$^{\Delta arm}$•**SpSMN-FL**•SpG8$^{\Delta loop}$•SpG7•SpG6 (monomeric MW 73.45 kDa)

Data collection was performed by SEC-SAXS strategy (see methods section 3.2.3.3.6) where the complexes were exposed to X-ray beam as they eluted from a Superdex 200 10/300 gelfiltration column. Data analysis was performed by ATSAS 3.0 software package. For all data interpretation and SAXS terminologies, please refer to methods section 3.2.3.3. For data collection parameters refer to annexure Table 8.3.

4.8.2 SEC-SAXS: Standard constructs are monodisperse dimers

Due to mutations that prevent YG-box oligomerization but allow a dimer formation, and the deletion of the unstructured region of SpSMN (residues 36-119), SpG2$^{\Delta arm}$•**SpSMN**$^{\Delta 36\text{-}119}$A134E and SpG2$^{\Delta arm}$•**SpSMN**$^{\Delta 36\text{-}119}$S130D are expected to be monodisperse dimeric samples (52.5 kDa) and exhibit globular nature. As shown in the chromatograms (Figure 4.11 A), both constructs exhibit the expected molecular weight distributions over the elution profile (scatter plots). Frames at the peak of each chromatogram were selected (blue and red circles) and final SAXS curves generated after averaging and buffer subtraction (B). Using Guinier analysis, radius of gyration (Rg) and the zeroth angle intensities I(0) were determined for each construct. The calculated molecular weights for each construct by three different methods (from Porod Invariant Qp, from MoW tool and from Volume of Correlation Vc) appear remarkably close to the expected molecular weights (B). Hence, these complexes were later used as standards for calculation of molecular weights from I(0) (see section 4.8.8).

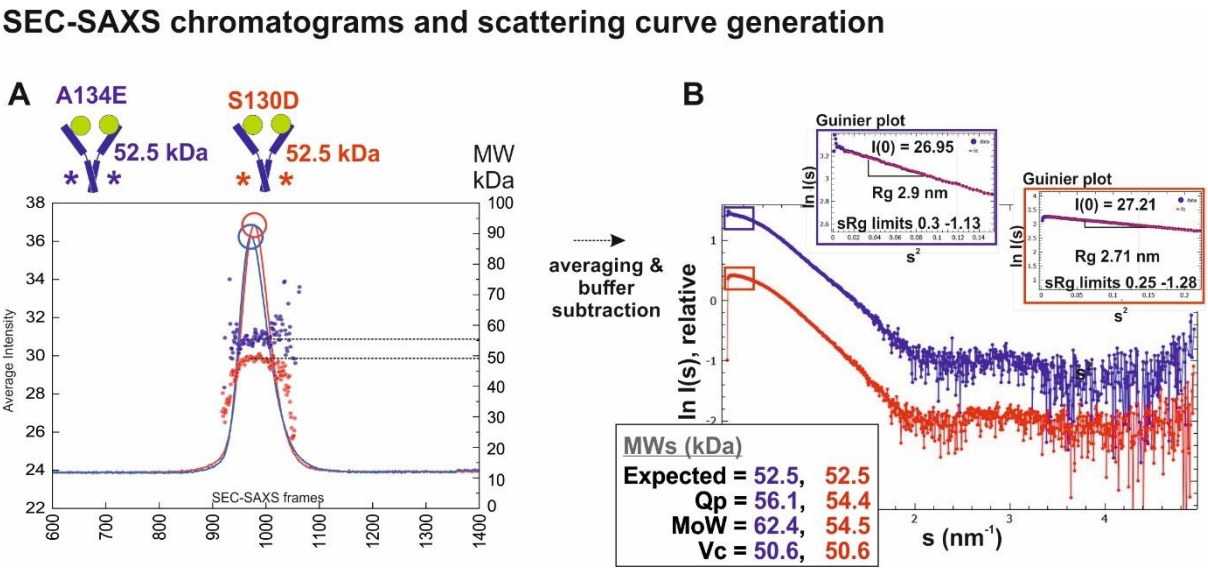


Figure 4.11: SEC-SAXS chromatograms and scattering curve generation of standards

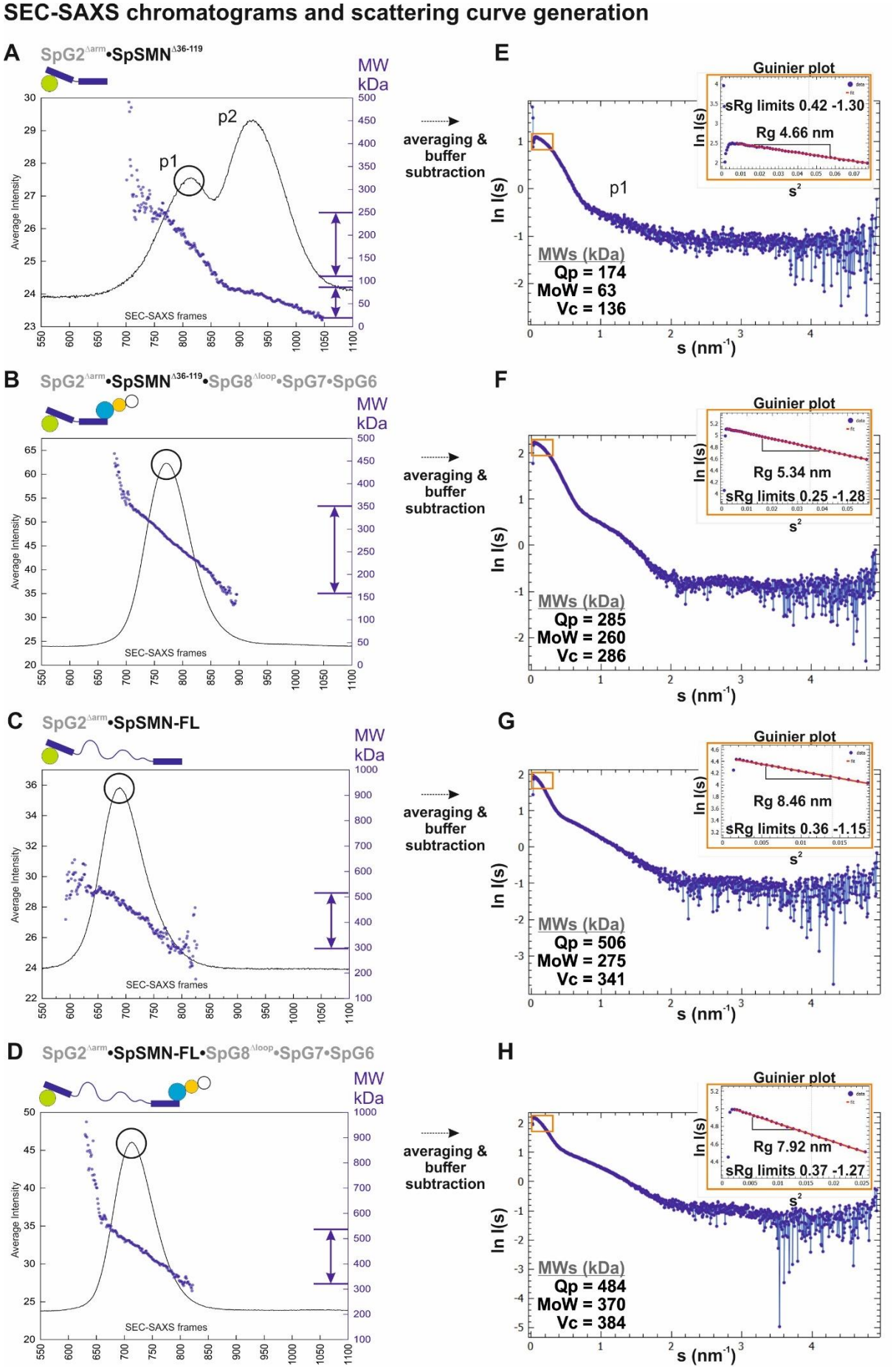
A: SEC-SAXS frames of standard constructs SpG2$^{\Delta arm}$•**SpSMN**$^{\Delta 36\text{-}119}$A134E and SpG2$^{\Delta arm}$•**SpSMN**$^{\Delta 36\text{-}119}$S130D showing MWs calculated for each frame (from porod invariant Qp) as scatter plot. Frames at the peak, indicated by circles, were averaged and background subtracted to generate final scattering curves depicted as ln I(s) vs s (B, blue and red plots). Quality was assessed by the linearity of Guinier region. Upper sRg limits were maintained below 1.3 and Rg calculated from slope of linear fit (blue and orange insets). From each final curve, MWs calculated using Porod invariant Qp, MoW tool, and volume of correlation Vc are shown in inset below the curves.

4.8.3 SEC-SAXS: SpSMN$^{\Delta 36\text{-}119}$ & SpSMN-FL complexes appear as polydisperse oligomers

SpG2$^{\Delta arm}$•**SpSMN**$^{\Delta 36\text{-}119}$ and SpG2$^{\Delta arm}$•**SpSMN**$^{\Delta 36\text{-}119}$•SpG8$^{\Delta loop}$•SpG7•SpG6 lack the unstructured region of SpSMN, and are expected to be relatively globular, whereas SpG2$^{\Delta arm}$•**SpSMN-FL** and SpG2$^{\Delta arm}$•**SpSMN-FL**•SpG8$^{\Delta loop}$•SpG7•SpG6 are expected to exhibit large apparent MW due to the unstructured region of SpSMN. All 4 constructs are expected to form higher order oligomers (larger than dimers) due to the wild-type YG-box sequence. As shown in the chromatograms (Figure 4.12, A-D), all 4 constructs appear to be polydisperse samples with a broad range of MW distributions over the elution profile. Of note, SpG2$^{\Delta arm}$•**SpSMN**$^{\Delta 36\text{-}119}$ (A) exists predominantly as dimers (p2) as well as higher order oligomers (p1). Frames at the peak of each chromatogram were selected (black circles) and final SAXS curves generated after averaging and buffer subtraction (Figure 4.12, E-H). From the slope of Guinier plot, radii of gyration were determined for each construct. Molecular weights for each construct were calculated by three different methods (from Porod Invariant Qp, from MoW tool and from Volume of Correlation Vc). Two important observations could be made based on these analyses. First, variability between the molecular weights obtained from the 3 methods is relatively large for SpG2$^{\Delta arm}$•**SpSMN**$^{\Delta 36\text{-}119}$ and SpG2$^{\Delta arm}$•**SpSMN-FL** than for the SpG8$^{\Delta loop}$•SpG7•SpG6 containing complexes. Second, in the presence of SpG8$^{\Delta loop}$•SpG7•SpG6, the radius of gyration Rg is higher for **SpSMN**$^{\Delta 36\text{-}119}$ complexes (5.34 nm) but lower for **SpSMN-FL** complexes (7.92 nm). This suggests a possible role of SpG8$^{\Delta loop}$•SpG7•SpG6 in controlling the biophysical properties of the whole complex, to a certain degree.

Figure 4.12: SEC-SAXS chromatograms and scattering curve generation of samples

SEC-SAXS frames for SpSMN$^{\Delta 36\text{-}119}$ complexes (A-B) and SpSMN-FL complexes (C-D) are shown on the left panels. The molecular weight calculated (from porod invariant Qp) for each scattering frame is shown as blue scatter plots. Upper and lower limits of MW distribution within the chromatogram are indicated with arrows on the right Y-axis. Selected data area for final scattering curve generation is shown as black circles. The final scattering curves after buffer subtraction are shown on the right panels as ln(s) vs s (blue plots). Quality of the generated scattering curves was assessed by the Guinier region linearity (orange inset). For determination of Rg for each curve, the upper sRg limit was maintained below 1.3. Using the scattering curves, MWs were calculated by three different methods: from the porod invariant Qp, from MoW tool, and from volume of correlation Vc.


SEC-SAXS chromatograms and scattering curve generation
A
SpG2Δarm•SpSMNΔ36-119
MW kDa
p1
p2
Average Intensity
SEC-SAXS frames
averaging & buffer subtraction
E
Guinier plot
sRg limits 0.42 -1.30
Rg 4.66 nm
ln I(s)
s²
p1
MWs (kDa)
Qp = 174
MoW = 63
Vc = 136
s (nm⁻¹)
B
SpG2Δarm•SpSMNΔ36-119•SpG8Δloop•SpG7•SpG6
MW kDa
Average Intensity
SEC-SAXS frames
averaging & buffer subtraction
F
Guinier plot
Rg 5.34 nm
sRg limits 0.25 -1.28
ln I(s)
s²
MWs (kDa)
Qp = 285
MoW = 260
Vc = 286
s (nm⁻¹)
C
SpG2Δarm•SpSMN-FL
MW kDa
Average Intensity
SEC-SAXS frames
averaging & buffer subtraction
G
Guinier plot
Rg 8.46 nm
sRg limits 0.36 -1.15
ln I(s)
s²
MWs (kDa)
Qp = 506
MoW = 275
Vc = 341
s (nm⁻¹)
D
SpG2Δarm•SpSMN-FL•SpG8Δloop•SpG7•SpG6
MW kDa
Average Intensity
SEC-SAXS frames
averaging & buffer subtraction
H
Guinier plot
Rg 7.92 nm
sRg limits 0.37 -1.27
ln I(s)
s²
MWs (kDa)
Qp = 484
MoW = 370
Vc = 384
s (nm⁻¹)

4.8.4 Properties of SpSMN-FL complexes: Guinier analysis predicts intrinsic disorder

As discussed earlier, according to the Guinier approximation $I(s) = I(0)\, e^{\wedge}(- (sRg)^2/3)$, SAXS data transformed as $\ln I(s)$ vs s^2 is linear up to a maximum limit of the term $sRg<1.3$. Hence, particles with larger Rg satisfy this criterion up to smaller values of the angular momentum s and vice versa. Intrinsically disordered/highly flexible proteins exhibit large radius of gyration (Rg) compared to globular and compact proteins of similar molecular weights and are restricted to smaller values of the angular momentum s. SpG2$^{\Delta arm}$•**SpSMN**$^{\Delta 36\text{-}119}$•SpG8$^{\Delta loop}$•SpG7•SpG6 (260-286 kDa, Figure 4.12 F) and Catalase (240 kDa) are globular complexes which follow Guinier law up to angular ranges 0.245 and 0.317 nm^{-1}, respectively (Figure 4.13). SpG2$^{\Delta arm}$•**SpSMN-FL** and SpG2$^{\Delta arm}$•**SpSMN-FL**•SpG8$^{\Delta loop}$•SpG7•SpG6, which are of comparable molecular weights to the previous two (~228 and ~295 kDa respectively, see Figure 4.16 D), follow Guinier law only up to angular ranges of 0.155 and 0.164 nm^{-1}, respectively. This behavior is highly indicative of intrinsic disorder and illustrates that SpSMN containing complexes behave as intrinsically disordered proteins.

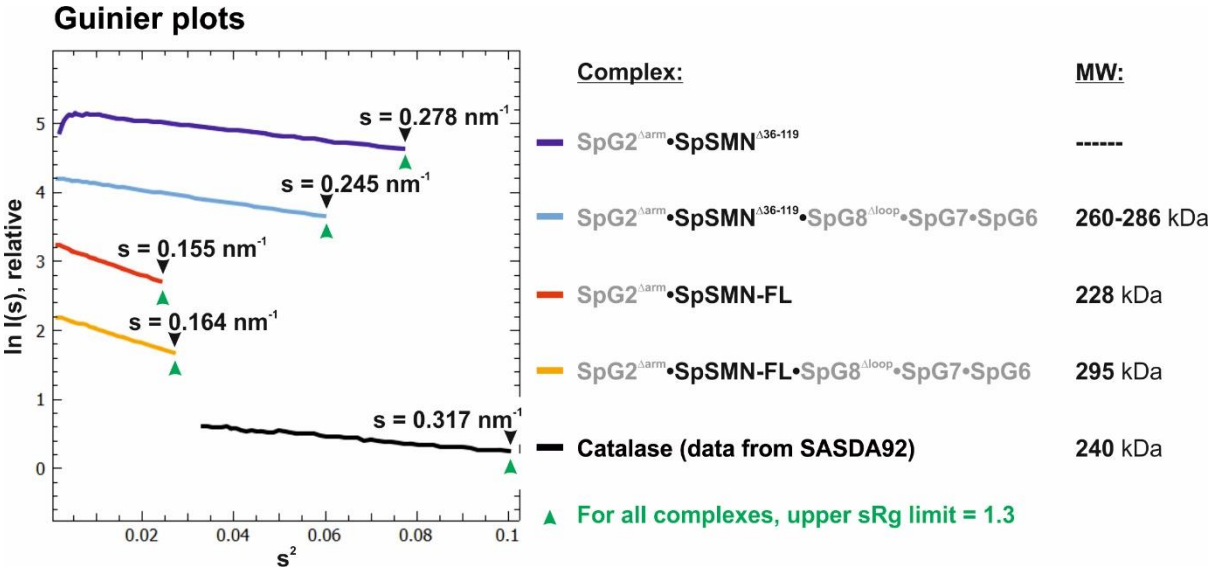


Figure 4.13: Maximum angular ranges (s) of the linear fit within Guinier region.
The Guinier plots, ln I(s) vs s^2, of all complexes are depicted on a relative scale. For all complexes, the upper sRg limit was fixed at 1.3 (green arrowhead). From a linear fit, the Rg was calculated and the maximum angular range (s) for the linear region was determined (black arrowhead). The SpSMN-FL complexes (red and orange) satisfy Guinier linearity up to a limited angular (s) range (0.155 nm^{-1} and 0.164 nm^{-1}) compared to SpSMNΔ36-119 complexes of comparable molecular weight (dark blue and light blue, 0.278 nm^{-1} and 0.245 nm^{-1}, respectively). For comparison, the Guinier plot and angular range of a highly globular protein of similar size (Catalase, 240 kDa) is shown in black (data taken from SASBDB: SASDA92).

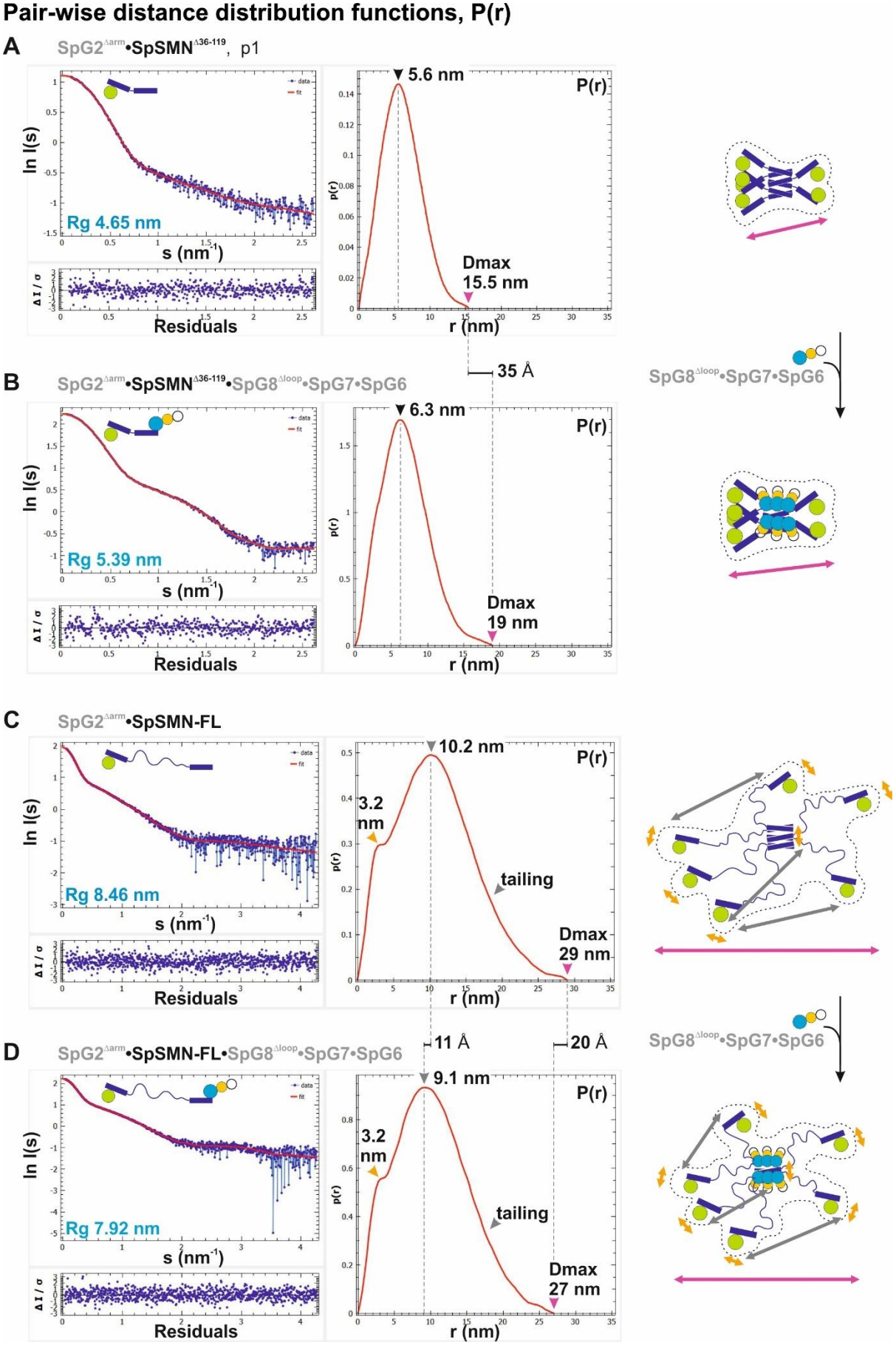
4.8.5 Properties of SpSMN-FL complexes: P(r) function reveals multidomain architecture, IDP-like extended conformations

Pair-wise distance distribution functions derived from SAXS curves can reveal the domain architectures of macromolecules. Globular proteins typically yield a symmetric, gaussian P(r) curve with a single peak, ending smoothly in a concave curvature at a specific Dmax (maximum dimension of the particle). Multidomain proteins with flexible linkers yield asymmetric curves with multiple peaks. Extended tailing is an indication of flexibility arising from large number of conformations. P(r) curves of SpSMN$^{\Delta 36\text{-}119}$ and SpSMN-FL complexes were derived from their scattering data and are shown in Figures 4.14 A-D, middle panel. Fit of the generated P(r) curves to the experimental data as well as the residual plots are shown in the left panels. As expected, P(r) curves of both SpSMN$^{\Delta 36\text{-}119}$ complexes display characteristics of globular proteins such as a single peak and nearly gaussian curve (Figures 4.14 A-B). The P(r) curves of SpG2$^{\Delta arm}$•**SpSMN-FL** and SpG2$^{\Delta arm}$•**SpSMN-FL**•SpG8$^{\Delta loop}$•SpG7•SpG6 (Figure 4.14 C-D), however, are asymmetric, exhibit multiple peaks typical for multidomain proteins connected with linkers, and display an extended tail ending in large Dmax values (29 nm and 27 nm respectively, pink arrows) typical for IDPs adopting large number of conformations. The first peak in the P(r) curves (3.2 nm) can be attributed to globular pair-wise distances within SpG2$^{\Delta arm}$ and YG-box domains (orange arrows). The second peak in the curves (10.24 nm and 9.1 nm, respectively) can be attributed to pair-wise distances between globular domains (grey arrows). It must be noted, that while the initial peak (3.2 nm) can be assumed to be a distinctive value arising from the globular domains, the second peak and Dmax values are average over many conformations. Interestingly, for SpSMN$^{\Delta 36\text{-}119}$ complexes (Figure 4.14 A-B), all physical parameters such as Rg, peak of P(r) and Dmax, are higher in the presence of SpG8$^{\Delta loop}$•SpG7•SpG6, but for SpSMN-FL complexes (Figure 4.14 C-D) all three parameters decrease in the presence of SpG8$^{\Delta loop}$•SpG7•SpG6.

Figure 4.14: Pairwise distance distribution functions, P(r)

The P(r) functions of SpSMN$^{\Delta 36\text{-}119}$-complexes (A-B) and SpSMN-FL -complexes (C-D) are shown as red curves (middle panel). The fit of the P(r) functions to the experimental scattering plots (blue plots) are shown on the left panels. The goodness of fit can be assessed by the residual plots under each curve. The indicated Rg values for each complex are calculated from Guinier analysis. A representative cartoon of each complex is shown on the right panels. While the SpSMN$^{\Delta 36\text{-}119}$-complexes show a gaussian P(r) function (indicative of globularity) with a single peak (black arrow-head), the SpSMN-FL complexes show multiple peaks and an extended tail ending in high Dmax values (magenta). For-

Pair-wise distance distribution functions, P(r)



-SpSMN-FL complexes, the difference in Dmax and P(r) peak values between complexes is shown by dotted grey lines (in angstroms). While for SpSMN$^{\Delta 36\text{-}119}$-complexes the values for Rg, Dmax and P(r) peak increase upon the addition of SpG8•SpG7•SpG6, all these parameters decrease for SpSMN-FL-complexes. The initial peak at 3.2 nm remains unchanged (orange arrowhead) for both complexes which might be attributed to the globular SpG2 domains and the YG-box bundle.

4.8.6 Properties of SpSMN-FL complexes: Dimensionless Kratky plot exhibits dual behavior, qualifies residues 36-119 as unstructured & flexible

Dimensionless Kratky plots provide both qualitative and quantitative estimation of macromolecular flexibility. Since these plots are normalized for macromolecular size (by multiplying data with radius of gyration Rg) and concentration (by dividing data with zeroth angle intensity I(0)), they allow for a quantitative comparison of the degree of flexibility between macromolecules of very different sizes. Dimensionless Kratky plots of globular proteins (discrete electron density contrast) yield a bell-shaped curve with a distinct maximum at (√3, 1.104), and eventual decay of signal to zero at higher angles. Fully unfolded proteins do not show a distinct maximum and the signal rises continuously at higher angles due to diffuse electron density contrast arising from flexibility. Multidomain proteins connected with flexible linkers, however, exhibit dual behaviour. All **SpSMN$^{\Delta 36\text{-}119}$** complexes exhibit a bell-shaped curve with a well-defined maximum at (√3, 1.104) typical for globular proteins (Figure 4.15 B-C). SpG2$^{\Delta arm}$•**SpSMN-FL** and SpG2$^{\Delta arm}$•**SpSMN-FL**•SpG8$^{\Delta loop}$•SpG7•SpG6, however, exhibit dual behavior with an initial peak at (√3, 1.104) followed by rise in signal (indicator of disorder and flexibility) between sRg=3.5 - 7, and eventual decay at higher angles (Figure 4.15 A). These observations qualify residues 36-119 to be the dominant contributor towards the overall flexibility of **SpSMN-FL** complexes. The relatively faster decay of signal at higher angles (beyond sRg 3.5) for complexes with SpG8$^{\Delta loop}$•SpG7•SpG6 suggests a reduced flexibility in the presence of SpG8$^{\Delta loop}$•SpG7•SpG6.

4.8.7 Properties of SpSMN-FL complexes: SpG8$^{\Delta loop}$•SpG7•SpG6 induces molecular compaction and mitigates flexibility

Information obtained from Guinier analyses, P(r) functions, and Dimensionless Kratky plots in the previous sections, strongly indicate that SpG2$^{\Delta arm}$•**SpSMN-FL** undergoes molecular compaction and gains some degree of rigidity upon binding to SpG8$^{\Delta loop}$•SpG7•SpG6 (refer to sections 4.8.4, 4.8.5 and 4.8.6). First, the Guinier

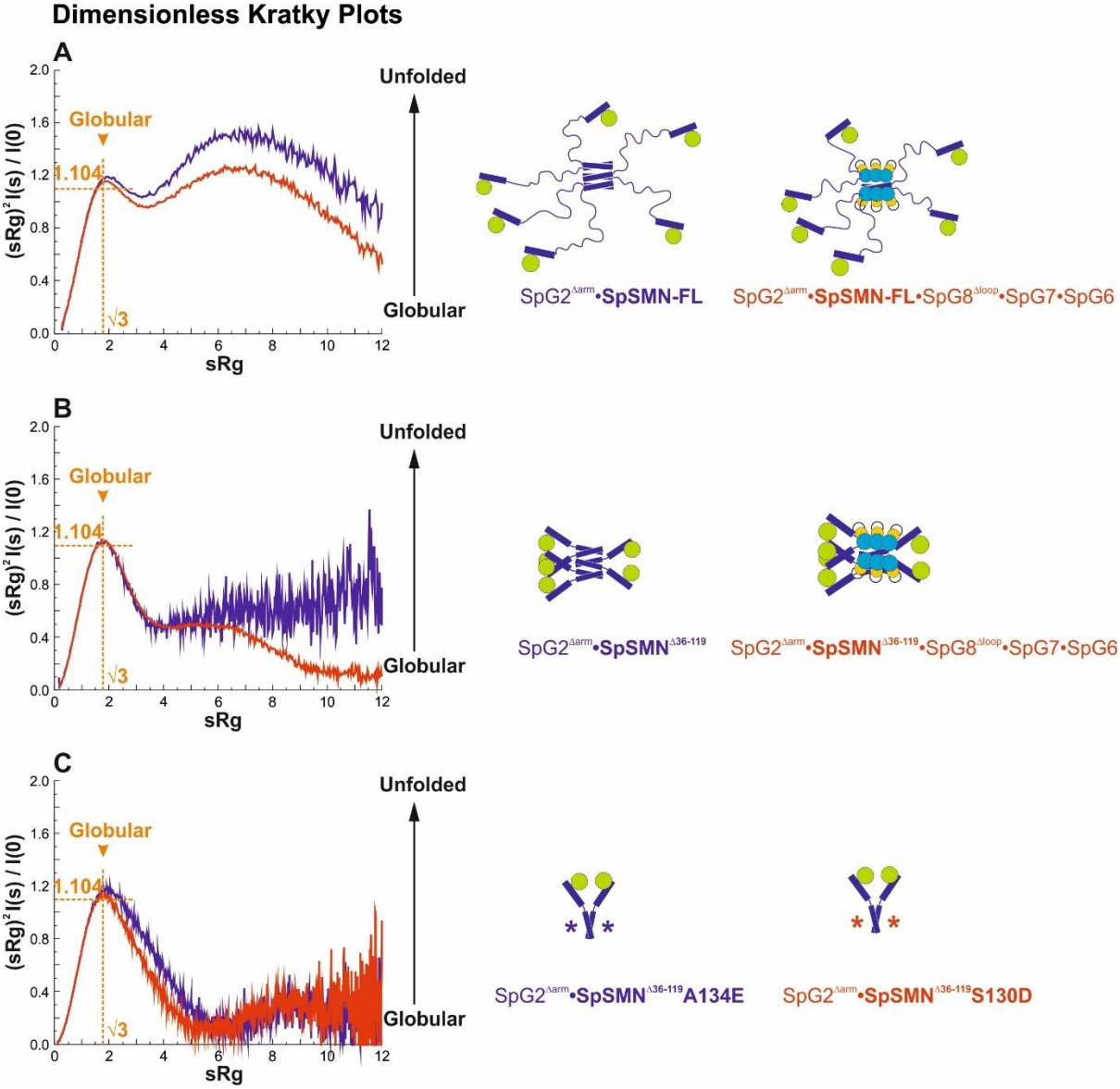


Figure 4.15: Dimensionless Kratky plots

For the generation of these plots from the traditional Kratky plot [$s^2I(s)$ vs s], Rg and I(0) were obtained from the Guinier analysis of each complex. Dimensionless Kratky plots are shown in A: SpSMN-FL complexes, B: $SpSMN^{\Delta 36\text{-}119}$ complexes, C: globular standards. The typical peak at (√3, 1.104) for globular particles are shown in orange dashed lines. For the globular standards (C), the signal decays rapidly after the peak. For $SpSMN^{\Delta 36\text{-}119}$ complexes (B), the signal decays rapidly after the peak compared to that of the SpSMN-FL complexes (A). The relative trajectory of signal after the globular peak can be used to assess the quantitative difference in flexibility (e.g., relatively globular, or relatively flexible) between different complexes of different size and concentration.

angular range of $SpG2^{\Delta arm}$•**SpSMN-FL**•$SpG8^{\Delta loop}$•SpG7•SpG6 is slightly higher (s=0.164 nm^{-1}) than that of $SpG2^{\Delta arm}$•**SpSMN-FL** (s=0.155 nm^{-1}) and the average radius of gyration (Rg) of $SpG2^{\Delta arm}$•**SpSMN-FL**•$SpG8^{\Delta loop}$•SpG7•SpG6 (7.92 nm) is 5.4 Å lower than that of $SpG2^{\Delta arm}$•**SpSMN-FL** (8.46 nm). Second, the first P(r) peak remains unchanged (3.2 nm) for both complexes whereas the second peak is 11 Å smaller for $SpG2^{\Delta arm}$•**SpSMN-FL**•$SpG8^{\Delta loop}$•SpG7•SpG6, suggesting an overall

decrease in interdomain pair-wise distances. This is also accompanied by a ~20 Å reduction in average Dmax for the pentamer. Third, compared to SpG2$^{\Delta arm}$•**SpSMN-FL**, the significant downturn of signal after sRg=3.5 in the Dimensionless Kratky plot of SpG2$^{\Delta arm}$•**SpSMN-FL**•SpG8$^{\Delta loop}$•SpG7•SpG6 (Figure 4.15 A) is indicative of relatively less diffuse electron density contrast and hence loss of flexibility.

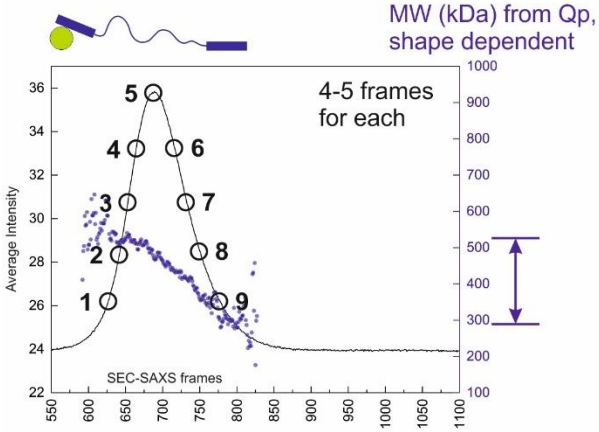
4.8.8 Properties of SpSMN-FL complexes: SpSMN-FL loses its oligomeric state upon SpG8$^{\Delta loop}$•SpG7•SpG6 binding

In the SEC-SAXS chromatograms (Figure 4.12 A-D), the depicted molecular weights at each frame is calculated from the volume of particles which is related to Porod Invariant (Qp) of the particles (see section 3.2.3.3.3). Qp is the area under the scattering curve transformed as Kratky plot, $s^2I(s)$ vs s, and hence shape dependent. As shown already, for SpSMN-FL complexes, the area under the Kratky plot is extremely undefined (Figure 4.15 A) due to the unstructured region of SpSMN (residues 36-119) resulting in slower decay of signal at higher angles. This results in inaccurate molecular weight calculations from inaccurate Qp (Figure 4.15 A). Hence, the depicted molecular weights calculated using Qp on the SEC-SAXS chromatograms for SpSMN-FL complexes (Figure 4.12 C-D) are largely inaccurate because of their dependence on the shape of the particles. In order to circumvent this problem, the molecular weights at regions 1-9 of the SEC-SAXS chromatograms (Figure 4.16 A-B) were determined according to the equation shown in C, which relies on the shape independent parameter, the zeroth angle intensity I(0). The I(0) and concentrations of the complexes (standards and SpSMN-FL complexes (for each of the 9 data regions)) were determined from Guinier analysis and UV-trace of the chromatogram, respectively (Figure 4.16 D). The information was computed into the formula shown in C, and the molecular weights for each of the 9 data regions were calculated. As shown in the table D, SpG2$^{\Delta arm}$•**SpSMN-FL**, on average, exists as a hexamer. SpG2$^{\Delta arm}$•**SpSMN-FL**•SpG8$^{\Delta loop}$•SpG7•SpG6, however, is tetrameric. Since the YG-box is the oligomeric domain within the complexes, and SpG8 N-terminus binds directly to the YG-box, this data suggests that upon binding to SpG8, SpSMN loses its oligomeric state *in vitro*.

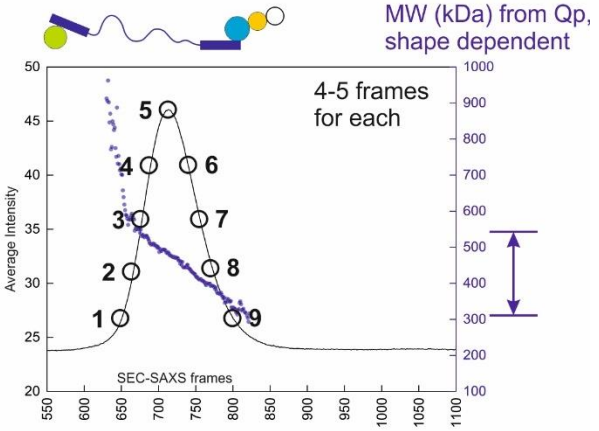
Figure 4.16: Molecular Weight calculations for SpSMN-FL complexes from I(0)
A-B: SEC-SAXS frames of SpSMN-FL complexes, showing 9 selected data regions (4-5 frames each)-

Molecular weight calculations from I(0)

A $SpG2^{\Delta arm}$•SpSMN-FL



B $SpG2^{\Delta arm}$•SpSMN-FL•$SpG8^{\Delta loop}$•SpG7•SpG6



C MW from I(0), shape independent → $MW_{unknown} = MW_{standard} \times \frac{Intensity(0)_{unknown} / Concentration_{unknown}}{Intensity(0)_{standard} / Concentration_{standard}}$

D

	Intensity (0) (arbitrary units)	Concentration (mg/ml)	Molecular Weight (kDa)	Oligomer
Standards				
$SpG2^{\Delta arm}$•$SpSMN^{\Delta 36-119}$•A134E	26.95	1.28	52.5 (known)	2 mer
$SpG2^{\Delta arm}$•$SpSMN^{\Delta 36-119}$•S130D	27.21	1.28	52.5 (known)	2 mer
$SpG2^{\Delta arm}$•SpSMN-FL				**(MW / 35.5)**
1	18.07	0.2	225.34	6.3 mer
2	31.94	0.331	239.50	6.7 mer
3	49.43	0.493	248.63	7.0 mer
4	80.98	0.774	253.71	7.3 mer
5	85.89	0.882	242.87	6.8 mer
6	67.90	0.714	236.06	6.6 mer
7	45.05	0.499	226.50	6.4 mer
8	25.07	0.313	198.82	5.6 mer
9	15.53	0.212	182.70	5.1 mer
			average = 228.24 kDa	**6.4 mer**
$SpG2^{\Delta arm}$•SpSMN-FL•$SpG8^{\Delta loop}$•SpG7•SpG6				**(MW / 73.45)**
1	20.34	0.132	382.49	5.2 mer
2	51.53	0.384	331.10	4.5 mer
3	85.30	0.720	294.08	4.0 mer
4	120.83	1.081	277.71	3.8 mer
5	156.21	1.470	263.78	3.6 mer
6	114.24	1.081	262.59	3.6 mer
7	77.47	0.725	265.24	3.6 mer
8	40.85	0.355	285.63	3.9 mer
9	13.51	0.114	294.17	4.0 mer
			average = 295.19 kDa	**4.0 mer**

-which were individually processed. C: Formula for MW calculation from I(0) and concentration, using I(0), concentration, and MW of known standard proteins. D: I(0) for standard proteins calculated from data shown in Figure 4.11. I(0) for both SpSMN-FL containing complexes were calculated from each selected data region (regions 1-9). Concentrations for each individual region was calculated from the corresponding UV280 trace of the chromatogram and the extinction coefficients of the complexes. Oligomeric states at each data region was calculated by dividing the monomeric MW of the corresponding complex (SpG2Δarm•SpSMN-FL = 35.5 kDa, SpG2Δarm•SpSMN-FL•SpG8Δloop•SpG7•SpG6 = 73.45 kDa).

4.8.9 Properties of SpSMN-FL complexes: Ab Initio models of SpSMN-FL complexes show extended structures originating from a central node

Although SpSMN-FL complexes are now shown to be highly flexible and therefore a single conformation cannot represent their true nature in solution, *ab initio* structural modeling was nevertheless attempted using DAMMIF. 20 models each for SpG2$^{\Delta arm}$•**SpSMN-FL** and SpG2$^{\Delta arm}$•**SpSMN-FL**•SpG8$^{\Delta loop}$•SpG7•SpG6 were generated. 4 representative models for each complex are shown in Figure 4.17 A-B. Computed scattering curves from each model shows exceptionally good fit ($Chi^2 < 1.0$) to the original scattering data (Figure 4.17 curves). Fit of the computed curves from strikingly different models to the same scattering data can be conveniently attributed to SAXS scattering data being spherically averaged. Nevertheless, a common feature of all the models, extended arm like structures connected at a roughly central node, further corroborates all previous data supporting a central, antiparallel, oligomeric YG-box domain of SpSMN, where the unstructured regions extend towards opposite directions.

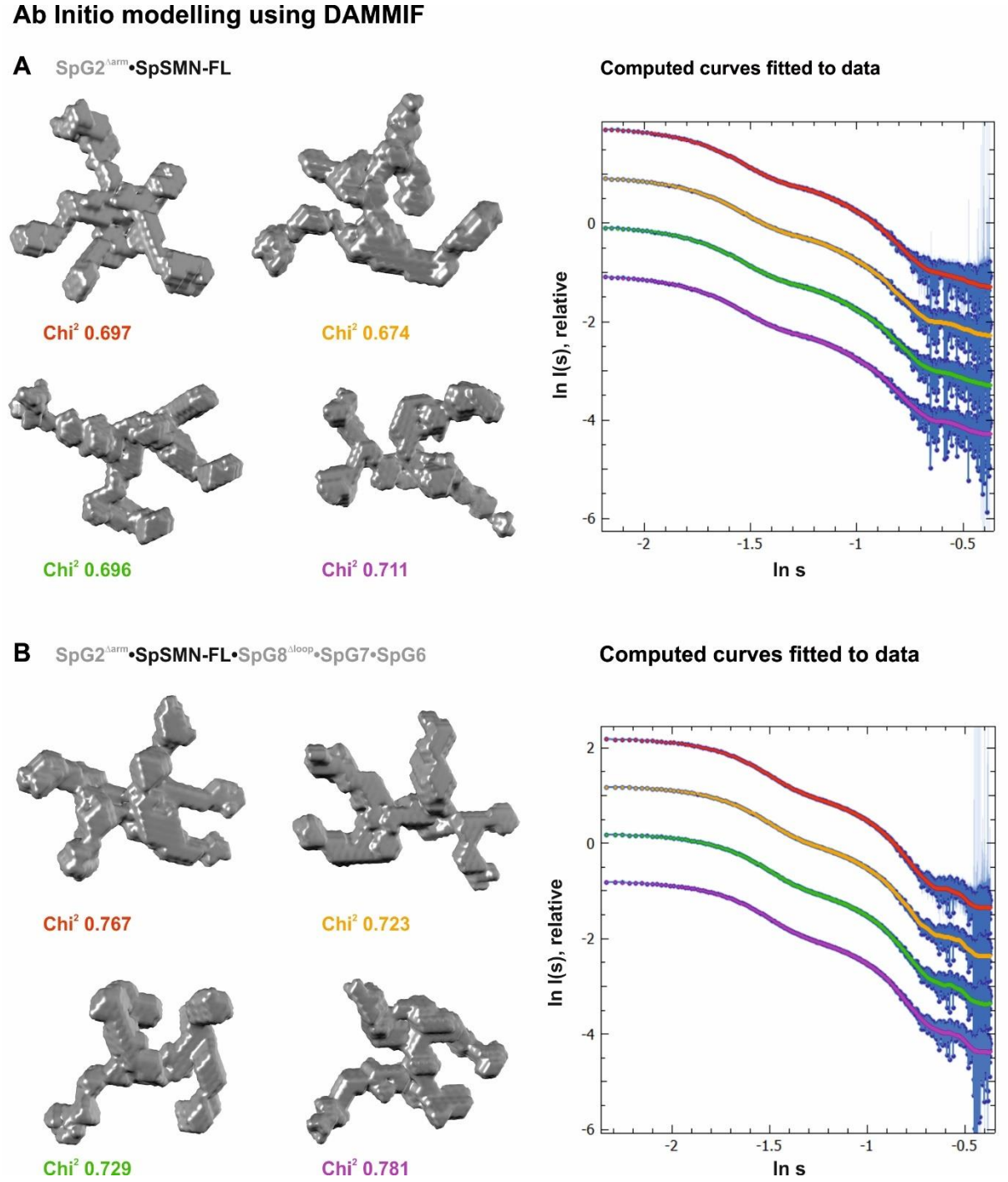


Figure 4.17: Ab initio modeling of SpSMN-FL complexes using DAMMIF
Four representative models of 20 generated models are shown with Chi² values. Each model exhibits nodal behavior where extended structures originate from a common point. The models were displayed using UCSF Chimera. Fit of the computed curve from each model (with corresponding colors) to the experimental solution scattering data (blue plots) are shown on the right panels.

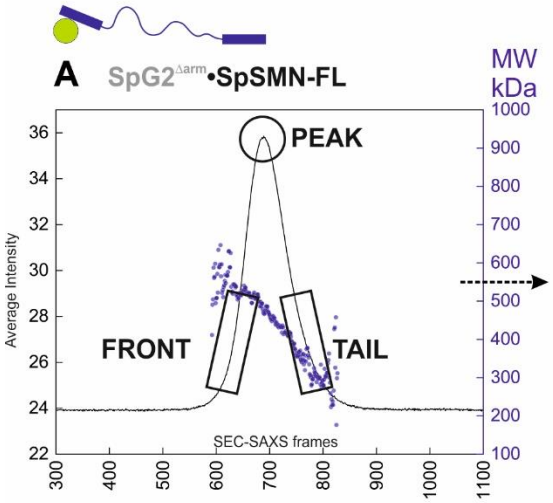
4.8.10 Properties of SpSMN-FL complexes: EOM analysis identifies hexameric SpG2$^{\Delta arm}$•SpSMN-FL throughout SEC-SAXS chromatogram

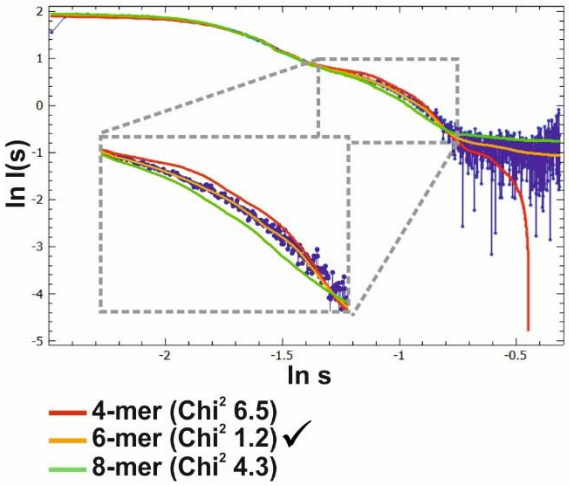
Existence of the oligomeric YG-box structure of SpSMN and available structure of Gemin2 subunit from *Drosophila melanogaster* (PDB ID: 4V98) made it possible to generate theoretical conformers of oligomeric SpG2$^{\Delta arm}$•SpSMN-FL (a pool of 10,000 conformers each for tetramers, hexamers and octamers) with random conformations of SpSMN's unstructured region (residues 36-119). Each theoretical pool was individually subjected to EOM analysis using SAXS data from the peak, front and tail regions of SEC-SAXS chromatogram (Figure 4.18 A). Fit of the computed scattering curves of the resulting final ensemble of conformers to the experimental data was analysed (Figure 4.18 B-D). Optimized ensembles of hexameric SpG2$^{\Delta arm}$•SpSMN-FL (orange) yielded excellent fits (peak:Chi^2 1.2, front:0.7 and tail:0.9) to data from all three regions of chromatogram (Figure 4.18 B-D, orange curves). Octameric (green curves) and tetrameric (red curves) SpG2$^{\Delta arm}$•SpSMN-FL yielded excellent fits to data from the front and tail of the chromatogram, respectively (each with Chi^2 0.8), but not to the peak (Chi^2 4.3 and 6.5, respectively). Interestingly, the lack of fit of the computed curves (Figure 4.18 B-C, inset) is especially emphasized in the region of the scattering which precisely corresponds to the region shown (dotted box) in the Dimensionless Kratky plot (Figure 4.18 E). As demonstrated earlier in section 4.8.6, this region of scattering is predominantly the contribution from the unstructured region of SpSMN. Optimized ensembles of hexameric SpG2$^{\Delta arm}$•SpSMN-FL comprised approximately twice the number of conformers compared to those of octamers and tetramers (Figure 4.18 F). This suggests a possibility that a hexameric SpG2$^{\Delta arm}$•SpSMN-FL could be the most dynamic oligomeric state for the complex.

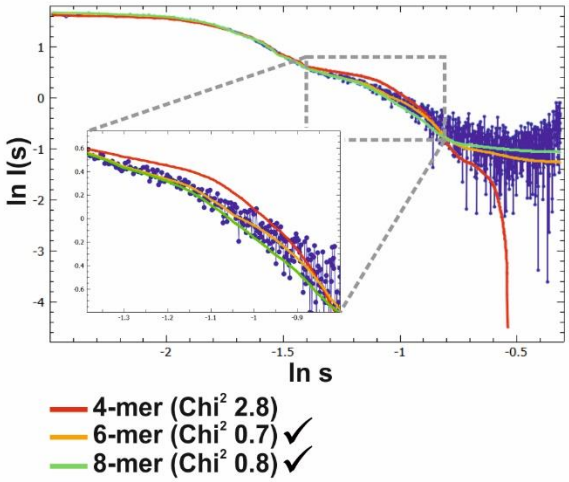
Figure 4.18: EOM analysis of SpG2$^{\Delta arm}$•SpSMN-FL

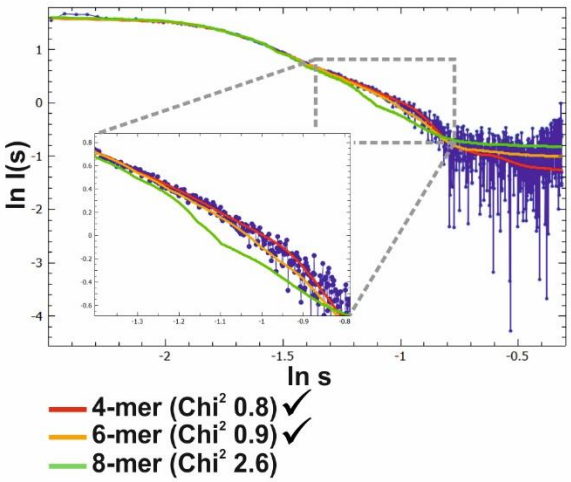
SAXS curves were generated using data from the peak, front and tail of the SEC-SAXS chromatogram (A). Pools of 10,000 theoretical conformers, each for Tetramers, hexamers, and octamers of SpG2$^{\Delta arm}$•SpSMN-FL were generated. Each pool was individually subjected to EOM optimization against SAXS curves generated from peak, front and tail. Fit of the computed curves from the optimized ensembles for each oligomer to the experimental SAXS curves are shown in B-D (as double log plots). Chi^2 values are shown on the right of the panels. Goodness of fit in an intermediate region of the scattering curve is emphasized in the insets (B-D, dotted box). The Dimensionless Kratky plot of SpG2$^{\Delta arm}$•SpSMN-FL (E) showing the corresponding region (dotted box) of the scattering curve emphasized in the insets in B-D. Total number of conformers in the final ensembles for each oligomer in shown in F.

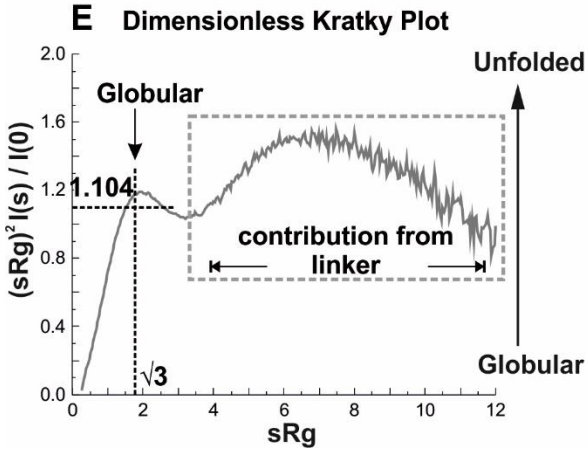
EOM analysis











F Number of conformers per ensemble

	PEAK	FRONT	TAIL
4-mer	—	—	4
6-mer	9	7	3
8-mer	—	4	—

4.8.11 Properties of SpSMN-FL complexes: EOM ensembles classify hexa- and octa-meric SpG2$^{\Delta arm}$•SpSMN-FL as fully flexible (R^{flex}ensemble > R^{flex}pool)

Comparison of the size distributions of the optimized final ensemble of conformers to the original pool of 10,000 theoretical conformers may in some cases reveal the propensity (or the lack thereof) of an intrinsically disordered protein towards a certain overall dimension. Size distributions within the original pool of theoretical conformers follow a normal distribution (Figure 4.19 A-C, dashed curves). After the EOM run, the combined size distributions (from peak, front and tail) within optimized ensembles of hexameric SpG2$^{\Delta arm}$•SpSMN-FL (Figure 4.19 A) cover a much broader range of dimensions compared to octamers and tetramers (Figure 4.19 B-C). The dimensions of hexamers range from predominantly 21 nm at the peak and tail regions to predominantly 26 nm at the front region (Figure 4.19 A, dashed lines). The term Rflex (a measure of the degree of flexibility of a system) for the hexameric ensembles being significantly higher than the Rflex of the original pool, classifies hexameric SpG2$^{\Delta arm}$•SpSMN-FL as a fully flexible system. Octamers at the front region and tetramers at the tail region are predominantly of 28 nm and 24 nm, respectively (Figure 4.19 B-C). Rflex values lower than that of pool also classifies octamers as fully flexible, whereas Rflex for tetramers is significantly smaller than that of pool and therefore is not classified as fully flexible.

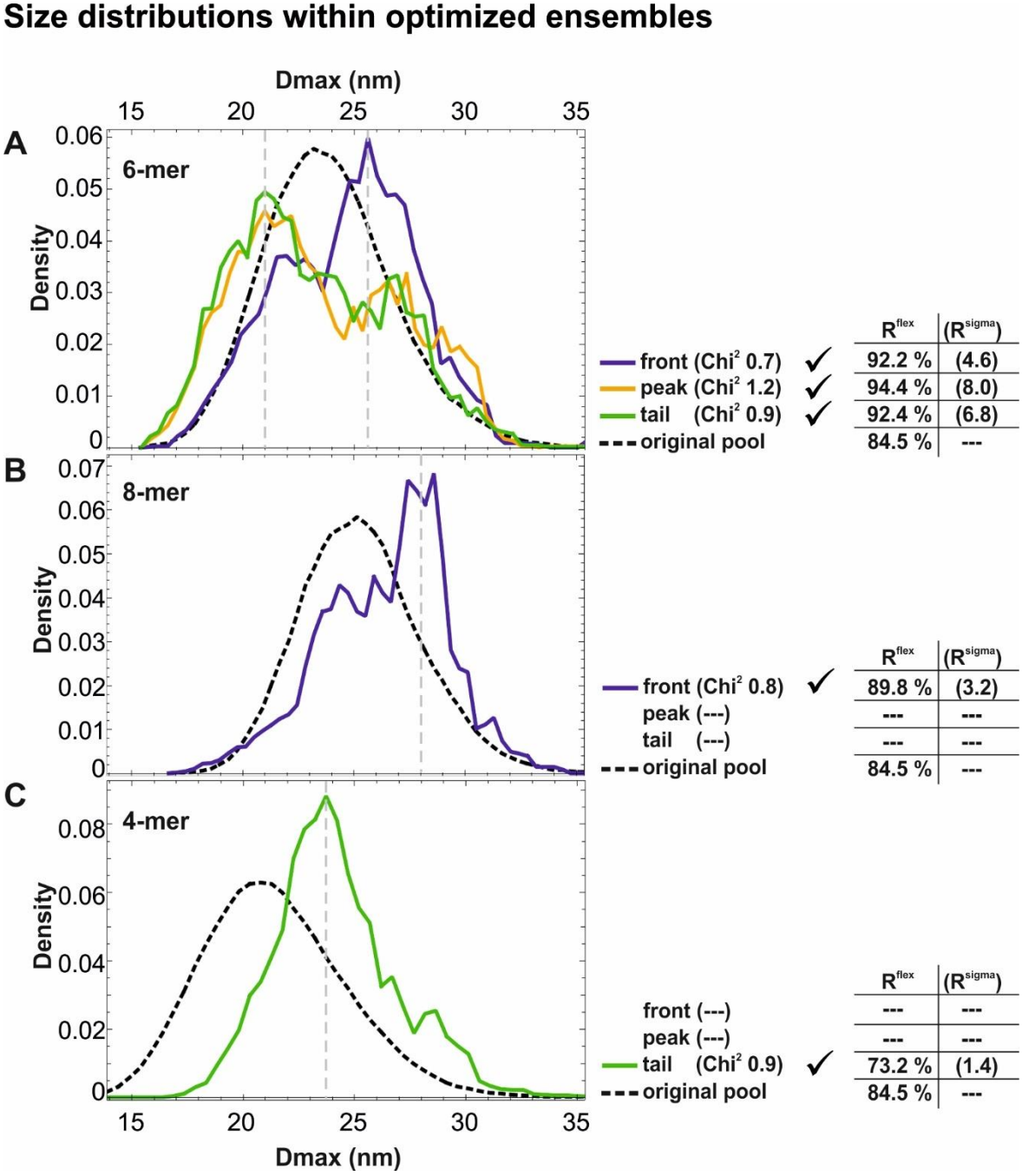


Figure 4.19: Size distributions within optimized ensembles

Within the optimized ensembles, size (maximum particle dimension Dmax) distributions of the selected conformers that fit to the experimental data are shown for hexamers (A), octamers (B) and tetramers (C). The predominant particle size within each optimized ensemble is shown as grey dotted lines. Size distributions within the theoretical pool of 10,000 conformers is shown as black dotted curves. Chi^2 values for each ensemble is shown on right of panel. Rflex (a measure of the degree of flexibility) and Rsigma (variance of ensemble distribution with respect to pool). When $Rflex^{ensemble} > Rflex^{pool}$ (fully flexible system), Rsigma should be >1, and vice versa.

5. DISCUSSION

5.1 Introductory notes

The biogenesis of the Sm-core of UsnRNPs occurs both in the nucleus and the cytoplasm of eukaryotic cells. The UsnRNA and Sm proteins' genes are first transcribed by Pol II and exported into the cytoplasm. Once in the cytoplasm, the UsnRNAs undergo further maturation steps and the Sm proteins are translated from their respective mRNAs by the ribosome. Then, the Sm proteins and the UsnRNAs assemble into mature Sm-core particles by an assisted assembly process and are imported back into the nucleus where they dissociate from the assembly factors and are targeted to cajal bodies for further maturation steps. The cytoplasmic phase has been understood in some detail due to a comprehensive biochemical investigation by several groups. These results revealed an elaborate assembly machinery whose function can be divided into an early and a late phase. The early phase is dominated by the PRMT5 complex and serves at least two functions. This complex recruits newly synthesized Sm proteins by means of the action of the assembly chaperone pICln, which picks up newly synthesized Sm proteins from the ribosome exit tunnel and serves to prevent non-cognate interactions. Secondly, PRMT5 bound Sm proteins acquire symmetric dimethylation marks on designated arginine residues in their C-terminal tails. The late cytoplasmic phase involves the action of the SMN complex which removes the assembly chaperone (pICln) induced kinetic trap and loads the Sm protein intermediates onto the Sm-site of UsnRNA, thus leading to the generation of the mature Sm-core.

The PRMT5 complex has been thoroughly analyzed by structural biological techniques. However, a comprehensive structural understanding of the SMN complex remains elusive. Although structures of some of the subunits of the human SMN complex and/or domains thereof exists, attempts to crystallize or to prepare suitable samples for cryoEM of the whole SMN complex (all 9 subunits) have not been successful. To gain structural insight into the SMN complex, I chose the SMN complex of *S. pombe*, as a model system due to its minimalistic composition. In collaboration with the lab of Dr. Remy Bordonne (Institute of Molecular Genetics of Montpellier, CNRS, France) we could show that this complex consists of 5 core subunits only, namely SpSMN, SpGemin2, SpGemin8, SpGemin7, and SpGemin6. While the SpGemin2 and SpSMN subunits have been shown to be the bona fide orthologs of

their human counterparts, the interaction pattern of the whole SpSMN complex was not known at the beginning of this thesis. The newly discovered SpGemin8, SpGemin7 and SpGemin6 subunits have been shown to be essential for viability (data not published). Although these subunits share low sequence identity with their human counterparts, their secondary structure predictions show high structural homology (see Figure 1.4 A and B). The only known structure of the SpSMN complex till date is a dimeric YG-box domain of SpSMN fused to an MBP-tag (Figure 1.5 D).

In this work, SpSMN sub-complexes were recombinantly co-expressed and purified from *E. coli*. The sub-complexes as well as the reconstituted pentamer were characterized by gelfiltration for their oligomeric states and hydrodynamic properties. Using X-ray crystallography, the structural basis of higher order oligomers of the YG-box was elucidated. Using small angle X-ray scattering coupled with gelfiltration (SEC-SAXS), various biophysical properties such as, oligomeric states, intrinsic disorder (Guinier analysis), domain architecture (p(r) function), flexibility (Kratky plot), and conformational distributions (EOM) were determined.

5.2 *S. pombe* possesses an elaborate SMN complex than previously thought

Except plants and unicellular eukaryotes, most eukaryotic species (metazoans) contain an elaborate SMN complex consisting of at least Gemin2, SMN, Gemin8, Gemin7, and Gemin6 (Kroiss et al, 2008). Metazoan genes are intron dense (Miao et al, 2006) where the majority of genes contain at least 2-8 introns. In such cases, the crucial role of splicing in gene expression necessitates a highly functional and efficient UsnRNP biogenesis machinery, which includes the SMN complex as a central assembly hub. Interestingly, *S. cerevisiae*, which is one of the least intron dense unicellular eukaryotes with only 3.8% of its genes containing introns (Barras et al, 2008), possesses only a Gemin2 homologue, whose function is not yet clear. All other gemins, and even SMN, is absent from its genome, suggesting that S. cerevisiae assembles U snRNPs through a profoundly different mechanism than other eukaryotes. On the contrary, over 50% of *S. pombe* genes contain introns. Until recently, it was believed that *S. pombe* contains only the orthologs of Gemin2 and SMN (Hannus et al, 2000; Owen et al, 2000; Paushkin et al, 2000). This changed with advanced sequence homology tools, which allowed us in collaboration with Dr. Remy Bordonne to discover putative homologues of Gemin6, 7 and 8 (denoted SpGemin8,

SpGemin7 and SpGemin6). Using the already established consensus interaction map of the human SMN complex (Figure 1.5 A, Otter et al. 2007), I designed di- and tr-cistronic constructs for recombinant co-expression and purification. In brief, the complexes His-SpGemin2•SpSMN, SpGemin2•His-SpSMN•SpGemin8, His-SpGemin8•SpGemin7•SpGemin6 and His-SpGemin7•SpGemin6 were successfully purified, which demonstrates that the SpSMN complex components exhibit highly similar interaction pattern to those of the human SMN complex. In addition, the SpGemin8 subunit was shown to be the connecting link between SpSMN and SpGemin7 by gelfiltration binding assays. Thus, the three proteins were proven to be bona fide orthologs of their human counterparts. Although the precise functions of SpGemins 8, 7 and 6 remain unclear, in an intron dense genome like that of *S. pombe,* they may serve to distribute the functional load during UsnRNP biogenesis. Together, these results show that the *S. pombe* possesses an elaborate SpSMN complex consisting of all 5 core subunit homologs of the human SMN complex (Figure 5.1).

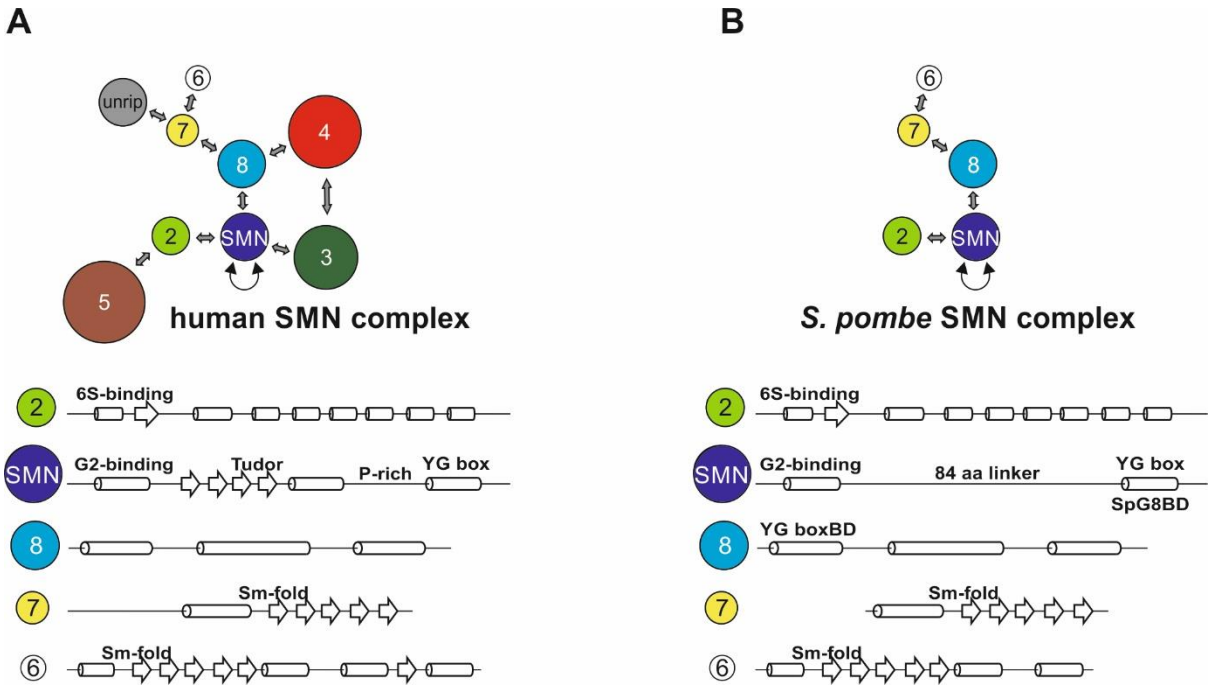


Figure 5.1: Human and *S. pombe* SMN complexes
Interaction map and secondary structural elements of the human (A) and *S. pombe* (B) SMN complexes. The newly discovered SpGemins 8, 7 and 6 display remarkable structural homology with their human counterparts. SpGemin7 and 6 contain the Sm-fold. SpGemin8 is entirely helical. The N-terminal helix is the YG-box binding domain of SpGemin8. (BD=binding domain)

5.3 SpSMN is the only contributor to the large hydrodynamic properties of the SpSMN complex

The purified complexes SpG2•SpSMN, SpG2•SpSMN•SpG8, SpG8•SpG7•SpG6 and SpG7•SpG6 were analyzed by gelfiltration chromatography to study their molecular weights and homogeneity. The crystal structures of the human orthologs of SpG7•SpG6 (Ma et al, 2005) shows heterodimeric, globular complex. Indeed, as demonstrated in section 4.1.3, SpG7•SpG6 elute from a gelfiltration column in the low molecular weight region between 66 and 150 kDa marker, indicating a compact and globular shape. This suggest that SpG7•SpG6 (MW 21 kDa) forms higher order oligomers. Since SpG7•SpG6 are structurally similar to Sm proteins and contain an Sm-like fold, they may form ring shaped oligomeric structures. Interestingly SpG8•SpG7•SpG6 complex elutes near the 44 kDa marker which is remarkably close to the monomeric molecular weight of the trimeric complex (41 kDa). This suggests that in the presence of SpG8, the oligomeric propensity of SpG7•SpG6 is completely prevented. This also suggests that the binding site of SpG8 to SpG7 may directly interfere with the oligomeric surface of SpG7•SpG6 complex. All complexes containing SpSMN full-length protein including SpG2•SpSMN, SpG2•SpSMN•SpG8, and the reconstituted pentameric complex SpG2•SpSMN•SpG8•SpG7•SpG6 exhibit remarkably similar hydrodynamic properties and elute near the 669 kDa marker from gelfiltration columns. This suggests that the SpGemins 8, 7 and 6, which are globular in nature and lack large unstructured regions, have no effect on the overall molecular size of the SpSMN complex. Although SpGemin2 is predominantly globular in structure (structures of human and fly orthologs are available: PDB 4V98, 5XJL), its N-terminal 6S binding arm with 80 residues almost identical in length to the SpSMN's unstructured linker region, which is 84 residues. Surprisingly, this region of SpG2 had no effect on the overall size of SpG2•SpSMN complex (see section 4.3.1). On the contrary, deletion of either SpSMN's unstructured linker or truncation of the C-terminal YG-box domain, had drastic effects on the overall size of SpG2•SpSMN complex (see section 4.3.1). This is clearly demonstrated by the >10-fold loss in apparent molecular size as judged by gelfiltration. Thus, the sole contributors to the overall size and hydrodynamic properties of the full-length SpSMN containing complexes were found to be the unstructured linker of SpSMN and its C-terminal YG-box oligomerization domain. In addition, results of this work show that the N-terminal helix of SpG8 is necessary and sufficient to interact with the YG-box of SpSMN (see section 4.6). In

summary, the SpSMN protein serves as the oligomeric core of the SpSMN complex and provides biding platforms for SpG2 at the N-terminus and SpG8•SpG7•SpG6 at the C-terminus. In addition, SpSMN also controls the oligomeric and hydrodynamic properties of the SpSMN complex through its YG-box domain and the unstructured linker, respectively. Hence, the primary focus of the work was on the SpSMN subunit. Using various techniques such as gelfiltration, X-ray crystallography and small angle X-ray scattering, the oligomeric properties of the YG-box, and the conformational properties of the unstructured linker (respectively) were studied.

5.4 SpSMN exists as dimers through octamers *in vitro*

The large hydrodynamic properties of complexes containing full length SpSMN is a consequence of both the unstructured linker and the C-terminal YG-box oligomerization domain of SpSMN. Flexibility resulting from the unstructured linker was the biggest obstacle in accurately determining the oligomeric state of SpSMN. For this reason, deletion mutants of SpSMN's unstructured linker were constructed and the resulting SpG2•SpSMN complexes were studied by gelfiltration. The existence of dimers and tetramers of SpSMN has been previously shown (Gupta et al, 2015). Results from our gelfiltration studies revealed that SpSMN can exist as dimers, tetramers, hexamers, octamers, and even >octamers *in vitro* (see section 4.4). These experiments were carried out at concentrations >>1 µM. It has been previously reported that at a concentration of 1 µM, SpSMN exists as an equilibrium mixture of dimers and tetramers (Gupta et al, 2015). Since the cellular concentration of SpSMN has been reported to be less than 1 µM (Gupta et el, 2015), it has been suggested that the dimeric form of SpSMN represents the functional oligomeric species in yeast. This is further corroborated by the fact that the distribution of SMN protein in yeast (both in cytoplasm and nucleus) is diffuse without any concentrated localizations (Paushkin et al, 2000). On the contrary, the role of SMN in maintaining the integrity of cajal bodies in human cells have been investigated to some extent (Neugebauer 2017). While the intrinsically disordered regions, RNA binding properties, and oligomerization of the human SMN complex may have a role in the liquid-liquid phase separation properties of cajal bodies, it is unclear what specific functions the intrinsically disordered region of SpSMN might play in yeast which lacks concentrated localizations of SpSMN. Nevertheless, this unstructured region of SpSMN might serve as binding sites for other SpSMN associated proteins which are yet to be discovered.

5.5 Structural basis of SpSMN oligomerization

YG-box dimerization has been well characterized in two structural studies (Figure 1.5 C-D). The reported structures show that the YG-box forms glycine zipper dimers through its 3x(YxxG) motif. Gupta et al (2015) also showed that higher order oligomers of SMN are formed by the multimerization of SMN dimers which likely means that dimers are the fundamental unit of higher order SMN oligomers. The authors also demonstrated by a disulfide linkage assay that the oligomeric interaction interface is different from the dimerization interface. In this work, the crystal structure of a fragment of SpSMN lacking the entire 84 residues long unstructured region ($SpSMN^{\Delta 36\text{-}119}$) was determined. This structure extends the YG-box helical region towards the N-terminus, well beyond the existing structures (Figure 4.5 B, Figure 5.2). As a result of this, a previously unseen salt-bridge between Asp121 and Lys125 at the N-terminal end could be characterized, which may be involved in intra-helix stability or have some specific function. In addition, the classical glycine zipper dimers were found to be stacked upon each other in an anti-parallel fashion (Figure 4.5 D). This new interface between dimers was proven to be the oligomeric interface in solution by mutational analysis of the two innermost residues Ser130 and Ala134 at the point of closest contact between dimers. Substituting these residues to bulkier residues Asp130 and Glu134 respectively, prevented oligomerization beyond a dimeric YG-box due to steric hindrance (Figure 4.6 B-E). This further corroborates the previous notion (Gupta et al, 2015) that the fundamental unit of YG-box oligomers are glycine-zipper dimers and the oligomeric and dimeric interfaces are distinct from each other. Dimer-dimer oligomerization occurs mostly through large hydrophobic interaction surface. At the residue level, there are mainly 4 amino acids which play crucial roles in oligomerization: Trp131, Ser130, Ala134 and Thr138 (see section 4.5.5). Structural alignments of the existing YG-box structures (MBP-SpYG-box dimer, salmon; MBP-hYG-box dimer, green) with the oligomeric structure from this work (blue and yellow) shows that the conformations of the sidechains are identical to those of the existing YG-box structures (Figure 5.2). First, the corresponding residue of yeast Ala134 in human YG-box (green) is Ser270 where, the OH group (red) is pointed away from the Tryptophan and would not sterically hinder the CH-π interactions necessary for oligomerization. Second, the corresponding residue of yeast Thr138 in human YG-box (green) is Thr274 which exhibits an identical conformation to form hydrogen bond with the Tryptophan residue. These observations rule out the possibility of these specific

interactions being crystallization artifacts (Figure 5.2) but rather argue for the structural basis of oligomerization in the human SMN complex.

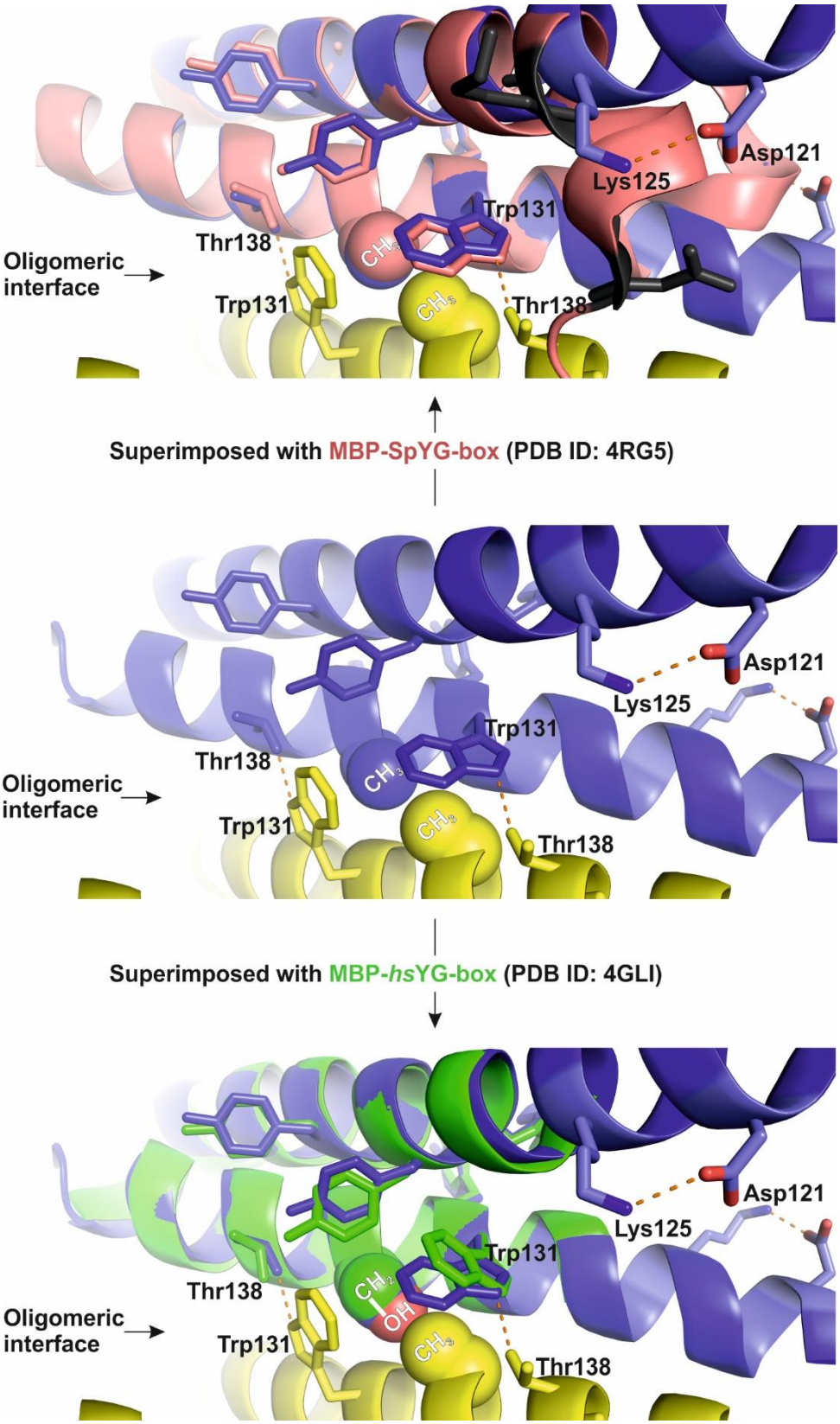


Figure 5.2: Structural alignments of human YG-box and existing SpYG-box to the characterized structure of this work.
The characterized oligomeric interface between the dimers (blue and yellow glycine-zipper dimers) of SpSMN YG-box is shown in the middle panel. Superimposition of the existing Sp YG-box glycine-zipper dimer structure (salmon color, PDB-ID 4RG5) and the human YG-box glycine-zipper dimer structure (green color, PDB-ID 4GLI) are shown in the top and bottom panels, respectively. The conformations of the residues Trp131 and Thr138 from the oligomeric structure, perfectly align with the corresponding residues of the existing dimeric structures. The corresponding for SpYG-box Ala134 in humans is Ser270 (sidechain -OH group shown as red sphere) which would not interfere with the oligomeric stacking.

5.6 The unstructured region of SpSMN imparts intrinsic disorder, flexibility, and dynamic properties to the SpSMN complex

Apart from the YG-box, the second feature that gives large hydrodynamic properties to SpSMN containing complexes is the unstructured region of SpSMN. As a result of anti-parallel SpSMN oligomers, the unstructured regions extend in opposite directions. As a consequence, the compact YG-box oligomer forms the core of the complex which is connected to the SpG2 globular subunits by long unstructured linkers. Such a complex is expected to be highly flexible and behave as intrinsically disordered protein. The only structural biological technique that can be used to study IDPs or highly flexible proteins is small angle X-ray scattering (SAXS). A detailed overview of SAXS basic principles, data collection, data processing and interpretation is described in methods section 3.2.3.3. Intrinsically disordered proteins follow the Guinier law up to shorter angular range compared to compact and globular proteins of similar molecular weight. Guinier analysis showed clearly that SpSMN complexes behave as highly disordered proteins (see section 4.7.4). The asymmetric p(r) curves with tailing for all complexes with full-length SpSMN shows typical behavior for multidomain proteins connected with flexible linkers (see section 4.7.5). As shown in the cartoon (Figure 4.13), the globular YG-box oligomeric core is connected with globular SpG2 subunits by SpSMN's unstructured linker. The flexibility of these unstructured regions was quantitatively assessed by dimensionless Kratky plots which showed a remarkable deviation in signal from the typical behavior of globular proteins (see section 4.7.6). Interestingly, SpG8•SpG7•SpG6 was found to modulate the overall biophysical properties of the SpSMN complex. This is clearly seen in the loss of Dmax in the p(r) curve (Figure 4.13 C-D) which is in line with a significant downturn of signal in the dimensionless Kratky plot (Figure 4.14 A). In addition to this, the presence of SpG8•SpG7•SpG6 seems to restrict SpSMN oligomerization to a tetramer (Figure 4.15). The EOM (Ensemble Optimization Method) analysis identified tetramers,

hexamers, and octamers as the oligomeric states of SpSMN (see section 4.7.10). The hexameric form was found to be the most dynamic form amongst these oligomers. This is clearly evident by the nearly double the number of conformers per ensemble compared to tetramers and octamers (Figure 4.17). In agreement with this, the size distribution analysis of the ensembles shows a significantly broader conformational space for the hexameric form compared to tetramers and octamers (see section 4.7.11).

Together, these results have enabled us to propose a structural architecture of the SpSMN complex (Figure 5.3). The core of the complex is an oligomeric SpSMN where parallel YG-box dimers are stacked upon each other in an anti-parallel fashion. In the absence of SpGemins-8, 7 and 6, the predominant SpSMN oligomer found *in vitro* was a hexameric species, whereas in the presence of SpGemins-8, 7 and 6, the predominant SpSMN oligomer found *in vitro* was a tetrameric species (see section 4.8.8). Binding of SpGemins-8, 7 and 6 to the YG-box alters the biophysical properties of the whole complex by restricting the oligomeric state and partially mitigating its flexibility (see section 4.8.6). The oligomeric SpSMN core subunit acts as a hub protein which provides the binding platforms for SpGemin2 at the N-terminus and for SpGemin8, 7 and 6 at the C-terminus.

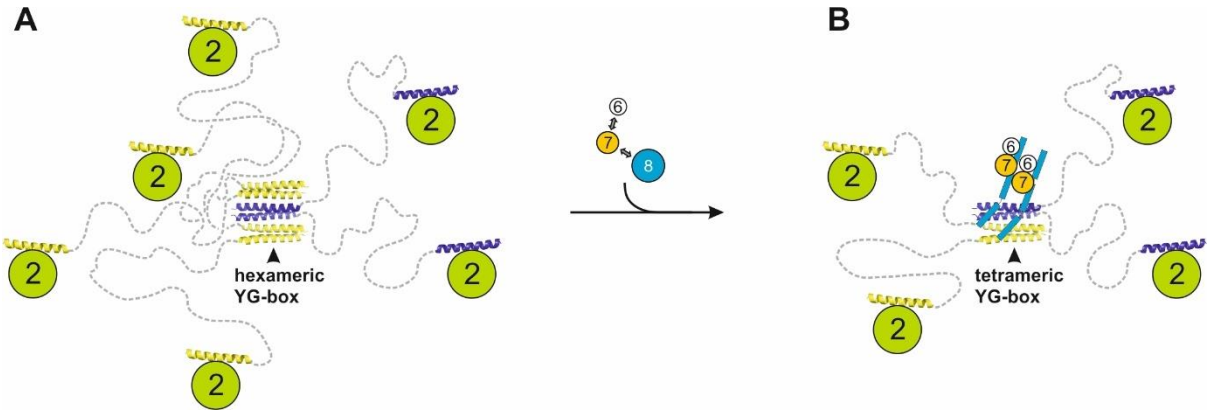


Figure 5.3: A structural model of the SpSMN complex
A. SpG2•SpSMN exists as a hexameric complex in the absence of SpG8•SpG7•SpG6. The SpSMN YG-box core is shown as anti-parallel oligomer of yellow and blue glycine-zipper dimers. The N-terminal SpG2 binding helix of SpSMN monomers are shown in corresponding colors. The SpG2 subunit is shown as a green cartoon. B. Binding of SpG8•SpG7•SpG6 to the SpG2•SpSMN via SpSMN's YG-box results in the formation of the hetero-pentameric SpSMN complex SpG2•SpSMN•SpG8•SpG7•SpG6, where SpSMN is restricted in its oligomeric state and is a tetramer. The unstructured linker of SpSMN is shown as dashed lines.

6. CONCLUSION

The SMN complex is a crucial biogenesis factor for spliceosomal UsnRNPs. Destabilization of the SMN complex due to deletions and/or mutations in the SMN gene, as observed in spinal muscular atrophy (SMA) perturbs the UsnRNP biogenesis resulting in defective splicing of pre-mRNAs. The extraordinarily complex biophysical properties of the SMN complex, which displays oligomeric and intrinsic disorder properties poses a major challenge towards its structural characterization. Due to its minimalistic composition, the *Schizosaccharomyces pombe* SMN complex was used in this work as a model system to study the oligomeric and intrinsic disorder properties of the complex. Through biochemical and structural analyses, SpSMN was shown to be the central oligomeric core of the pentameric SpSMN complex. A previously unseen oligomeric interface of the YG-box was structurally characterized, which showed in atomic detail that SpSMN forms higher order oligomers by multimerization of YG-box dimers in an anti-parallel fashion. The results highlight the role of the SpSMN subunit as the central organizer of the SpSMN complex, which not only controls the biophysical properties of the complex, but also provides binding sites for peripheral subunits. An especially important highlight of this work was the use of small angle X-ray scattering technique to comprehensively elucidate the dynamic behavior of the SpSMN complex as a function of the core SpSMN subunit. The obtained results make a strong case for a framework of experiments involving biochemical and small angle X-ray scattering experiments to study the properties and dynamics of the human SMN complex, which may unravel the mechanistic details of its functions.

7. REFERENCES

Berget SM, Berk AJ, Harrison T, Sharp PA. Spliced segments at the 5′ terminus of adenovirus-2 late mRNA: A role for heterogeneous nuclear RNA in mammalian cells. Cold Spring Harb Symp Quant Biol. 1977a;42:523-530.

Berget SM, Moore C, Sharp PA. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci. 1977b;74:3171-3175.

Bernadó P, Svergun DI. Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. Molecular biosystems. 2012;8(1):151-67.

Branscombe TL, Frankel A, Lee JH, Cook JR, Yang ZH, Pestka S, Clarke S. PRMT5 (Janus kinase-binding protein 1) catalyzes the formation of symmetric dimethylarginine residues in proteins. Journal of Biological Chemistry. 2001;276(35):32971-6.

Broker TR, Chow LT, Down AR, Gelinas RE, Hassel JA, Klessig DF, Lewis JB, Roberts RJ, Zain BS. Displacement loops in adenovirus DNA-RNA hybrids. Cold Spring Harb Symp Quant Biol. 1977;42:531-554.

Burge CB, Tuschl T, Sharp PA. The RNA World (Second Edition). Cold Spring Harbor Laboratory Press.1999;525-560.

Burghes AH, Beattie CE. Spinal muscular atrophy: why do low levels of survival motor neuron protein make motor neurons sick? Nature Reviews Neuroscience. 2009;10(8):597-609.

Chari A, Golas MM, Klingenhäger M, Neuenkirchen N, Sander B, Englbrecht C, Sickmann A, Stark H, Fischer U. An assembly chaperone collaborates with the SMN complex to generate spliceosomal SnRNPs. Cell. 2008;135(3):497-509.

Chaudhuri B, Muñoz IG, Qian S, Urban VS. Biological Small Angle Scattering: Techniques, Strategies and Tips. Springer. 2017;4.

Chow LT, Gelinas RE, Broker TR, Roberts RJ. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. Cell. 1977a;12:1-8.

Chow LT, Roberts JM, Lewis JB, Broker TR. A map of cytoplasmic RNA transcripts from lytic adenovirus type 2, determined by electron microscopy of RNA:DNA hybrids. Cell. 1977b;11:819-836.

Cordin O and Beggs JD. RNA helicases in splicing. RNA Biology. 2013;10:1, 83-95.

Fair BJ, Pleiss JA. The power of fission: yeast as a tool for understanding complex splicing. Current genetics. 2017;63(3):375-80.

Feltz C, Anthony K, Brilot A, Krummel DP. Architecture of the Spliceosome. Biochemistry. 2012;51:3321-3333.

Fischer U, Englbrecht C, Chari A. Biogenesis of spliceosomal small nuclear ribonucleoproteins. Wiley Interdisciplinary Reviews: RNA. 2011;2(5):718-31.

Fischer U, Liu Q, Dreyfuss G. The SMN–SIP1 Complex Has an Essential Rolein Spliceosomal snRNP Biogenesis. Cell. 1997;90:1023-1029.

Fischer U, Liu Q, Dreyfuss G. The SMN–SIP1 complex has an essential role in spliceosomal snRNP biogenesis. Cell. 1997;90(6):1023-9.

Franke D, Petoukhov MV, Konarev PV, Panjkovich A, Tuukkanen A, Mertens HD, Kikhney AG, Hajizadeh NR, Franklin JM, Jeffries CM, Svergun D. ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. Journal of applied crystallography. 2017;50(4):1212-25.

Gonsalvez GB, Praveen K, Hicks AJ, Tian L, Matera AG. Sm protein methylation is dispensable for snRNP assembly in Drosophila melanogaster. RNA. 2008;14(5):878-87.

Grimm C, Chari A, Pelz JP, Kuper J, Kisker C, Diederichs K, Stark H, Schindelin H, Fischer U. Structural basis of assembly chaperone-mediated snRNP formation. Molecular cell. 2013;49(4):692-703.

Gruss OJ, Meduri R, Schilling M, Fischer U. UsnRNP biogenesis: mechanisms and regulation. Chromosoma. 2017;126(5):577-93.

Gupta K, Martin R, Sharp R, Sarachan KL, Ninan NS, Van Duyne GD. Oligomeric properties of survival motor neuron· Gemin2 complexes. Journal of Biological Chemistry. 2015;290(33):20185-99.

Hannus S, Bühler D, Romano M, Seraphin B, Fischer U. The Schizosaccharomyces pombe protein Yab8p and a novel factor, Yip1p, share structural and functional similarity with the spinal muscular atrophy-associated proteins SMN and SIP1. Human Molecular Genetics. 2000;9(5):663-74.

Jacques DA, Trewhella J. Small-angle scattering for structural biology—Expanding the frontier while avoiding the pitfalls. Protein science. 2010;19(4):642-57.

Jędrzejowska M, Gos M, Zimowski JG, Kostera-Pruszczyk A, Ryniewicz B, Hausmanowa-Petrusewicz I. Novel point mutations in survival motor neuron 1 gene expand the spectrum of phenotypes observed in spinal muscular atrophy patients. Neuromuscular Disorders. 2014;24(7):617-23.

Jin W, Wang Y, Liu CP, Yang N, Jin M, Cong Y, Wang M, Xu RM. Structural basis for snRNA recognition by the double-WD40 repeat domain of Gemin5. Genes & development. 2016;30(21):2391-403.

Kikhney AG, Svergun DI. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. FEBS letters. 2015;589(19):2570-7.

Kolb SJ, Kissel JT. Spinal muscular atrophy. Neurologic clinics. 2015;33(4):831-46.

Kroiss M, Schultz J, Wiesner J, Chari A, Sickmann A, Fischer U. Evolution of an RNP assembly system: a minimal SMN complex facilitates formation of UsnRNPs in Drosophila melanogaster. Proceedings of the National Academy of Sciences. 2008;105(29):10045-50.

Kupfer DM, Drabenstot SD, Buchanan KL, Lai H, Zhu H, Dyer DW, Roe BA, Murphy JW. Introns and splicing elements of five diverse fungi. Eukaryotic cell. 2004;3(5):1088-100.

Martin R, Gupta K, Ninan NS, Perry K, Van Duyne GD. The survival motor neuron protein forms soluble glycine zipper oligomers. Structure. 2012;20(11):1929-39.

Matera AG, Wang Z. A day in the life of the spliceosome. Nature reviews Molecular cell biology. 2014;15(2):108-21.

Meister G, Bühler D, Pillai R, Lottspeich F, Fischer U. A multiprotein complex mediates the ATP-dependent assembly of spliceosomal U snRNPs. Nature cell biology. 2001;3(11):945-949.

Montzka KA, Steitz JA. Additional low-abundance human small nuclear ribonucleoproteins: U11, U12, etc. Proceedings of the National Academy of Sciences. 1988;85(23):8885-9.

Moore MJ, Proudfoot NJ. Pre-mRNA processing reaches back to transcription and ahead to translation. Cell. 2009;136(4):688-700.

Mylonas E, Svergun DI. Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering. Applied Crystallography. 2007;40:245-9.

Nguyen TH, Galej WP, Fica SM, Lin PC, Newman AJ, Nagai K. CryoEM structures of two spliceosomal complexes: starter and dessert at the spliceosome feast. Current opinion in structural biology. 2016;36:48-57.

Ohno M, Segref A, Bachi A, Wilm M, Mattaj IW. PHAX, a mediator of U snRNA nuclear export whose activity is regulated by phosphorylation. Cell. 2000;101(2):187-98.

Otter S, Grimmler M, Neuenkirchen N, Chari A, Sickmann A, Fischer U. A comprehensive interaction map of the human survival of motor neuron (SMN) complex. Journal of Biological Chemistry. 2007;282(8):5825-33.

Owen N, Doe CL, Mellor J, Davies KE. Characterization of the Schizosaccharomyces pombe orthologue of the human survival motor neuron (SMN) protein. Human molecular genetics. 2000;9(5):675-84.

Paushkin S, Gubitz AK, Massenet S, Dreyfuss G. The SMN complex, an assemblyosome of ribonucleoproteins. Current opinion in cell biology. 2002;14(3):305-12.

Pellizzoni L, Yong J, Dreyfuss G. Essential role for the SMN complex in the specificity of snRNP assembly. Science. 2002;298(5599):1775-9.

Rambo RP, Tainer JA. Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. Biopolymers. 2011;95(8):559-71.

Rambo RP, Tainer JA. Accurate assessment of mass, models and resolution by small-angle scattering. Nature. 2013;496(7446):477-81.

Receveur-Bréchot V, Durand D. How random are intrinsically disordered proteins? A small angle scattering perspective. Current Protein and Peptide Science. 2012;13(1):55-75.

Schütz P, Karlberg T, Van Den Berg S, Collins R, Lehtiö L, Högbom M, Holmberg-Schiavone L, Tempel W, Park HW, Hammarström M, Moche M. Comparative structural analysis of human DEAD-box RNA helicases. PloS one. 2010;5(9):e12791.

Selenko P, Sprangers R, Stier G, Bühler D, Fischer U, Sattler M. SMN tudor domain structure and its interaction with the Sm proteins. Nature structural biology. 2001;8(1):27-31.

Trewhella J, Duff AP, Durand D, Gabel F, Guss JM, Hendrickson WA, Hura GL, Jacques DA, Kirby NM, Kwan AH, Perez J. 2017 publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution: an update. Acta Crystallographica Section D: Structural Biology. 2017;73(9):710-28.

Tria G, Mertens HD, Kachala M, Svergun DI. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. IUCrJ. 2015;2(2):207-17.

Tripsianes K, Madl T, Machyna M, Fessas D, Englbrecht C, Fischer U, Neugebauer KM, Sattler M. Structural basis for dimethylarginine recognition by the Tudor domains of human SMN and SPF30 proteins. Nature structural & molecular biology. 2011;18(12):1414.

Wahl MC, Fischer U. The right pick: structural basis of snRNA selection by Gemin5. Genes & development. 2016;30(21):2341-4.

Wahl MC, Will Cl, Lührmann R. The spliceosome: design principles of a dynamic RNP machine. Cell. 2009;136(4):701-18.

Will CL and Lührmann R. Spliceosome Structure and Function. Cold Spring Harb Perspect Biol. 2011;3:a003707.

Xu C, Ishikawa H, Izumikawa K, Li L, He H, Nobe Y, Yamauchi Y, Shahjee HM, Wu XH, Yu YT, Isobe T. Structural insights into Gemin5-guided selection of pre-snRNAs for snRNP assembly. Genes & development. 2016;30(21):2376-90.

Yan C, Hang J, Wan R, Huang M, Wong CC, Shi Y. Structure of a yeast spliceosome at 3.6-angstrom resolution. Science. 2015;349(6253):1182-91.

Yan C, Wan R, Shi Y. Molecular Mechanisms of pre-mRNA Splicing through Structural Biology of the Spliceosome. Cold Spring Harb Perspect Biol. 2019;11:a032409.

Zhang R, So BR, Li P, Yong J, Glisovic T, Wan L, Dreyfuss G. Structure of a key intermediate of the SMN complex reveals Gemin2's crucial function in snRNP assembly. Cell. 2011;146(3):384-95.

Zhu T, Niu DK. Mechanisms of intron loss and gain in the fission yeast Schizosaccharomyces. PLoS One. 2013;8(4):e61683.

8. ANNEXURE

8.1 Amino acid sequences of SpSMN complex subunits

>SpGemin2, UniprotKB P0CU08

```
  1 MPSKRKRNPL QYQTSGSLDE ETNQRSAFPQ IDNNSASESL EYDIPLDGLD YLATVREEAR
 61 KLVPFVAARR EPETRETIPL RKLEIEAGKK SFDPFLRYLL NIIDKEGERL EQYMESSSLD
121 ASILPKNLQQ WRVYIEHKAP CWAILAVVDL ATVLEILESL SSWLEKDAID LQSQWIFCFC
181 YKLPELLNGE DISTLRSVLK SLRSTHTSFP ALQMSASALQ AVLVYRYGQK DLFQT = 235 aa
```

>SpSMN, UniprotKB Q09808

```
  1 MDQSQKEVWD DSELRNAFET ALHEFKKYHS IEAKGGVSDP DSRLDGEKLI SAARTEESIS
 61 KLEEGEQMIN QQTETTLEGD THIQQFADNK GLSDEKPETR AAETHQEFME VPPPIRGLTY
121 DETYKKLIMS WYYAGYYTGL AEGLAKSEQR KD = 152 aa
```

>SpGemin8, UniprotKB Q9USY8

```
  1 MSSEITEGDL QKFHDEHFNA KAVNLWNVAF AQNDRGGNSE SANVEYTQSV ERYPDGTIRT
 61 LTDEQILWFR ESEKRELMWK KEKEQLLKEK ELRQKALDKE RMVSSKPETN PKTPISLKEL
121 KDIEIYQNQF HYSAYEILEE EKILDNIFRK FTALPIKYWP ATPIRG = 166 aa
```

>SpGemin7, UniprotKB G2TRR1

```
  1 MAENNKKSTA YIQRMRTLKF YQKMASARIP ITVYLHDQRE VKAEFGAIDS SESKLAVSNL
 61 QTDWGVINRA VIRTGDVVGI EYNLVQEEGE L = 91 aa
```

>SpGemin6, UniprotKB G2TRK8

```
  1 MDSHTTEKRR GSYLRVLFKN GRLPVEGFLW NCDPLTGTLI LIQPLTPNTD EVEDTHYRFY
 61 GIMSDAIQTI EPDESMSPLN QQSLAEYDSL LTS = 93 aa
```

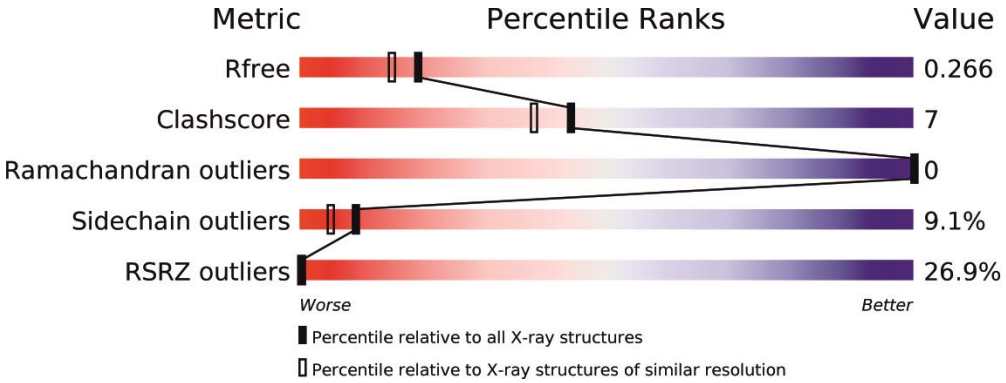


Figure 8.1: Percentile score for global validation metrics for SpSMN$^{\Delta 36\text{-}119}$ structure

The validation metrics were obtained using wwpdb.org/validation/validation-servers. Final resolution of the structure was at 2.16 Å.

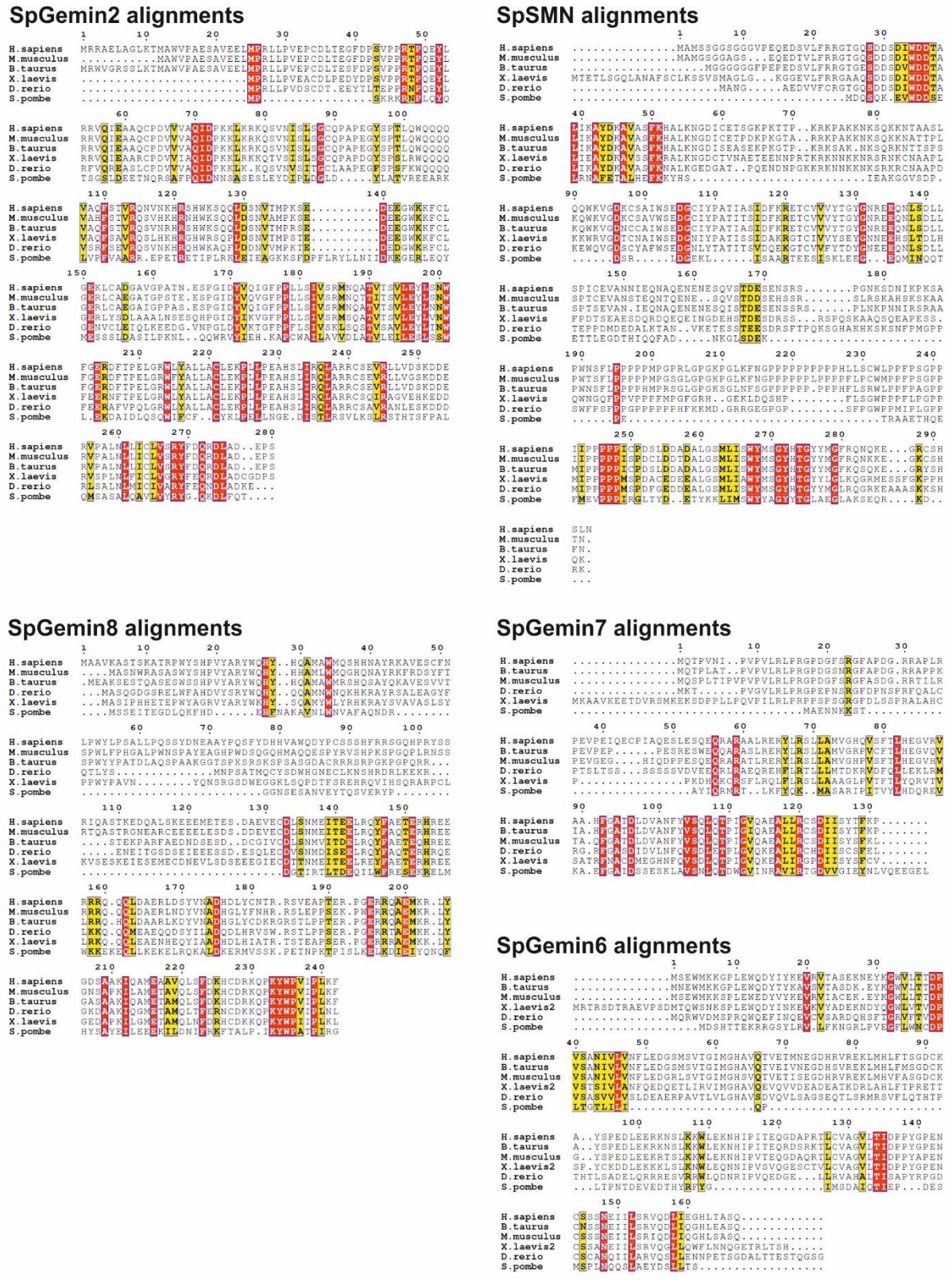


Figure 8.2: Multiple sequence alignments
Multiple sequence alignments of SpGemin2, SpSMN, SpGemin8, SpGemin7 and SpGemin6 with orthologs from human, *M. musculus, B. taurus, X. laevis, and D. rerio.* The alignments were performed using ESPript webtool.

Table 8.2: Crystallographic data for SpSMN$^{\Delta 36\text{-}119}$ structure

Wavelength (Å)	0.9762
Resolution range	37.09 - 2.158 (2.235 - 2.158)
Spacegroup	C2221
Unitcell	a=27.19, b=83.71 c=160.06; α=90, β=90, γ=90
Total reflections	67776 (7057)
Unique reflections	10282 (1010)
Multiplicity	6.6 (7.0)
Completeness(%)	0.99 (1.00)
MeanI/sigma(I)	10.85 (2.25)
Wilson B-factor	46.47
R-merge	0.1046 (0.824)
R-meas	0.1142 (0.8907)
R-pim	0.04472 (0.3344)
CC1/2	0.998 (0.714)
CC*	0.999 (0.913)
Reflections used in refinement	10263 (1010)
Reflections used for R-free	531 (31)
R-work	0.2217 (0.3705)
R-free	0.2663 (0.4644)
CC(work)	0.962 (0.803)
CC(free)	0.947 (0.372)
Number of non-hydrogen atoms	982
macromolecules	892
ligands	88
solvent	2
Protein residues	108
RMS (Å) (bonds)	0.007
RMS (angles)	1.41
Ramachandran favored (%)	99
Ramachandran allowed (%)	0.96
Ramachandran outliers (%)	0
Rotamer outliers (%)	8
Clashs core	6.61
Average B-factor	77.5
macromolecules	76.24
ligands	90.76
solvent	58.41
Number of TLS groups	8

Table 8.3: SAXS data collection parameters

Beamline	ESRF BM29 BioSAXS
Beam geometry	0.7 mm x 0.7 mm
Wavelength (Å)	1.54
s-range ($Å^{-1}$)	0.0032 - 0.4944
Exposure time (sec)	1 per frame, 1800 frames
Concentration range	SEC-SAXS analysis (Superdex 200 10/300)
Temperature (K)	293
Primary data reduction	BM29 online data analysis, PRIMUS
1D data processing	ATSAS 3.0 package: CHROMIXS, PRIMUS, GNOM
3D graphics representations	PyMOL, Chimera

9. ABBREVIATIONS

Å	Angstrom
AEBSF	4-(2-aminoethyl)benzenesulfonyl fluoride hydrochloride
APS	Ammonium persulfate
BV	Bed volume
CryoEM	Cryo-electron microscopy
CV	Column volume
Dmax	Particle maximum dimension
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
EOM	Ensemble optimization method
-FL	Full length
GFP	Green Fluorescent Protein
His-	Hexa-histidine tag
I(0)	Zeroth angle intensity
IDP	Intrinsically disordered Protein
kDa	Kilo dalton
LB	Luria Bertani
mAU	mili-absorption unit
MW	Molecular weight
Ni-NTA	Nickel Nitrilotriacetic acid
nm	Nanometer
P(r)	Pairwise distance distribution function
PCR	Polymerase chain reaction
pICln	Chloride conductance regulatory protein ICln
PISA	Proteins, Interfaces, Structures and Assemblies
PMSF	Phenylmethylsulfonyl fluoride
PPG	Polypropyleneglycol
PRMT5	Protein Arginine Methyl Transferase 5
Qp	Porod invariant Q
Rg	Radius of gyration
RNA	Ribonucleic acid
s	Angular momentum
SAXS	Small angle X-ray scattering

sDMA	symmetrically dimethylated arginine
SDS	Sodium dodecyl sulphate
SDS-PAGE	SDS Polyacrylamide gel electrophoresis
SEC	Size Exclusion Chromatography
Sm	Smith Antigen
SMA	Spinal Muscular Atrophy
SMN	Survival of Motor Neuron
SpG2	*S. pombe* Gemin2
SpG6	*S. pombe* Gemin6
SpG7	*S. pombe* Gemin7
SpG8	*S. pombe* Gemin8
SpSMN	*S. pombe* SMN
TAE	Tris acetate EDTA
TB	Terrific Broth
TEMED	Tetramethylethylenediamine
TEV	Tobacco etch virus protease
UsnRNA	Uridine-rich small nuclear ribonucleic acid
UsnRNP	Uridine-rich small nuclear ribonucleoprotein
Vc	Volume of correlation
Vp	Porod volume
WD45	WD repeat-containing protein 77 or MEP50
X^2	Chi^2

10. PUBLICATIONS

Vendelova E, Camargo de Lima J, Lorenzatto KR, Monteiro KM, Mueller T, **Veepaschit J**, Grimm C, Brehm K, Hrčková G, Lutz MB, Ferreira HB. Proteomic analysis of excretory-secretory products of Mesocestoides corti metacestodes reveals potential suppressors of dendritic cell functions. PLoS neglected tropical diseases. 2016 Oct 13;10(10):e0005061.

Veepaschit J,* Viswanathan A,* Bordonne R,* Grimm C, Fischer U. SMN complex is a multivalent hub for assembly of RNA-protein complexes: structural insights from the human and *S. pombe* SMN complexes (manuscript in preparation).

* These authors contributed equally to this work.

11. CURRICULUM VITAE

Jyotishman Veepaschit
Male; Born Nov 12, 1988
Nationality: Indian
E-mail: Jyotishman.Veepaschit@uni-wuerzburg.de

Current position:
PhD student in the laboratory of Prof Dr Utz Fischer, Department of Biochemistry, University of Würzburg, Germany.

Education:
2014-present: PhD student in the laboratory of Prof Dr Utz Fischer, Department of Biochemistry, University of Würzburg, Germany.

2011 – 2014: Master of Science, Master Degree Program of Protein Science & Biotechnology, Department of Biochemistry, University of Oulu, Oulu, Finland.

2006 – 2010: Bachelor of Technology in Biotechnology, Department of Biotechnology, SRM University, Chennai, India.

Research experience and technical skills:
Molecular biological techniques; packing and optimization of chromatography columns; affinity, ion-exchange, size-exclusion chromatographic techniques; crystallization and protein structure analysis; preparation of negative staining copper grids; sample preparation, optimization and screening of CryoEM samples by negative staining (Tecnai T12, 120 keV); sample preparation and data collection of CryoEM samples (Titan Krios, 300 keV); CryoEM data processing using Relion 3.0. Small-angle X-ray scattering data collection and processing using ATSAS 3.0.

Conferences and Workshops attended:
MicroScale Thermophoresis (Workshop) at the Graduate School of Life Sciences, University of Wuerzburg, Germany
10^{th} - 11^{th} July 2014

6th Murnau Conference on Structural Biology, Poster presentation
Title: “Towards the structural characterization of the UsnRNP assembly machinery”
14^{th} - 17^{th} September 2016

24th Annual Meeting of the German Crystallographic Society (DGK), Poster presentation
Title: “UsnRNP assembly machinery of Schizosaccharomyces pombe”
14^{th} - 17^{th} March 2016

Eureka! 10th International GSLS Students Symposium, Poster presentation
Title: “Biochemical analysis of newly discovered Gemins complete the central core of the SMN complex in Schizosaccharomyces pombe”
14^{th} - 15^{th} October 2015

Project Management (Virtual Workshop) at the Graduate School of Life Sciences, University of Wuerzburg, Germany
14th - 15th April 2020

Publications and Manuscripts in preparation:
Vendelova E, Camargo de Lima J, Lorenzatto KR, Monteiro KM, Mueller T, **Veepaschit J**, Grimm C, Brehm K, Hrčková G, Lutz MB, Ferreira HB. Proteomic analysis of excretory-secretory products of Mesocestoides corti metacestodes reveals potential suppressors of dendritic cell functions. PLoS neglected tropical diseases. 2016 Oct 13;10(10):e0005061.

Veepaschit J,* Viswanathan A,* Bordonne R,* Grimm C, Fischer U. SMN complex is a multivalent hub for assembly of RNA-protein complexes: structural insights from the human and S. pombe SMN complexes (manuscript under preparation).
* These authors contributed equally to this work.

Place, Date:

Signature:

12. ACKNOWLEDGEMENTS

First and foremost, I thank my God and Savior for His love in saving a wretch like me, giving the knowledge of His existence & pursuing my wayward heart to make me His.

I am immensely grateful to Prof. Utz Fischer for accepting me into his research group as a PhD student. During the 6 years of my work, he entrusted me with much of the laboratory and with many projects, even with some of the crucial experimental techniques central to our laboratory. I am also thankful for his tremendous patience with me, his positivity, optimism, and unrestrained support for all my ideas and research endeavors.

I convey my heartfelt gratitude to Dr Clemens Grimm for accepting me into his laboratory as a Master thesis student back in 2013, and then as a PhD worker into the crystallography group. I am also thankful to him for giving me so much freedom to carry out all my experiments and entrusting me with much of the laboratory and visiting students. I am thankful to him for giving me the opportunity to supervise practical sessions, especially the chromatography practicals. I love chromatography, and this was like a dream come true for me. Coming from a small village in India which no one has ever heard of, I got opportunities to visit one of world's brightest synchrotrons (ESRF) frequently with Clemens, and carryout cutting edge atomic research. Also, Clemens, your protein expression strategy is the best.

I am immensely thankful to Prof. Thomas Müller for agreeing to be one of my supervisors. I am grateful to him for his helpful insights into my results and great discussions during the few times I got to see him.

I thank Dr Farah Badbanchi-Fischer for all that she has done and still does for me. I thank her for her hard work and relentless efforts to provide all of us with the best resources and opportunities.

I thank the Graduate School of Life Sciences of the University of Wüzburg (especially Dr. Blum-Oehler) for their tremendous help and support throughout my PhD.

I express my heartfelt gratitude and respect for my supervisor turned colleague Dr Jann-Patrick Pelz, who taught me literally everything in the laboratory. Although I pestered him much with my constant questioning (they used to call me "der Kleine"), he always had patience with me. He always trusted me and wanted the best for me. I wish him the best for his life.

I express my heartfelt gratitude to Emilia Gärtner who takes care of our laboratory in the best way possible. I thank her for her hard work in going beyond her abilities to prepare all that we need in the laboratory. I thank her for her tremendous support during practicals. I thank her for doing so much for me and others, and just for being there all these years.

I greatly acknowledge the tremendous effort put in by some of my students, without whom I would not have learned anything myself: Yannic Lurz for helping me purify literally grams of proteins for SAXS, so much that we could not process; Maria Schultz for SUMO-Gemin8 work and for being so perfect in following my lead; Maria Gonzalez for pombe Sm protein work, and just for being such a nice pupil of mine; Nilofar Feizy and Leandro Buhlmann for PRMT5 work; Julia Karius and Simon Mungwa for some cloning and purifications.

I thank my former colleague Dr Aravindan Viswanathan for his company, for his help whenever I needed, and for creating some funny situations in the lab along with Dr Christina Plank and Dr Raji Meduri, who I also remember and thank for their company.

I thank Dr Archana Prusty for introducing me to polycistronic cloning which became a crucial tool for all my investigative work in this project. All this would not have been possible without that. I also thank her for the many insightful discussions over the years.

Words are not enough to describe how much I thank and appreciate the relentless support I received from Dr Isotta Lorenzi during my thesis writing. I thank her for all the suggestions, ideas, corrections, and most of all for her moral support and company.

I thank Dr Julia Bartuli for trusting me with all her CryoEM samples, from sample prep to data collection. I will always cherish our visits to the Café, Vitrobot, Bubi, and Berta. I also thank her for her support during my writing.

I thank Prof. Bettina Böttcher and her group members Tim, Cihan, Sam, Vanessa, and Christian, for teaching me and letting me use their Titan Krios, and trusting me with their expensive equipment.

I am thankful to my current (and some former) colleagues Michael, Georg, Bettina, Tanja, Jürgen, Hannes, Cornelius (also for translation), Manisha, Pradhipa, Clemens, Conny, Anja, Georg, Christina, Maritta, Florian, Tommy, Sanjay, Ankit, Christopher, Mona, Claudia, Sven, Lissy, Susanne, Andrea, Karl, Sonja, and all those who I didn’t name or forgot to do so, for providing a friendly working environment.

Finally, I thank my dad, mom, aunt, sister, and my brother-in-law, for standing by me, encouraging me, praying for me, and for wanting the best for me. They are the best people I could ever have in my life. I also thank members of my church both in Würzburg and at home, and all those who constantly prayed for me.