Biophysical characterization of squid' s sucker ring teeth proteins

Kumar Dhawan Akshita

2017

Kumar Dhawan Akshita. (2017). Biophysical characterization of squid' s sucker ring teeth proteins. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/69647>

<https://doi.org/10.32657/10356/69647>

NANYANG
TECHNOLOGICAL
UNIVERSITY

**Biophysical Characterization of Squid's Sucker
Ring Teeth Proteins**

KUMAR DHAWAN AKSHITA
SCHOOL OF BIOLOGICAL SCIENCES
2016

Biophysical Characterization of Squid's Sucker Ring Teeth Proteins

KUMAR DHAWAN AKSHITA

SCHOOL OF BIOLOGICAL SCIENCES

A thesis submitted to Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2016

Acknowledgement

Like great works, deep feelings always mean more than they are conscious of saying. With this thought, sincere gratitude and appreciation I would like to thank the following individuals for providing me with immense support, advice and constant encouragement throughout the process of developing this thesis. Thank you all for everything you did in your power to assist me in the most ambitious endeavor of my career thus far.

I would like to thank my three supervisors; each being an expert in his own field has helped in multiple ways to develop many facets of this piece of work (and myself). Julien, for giving me the opportunity to do PhD under his esteemed guidance, “introducing” me to the world of *Biomimicry* and nurturing my scientific outlook. I’m thankful for your understanding and the freedom that you provided me with, that allowed me to develop and learn at my own pace. Ali, for allowing me to “explore” the world of *Biomimicry* with him and his wonderful group. Your dedication and enthusiasm for scientific research is unsurpassed and the vast knowledge and incisive insight has been inspiring for all of us. You’ve been exceptionally patient with me to develop the project and my experimental skills gradually, more so because of my entirely computational background. Thank you for believing that I could reach this far. Chandra I can’t thank you enough for your superb guidance all through these years, from my masters till the present. You’ve guided me through many crises of confidence, specifically when I felt that I had nothing to contribute in this project or when I thought that I had hit a dead-end. Thanks for gently guiding me always back on the right path. Through many discussions you’ve instilled in me the ability to appreciate the big picture and the minute details of any complex problem, all at the same time.

At many points in the thesis you will come across the word “we”. This pronoun is used to refer to my supervisors and mentors whom I owe an enormous debt of gratitude in shaping up this work.

Learning so much required help and guidance from many experts apart from my supervisors. My earnest thanks to Paul, for guiding me and building up the foundation of this project; Shawn, for teaching me the nuances of gene cloning and giving me the wonderful opportunity to work with him; Harini, for helping me at every step to derive the NMR structure of the protein; Ragav, for guiding me with MD simulations, for his critique and being patient to resolve unforeseen errors, I would often bump into. The discussions with you all really helped to conceptualize the ideas and crystallize my thought process.

I would also like to extend my thanks to all the members of Julien, Ali and Chandra's group for their friendship and providing a wholesome conducive working environment. My appreciation to Shu Hui, for her care, affection and quiet understanding; Shahrouz for always setting up high standards of research; Dawei, for teaching me the basic lab skills right from holding the pipette to protein expression; Fu Jing, for her ever positive attitude and help with TEM; Dario for helping me set up my first crystal screen and sharing the excitement of it; Jun Jie, for teaching me how to make the best SDS-PAGE, putting things in order and making sure they work; Maryam, for the will strive; Clarinda, for her very calming spirit; Abbas, for keeping me in high spirits; Cai Hao, for his humor; Deepan, Bartoz and Ping Yuan for many valuable discussions. My gratitude to Konstantin for agreeing to be on my committee and for helpful feedbacks in designing the experiments. I acknowledge School of Biological Sciences, NTU and Ministry of Education Tier 2 Grant for funding my project.

I feel fortunate to gratefully acknowledge the support of some wonderful friends here. Thanks Kini for sharing this grad journey, which was no less than a roller coaster ride. You've offered unlimited encouragement every single day. Thank you for your friendship and all the good times we've spent. Thanks Dilraj for not answering my scientific queries (most of the times), forcing me to learn things by exploring myself and of course for your constructive criticism. Thanks Garima for your generous care throughout the research tenure. Thank you guys for being besides me and for the unstinting support during happy and hard moments to motivate me.

I know that I couldn't have completed this project without the constant support of my Mom and Dad. You both offered me with heap loads of moral support (like always). You have always encouraged me and expressed confidence in my abilities when I felt I lagged them the most. Thank you for showing faith in me and giving me the liberty to choose what I desired even when I was pursuing a journey across uncharted seas. You both keep me sane and well grounded and help put things in perspective always. Thank you for everything and much more! One might say parents are such and it's their job, well if it is, then you both are best at it. Love you both to bits. Special thanks for my unbelievably supportive husband, Lakshya. You've always demonstrated rare and amazing patience level with my anxiety attacks, multiple stresses, long and unearthly working hours, often hearing - rather too often "not right now!" and me as such. Thank you for your unfailing love, support, and understanding during these years, for believing in me and sharing my wish to complete this project and yet caring enough to love me. Thank you, my three guardian angels!

Thanks to Albert Camus for letting me know that: *"For if there is a sin against life, it consists perhaps not so much in despairing of life as in hoping for another life and in eluding the implacable grandeur of this life"*. If I may be permitted to take liberty with this quotation and replace the word "life" with "thesis", this to me sounds appropriate here. *Thus my gratitude to this work and thesis itself!*

*To the Energy
that sustains life*

Table of Contents

Acknowledgement	i
Table of Contents	v
List of Figures.....	x
List of Tables	xiii
Abbreviations	xiv
Abstract.....	xvi
1 Literature Review	1
1.1 Biomimicry	2
1.2 Proteins: The elementary building blocks of life.....	2
1.3 Proteins in Biological Materials	3
1.3.1 Structure-Property Paradigm	3
1.3.2 Significance of Hierarchy	5
1.4 Biological Protein Materials	7
1.4.1 Silk	8
1.4.2 Collagen	11
1.4.3 Elastin	13
1.4.4 Mussel Byssus	16
1.5 From Natural materials to tailored materials	19
2 Introduction to Squid's Natural Bio-tool: The Sucker Ring Teeth.....	21
2.1 Dosidicus gigas	21
2.2 Un-mineralized Extracellular Hard Tissues in Squid	22
2.3 Unique Features of Jumbo Squid's SRT	23

2.3.1	Micro-anatomy of suckers	23
2.3.2	Mechanical Properties of SRT	24
2.3.3	Biochemical Composition of SRT	25
2.3.4	The Suckerin Protein Family	26
2.3.5	Molecular Architecture of SRT	28
2.3.6	Contribution of β -sheets towards SRT's mechanical properties.....	32
2.3.7	Thermoplastic properties of SRT	33
2.4	Bio-materials from SRT and Suckerin	35
2.5	Objective and Scope of the Project.....	36
3	Methods Applied To Characterize The Structure of Suckerins.....	38
3.1	X-ray diffraction on native SRT fibers	38
3.2	Cloning, expression and purification of suckerin proteins	39
3.2.1	Suckerin-19	39
3.2.2	Sepio Monomer.....	40
3.3	Identification of Recombinant Proteins	43
3.3.1	Polyacrylamide Gel Electrophoresis.....	43
3.3.2	MALDI-ToF	44
3.4	Determination of Protein Concentration.....	45
3.5	Refolding of Suckerin Proteins.....	45
3.6	DLS Measurements	47
3.7	CD Spectroscopy	48
3.8	FTIR Spectroscopy	49
3.9	TEM.....	50
3.10	Theoretical Suckerin Protein Structure Prediction	51
3.10.1	Secondary Structure Prediction.....	51
3.10.2	Tertiary Structure Prediction.....	52

3.10.3	Model refinement using MD Simulations.....	54
3.11	X-ray Crystallography	61
3.11.1	Crystal Screening	61
3.12	NMR spectroscopy	62
3.12.1	One dimensional NMR spectroscopy	62
3.12.2	Heteronuclear Correlation Spectroscopy	63
3.12.2.1	Two Dimensional ¹⁵ N-HSQC	63
3.12.2.2	Three Dimensional Heteronuclear NMR.....	64
3.12.3	NMR acquisition and processing	66
3.12.4	Structure calculations.....	66
3.12.4.1	Automated structure calculation by CYANA-FLYA	67
3.12.4.2	Secondary structure chemical shifts	67
3.12.5	Hydrogen/Deuterium exchange strategy.....	68
4	Biophysical Characterization of SRT derived fiber and Suckerin protein. .	69
4.1	Introduction.....	69
4.2	Results and discussion	72
4.2.1	Biophysical characterisation of processed SRT fibers.....	72
4.2.2	Biophysical Characterization of Suckerin-19	73
4.2.2.1	Primary Structure Analysis.....	73
4.2.2.2	Recombinant Expression and Purification.....	74
4.2.2.3	Solubility, Refolding and Secondary Structure Characterization.....	75
4.2.2.4	NMR Characterization	81
4.2.2.5	TEM characterization	84
4.2.3	Design and characterization of Suckerin-19 constructs.....	85
4.3	Conclusion	87

5	Theoretical Structure Prediction of Suckerin-19 protein	89
5.1	Introduction.....	89
5.2	Results and Discussion	91
5.2.1	Secondary Structure Prediction.....	91
5.2.2	Peptide Simulations	92
5.2.3	Full Length <i>Ab initio</i> Modeling	99
5.2.4	Model refinement with MD	102
5.3	Conclusion	106
6	Biophysical Characterization of Sepio Monomer	107
6.1	Introduction.....	107
6.2	Results and Discussion	109
6.2.1	Cloning, Expression and Purification of Sepio monomer	109
6.2.2	Renaturation of Sepio monomer	110
6.2.3	Preliminary Spectroscopic Characterization.....	111
6.2.3.1	Dynamic Light Scattering.....	111
6.2.3.2	CD Spectroscopy	112
6.2.3.3	FTIR spectroscopy.....	113
6.2.3.4	¹ H-NMR spectroscopy.....	114
6.2.4	Crystal Screening.....	116
6.2.5	NMR Spectroscopy	117
6.2.5.1	Backbone and side-chain NMR assignment	117
6.2.5.2	NMR structure calculation and validation	119
6.2.5.3	Structure of Sepio monomer	119
6.2.5.4	Hydrogen/deuterium exchange	124
6.2.5.5	Secondary Chemical Shifts.....	126
6.3	Conclusion	128

Conclusion and Perspective	129
Bibliography	135
Publications	154
Appendix A	155
Appendix B	164

List of Figures

Chapter 1

Figure 1.1. Different levels of protein structural organization.	2
Figure 1.2. Structure-Property-Process relationship in silk biopolymer design.	4
Figure 1.3. Hierarchical self-assembly of amyloids at structural level and in materials..	7
Figure 1.4. Formation of geometrically confined structures in silk.....	9
Figure 1.5. Structure of collagenous tissues..	12
Figure 1.6. Schematic representation of the formation of elastin fibers.	15
Figure 1.7. Schematic view of protein organization in mussel byssus.	18
Figure 1.8. Sequence-Structure-Property-Process relationship of silk, collagen and elastin.	20

Chapter 2

Figure 2.1. <i>Dosidicus gigas</i> and its geographical distribution.....	22
Figure 2.2. Representations of <i>D. gigas</i> sucker ring teeth..	23
Figure 2.3. Microstructural features of <i>D. gigas</i> SRT	24
Figure 2.4. Micro-mechanical properties measured on <i>D. gigas</i> SRT cross-section.	25
Figure 2.5. Chemical and elemental decomposition of <i>D. gigas</i> SRT.....	26
Figure 2.6. Primary structure organization in suckerin protein family.....	27
Figure 2.7. Characterization of protein structure in <i>D. gigas</i> SRT by Raman spectroscopy and WAXS.	30
Figure 2.8. Characterization of native SRT by WAXS and SAXS.	31
Figure 2.9. Correlation between the mechanical properties and secondary structure of <i>D. gigas</i> SRT.....	32
Figure 2.10. Influence of temperature on SRT structure and mechanical properties	34

Chapter 4

Figure 4.1. Structural features of SRT as obtained by <i>in situ</i> X-ray scattering	70
Figure 4.2. Biophysical characterization of suckerin-19 in solution	71
Figure 4.3. Characterization of native SRT fiber using WAXS and SAXS	73
Figure 4.4. Primary sequence and recombinant expression of suckerin-19	75
Figure 4.5. Preliminary spectroscopic characterization of suckerin-19 in 50 mM MES at pH 5.5.....	76
Figure 4.6. Temperature dependent CD spectrum for suckerin-19 in solution..	79
Figure 4.7. Schematic oligomerization state and structure of suckerin-19.....	80
Figure 4.8. Secondary structure characterization and particle size distribution of suckerin-19 in various concentrations of MES buffer at pH 5.5.	81
Figure 4.9. 1D NMR spectrum of Suckerin-19.....	82
Figure 4.10. ¹⁵ N-HSQC spectra of suckerin-19 in 50 mM MES, pH 5.5	83
Figure 4.11. TEM micrographs of suckerin-19	85

Chapter 5

Figure 5.1. Consensus secondary structure predictions.	92
Figure 5.2. Local β -sheet conformations adopted by suckerin-19 (BP1-BP8) peptides, predicted using MD simulations.	95
Figure 5.3. Secondary structure conformations of suckerin-19 (BP1-BP8) peptides predicted during the MD simulation.	96
Figure 5.4. Local conformations of suckerin-19 peptides (SP1-SP4) with short Gly-rich domain	97
Figure 5.5. Structure acquired by assembling SP1+SP2 peptide depicting β -sheet conformation.....	98
Figure 5.6. Tertiary structure of suckerin-19 predicted using I-TASSER.....	102

Figure 5.7. Structural evolution of IT1-model observed during MD simulation.....	103
Figure 5.8. Characterization of secondary structure by FTIR and elastic modulus in suckerin-19 films.	104
Figure 5.9. Characterization of Tyrosine residues in IT2-model MD simulations..	105

Chapter 6

Figure 6.1. Primary structure features and material properties of sepio constructs.	108
Figure 6.2. Recombinant expression of sepio monomer.....	110
Figure 6.3. Size exclusion chromatogram of sepio monomer.	111
Figure 6.4. Spectroscopic characterization of sepio monomer	112
Figure 6.5. ATR-FTIR analysis of sepio monomer.	114
Figure 6.6. 1D NMR spectrum (amide region) of sepio monomer.....	115
Figure 6.7. ¹⁵ N-HSQC spectrum for sepio monomer	118
Figure 6.8. NOE connectives of sepio monomer in 20 mM sodium acetate pH 5.0	120
Figure 6.9. Ramachandran plot of sepio monomer.....	122
Figure 6.10. Tertiary structure of sepio monomer in 20 mM sodium acetate pH 5.0, in cartoon representation determined using NMR..	123
Figure 6.11. H/D Ex of sepio monomer in deuterated 20 mM sodium acetate, pH 5.0 recorded at 298 K and various time steps.	125
Figure 6.12. Secondary chemical shifts (SCS) plot for residues in sepio monomer derived from the ¹³ C α and ¹³ C β chemical shifts.	127

List of Tables

Chapter 3

Table 3-1. Buffers and their respective volumes used for gel casting.	43
Table 3-2. Annotated list of buffers used to obtain soluble suckerin-19 protein.....	46
Table 3-3. Annotated list of buffers used to obtain soluble sepia monomer	46

Chapter 4

Table 4-1. Constructs derived from suckerin-19 and other suckerin proteins cloned and expressed.....	86

Chapter 5

Table 5-1. Modular units that assemble to form suckerin-19 on which MD simulations were performed.	94
Table 5-2. Modular units with reduced Gly-rich domains on which MD simulations were performed.	96

Chapter 6

Table 6-1. Estimated secondary structure of sepia monomer from BeStSel fitting of experimental CD spectrum.	113
Table 6-2. List of backbone and side-chain NOEs	119
Table 6-3. Structural statistics of sepia monomer.	121

Abbreviations

1D	One dimensional
2D	Two-dimensional
3D	Three-dimensional
AAA	Amino acid analysis
aa	Amino acid
AFM	Atomic force microscopy
Ala	Alanine
APS	Ammonium persulfate
ATR	Attenuated total reflection
<i>B. mori</i>	<i>Bombyx mori</i>
CBR	Coomassie blue R
CD	Circular dichroism
<i>D. gigas</i>	<i>Dosidicus gigas</i>
DLS	Dynamic light scattering
DNA	Deoxy ribonucleic acid
DSC	Differential scanning calorimetry
<i>E</i>	Young's modulus or elastic modulus
<i>E. coli</i>	<i>Escherichia coli</i>
EDX	Energy-dispersive-X-ray spectroscopy
FPLC	Fast protein liquid chromatography
FTIR	Fourier transform infrared spectroscopy
Gly	Glycine
<i>H</i>	Hardness
H/D-Ex	Hydrogen-Deuterium Exchange
His	Histidine
HPLC	High pressure liquid chromatography
IB	Inclusion bodies
IPTG	Isopropyl β -D-1-thiogalactopyranoside
LB broth	Luria-Bertani broth
LCST	Lower critical solution temperature
Leu	Leucine

MALDI-ToF	Matrix- assisted laser desorption ionization time-of flight
MD	Molecular dynamics
MES	2-[N-Morpholino]ethanesulfonic acid
Mins	Minutes
MW	Molecular weight
NMR	Nuclear magnetic resonance
PCR	Polymerase chain reaction
PBS	Phosphate buffer saline
PEEK	Polyether ether ketone
pI	Isoelectric point
PMMA	Poly(methyl methacrylate)
PMSF	Phenylmethysulfonyl fluoride
Pro	Proline
RNA-seq	RNA sequencing
SAXS	Small-angle X-ray scattering
SDS-PAGE	Sodium dodecyl sulfate-polyacrylamide gel electrophoresis
<i>S. esculenta</i>	<i>Sepio esculenta</i>
SEC	Size exclusion chromatography
Ser	Serine
<i>S. lessoniana</i>	<i>Sepioteuthis lessoniana</i>
SRT	Sucker ring teeth
TEM	Transmission electron microscopy
TEMED	Tetramethylethylenediamine
Thr	Threonine
Tyr	Tyrosine
Val	Valine
WAXS	Wide-angle X-ray scattering

Abstract

Sucker ring teeth (SRT) are natural biopolymers entirely composed of proteins (called “suckerins”) exhibiting a block-copolymer-like structure. Despite being stabilized exclusively through the cooperative action of weak interactions, they feature exceptional mechanical properties that rival those of robust engineered polymers. Suckerins share sequence homology to silk and adopt nano-confined β -sheets embedded within an amorphous Gly-rich network. This study combines experimental and computational investigations to identify suckerin protein structure, and bridge it with function/mechanical properties. We illustrate the structural organization of suckerin assembly in processed SRT fibers. We further characterize the structural morphology of the suckerin-19 and its derived construct using different biophysical methods. The conformational propensity of specific sheet-forming domains is predicted using computational modeling and validated using NMR spectroscopy. We elucidate the influence of primary sequence on the β -sheet enriched structures and on the dynamic amorphous domains, to speculate the origin of the mechanical properties.

1 Literature Review

Earth has harbored an increasing diversity of life forms which have evolved for ~3.8 billion years. In this respect evolution has played and continues to play a decisive role in selection and development. Failures are fossils and what surrounds us is the secret to survival. From soft unicellular organisms, life gradually evolved into more complex multi-cellular animals and plants [1]. Since Nature abhors wasting energy, sophisticated materials and efficient structures have evolved under the unforgiving hand of natural selection over successive generations [2, 3]. Nature provides a wide range of extra cellular biological materials, which are multifunctional as they combine biological, mechanical, and other distinctive functions that may serve as a source of bio-inspiration to design solutions for a given set of requirements and constraints. And thus, even with a limited cocktail of building blocks, organisms managed to produce a trove of ingenious materials that work with unparalleled efficiency, simplicity, and durability [3, 4]. Living creatures around us constitute a mixture of biological materials and because of these materials cells function, plants stand tall, eyes capture and interpret light, while animals/birds move and fly.

Separating the mechanical from the biological functions of such natural materials and comparing their material property charts with synthetic materials using so called Material Selection Charts (also called the “Ashby” maps) significant features can be observed in their performances and mechanical properties [5]. Many man-made materials like high performance ceramics and alloy display high strength and toughness when compared to the best natural materials. However, the exceptional silk displays extraordinary toughness and elastic modulus of 10 GPa, similar to kevlar, suggesting that such natural materials can also have spectacular properties, generally similar to synthetic materials [6]. Moreover, comparing the structural design of synthetic vs. natural materials, it is worth noting that all biological materials are composed of basic ingredients made with a limited chemical palette processed at ambient temperatures with minimum energy requirements [7].

1.1 Biomimicry

As realized, Natural materials provide many opportunities to learn from the biological world *e.g.*: on hierarchical structuring, growth and functional adaptation, self-healing or damage repair. When scientists use Nature as their model, characterize it and eventually aim to duplicate it for engineering it is called “*Biomimicry*” [8, 9]. Through careful observations and dedicated studies, biomimetic researchers unfold the fundamental, ingenious and intricate principles underlying natural phenomena. They derive inspiration from nature’s elegance, its time tested patterns and strategies, to sustainably engineer (or mimic) the form or function of natural materials and their designs [9]. Thus, a serendipitous observation of nature followed by a systematic approach involving the study of natural materials and a thorough application of engineering principles can be efficiently applied to further develop bio-inspired concepts that form the fundamentals of biomimetic research.

1.2 Proteins: The elementary building blocks of life

Proteins are widely known as the building blocks of life, which are assembled in the cell. They form a wide range of biological load bearing structures from cells to silk, from bone to tendon, or skin to name just a few. Protein based materials are found in abundance in biological systems and perform some crucial functions or processes inside and outside of the cells and tissues [10]. Structurally, proteins can be organized into four distinct levels [10] as shown in figure 1.1.

Figure 1.1. Different levels of protein structural organization. Primary structure, which is a linear chain of amino acids connected through amide bonds; the secondary structure is repetitive local conformations (most commonly, α -helices and β -sheets); the tertiary structure comprises of the overall fold or the spatial arrangement of the polypeptide chain and the quaternary structure which determines the specific assembly of multiple polypeptide chains. Adapted from Ref. [11]

1.3 Proteins in Biological Materials

Commonly, biological protein materials can be distinguished based on their spatial location within their associated tissue *i.e.*, they either serve as intracellular or extracellular load bearing elements. Cells secrete extracellular proteins like silk, collagen, and elastin into the surrounding microenvironment. Intracellular protein materials provide structural support and architecture to cells, including actin, microtubules or the intermediate filaments vimentin that forms cells cytoskeleton or the lamin that forms the cell's nuclear envelope. There are other notable protein based materials that appear during specific biological processes such as fibrin during blood clotting [10, 12-15]. All protein materials virtually show hierarchical architectural organization. While some structural features are commonly seen and conserved in different tissues (e.g., α -helices, β -sheets or tropo-collagen molecules), others are highly specific (like tendon fascicles or β -sheet nano-crystals in silk) [10].

Recently, with the rising interest in biomimetics, the materials science of proteins is being explored with an aim to comprehend the underlying principles that relate the structure (of these biological proteins) and process (the manner in which proteins are arranged or assembles), which ultimately influences the property (strength, robustness, elasticity) of any biological material. Understanding the material properties of these hierarchical structures and studying their influence on the molecular and microscopic properties, using mechanistic insights on the basis of structure-process-property relationships in the biological context, can provide deeper understanding as well as help develop novel material design concepts.

1.3.1 Structure-Property Paradigm

Rigorous understanding of the fundamental elements of life and how they work through biological materials and systems can enable us to seamlessly integrate concepts from biology to material design. In order to develop functional materials, their synthesis, processing, control, implementation and analysis represents major scientific and engineering challenges. To extend development and performance of bio-inspired materials it is essential to bridge the “structure-process-property” gap, as breakthroughs can be accessed efficiently when the fundamentals are well understood. This paradigm has guided material science research for decades. As an example, α -helical domains commonly found in protein materials plays a critical role in

biophysical processes such as mechano- sensation or transduction and provide mechanical stability to cells (e.g. the intermediate filament networks) [16, 17]. Similarly, β -sheets are critical load-bearing building blocks of amyloid and spider silk proteins [18] (Figure 1.2).

Figure 1.2. Structure-Property-Process relationship in silk biopolymer design. The links depict relationships between the different fields that can be modulated to achieve the targeted properties. Adapted from Ref. [19]

For biological materials, various challenges still remain, specifically in order to develop rigorous understanding of structure-property relationships and its correlation with the mechanical response. For instance, in biological materials, bond energies are often comparable to thermal energy, as seen in hydrogen bonding, the most abundant chemical bond in biology. The prevalence of high viscoelastic behavior as observed in many biological materials is a response to mechanical deformation which is intrinsically time-dependent [20]. Such mechanical features and their correlation with the atomistic chemical reactions in protein materials is not only essential for the biological processes but also for *de novo* design and fabrication of bio-inspired materials [21]. From studies done in other fields of materials science, it is an established fact that nano or microscopic structural properties control the macroscopic behavior of materials. For example confinement or reduction in grain size strengthens crystalline metals [22, 23], whereas in biology hierarchical multi scale structures is the key to understand the mechanical response. Thus characterizing biological materials to

gain similar insights on the fundamental principles (like assembly or deformation) is an important frontier of research.

In order to elucidate the significance and essential role of small-scale features and their influence on macroscopic properties, an integrated experimental and simulation approach can be applied as a promising strategy to analyze these materials. While experimental methods have gained unparalleled accuracy, simulations have evolved as a predictive tool that can complement these experimental methods. Understanding from the small atomistic scales can be further extended up to macroscopic scales of entire tissues by explicitly comprehending structural features at each hierarchy with reference to cross-scale and inter-scale interactions. Even though this is challenging, successive studies with integration of experimental and computational work can provide promising potential and possibilities to trigger an optimized material design, transforming materials science as a discipline [19, 24].

1.3.2 Significance of Hierarchy

Hierarchy is observed in many different biological systems. Hierarchy in a system is defined to be composed of stable, perceptible sub-elements that merge by a superordinate relation [25]. In a complex hierarchical system the lower level features largely influence those at higher levels, which consequently have an impact on the behavior of the system. Hence, an interaction at and between various length scales is a crucial concept to be understood in a hierarchical system.

In Nature hierarchy is used to create multifunctional (ability to provide mechanical strength and control biochemical processes) materials and to overcome the intrinsic weakness of the building blocks. It takes the advantage of (sub) molecular features that are characterized by weak interactions. At the highest level of organization, the biomaterials display complex architecture, however at the fundamental level of hierarchy they are generally composed of simple, and few building blocks. Moreover the design strategy of these biological materials ranges from a choice of relatively fewer elements. For instance, nature uses lighter elements like C, N, O, H, Ca, Zn, Fe, S, P etc. for synthesis of bio-minerals and biopolymers like polymeric carbohydrates and proteins, which possess remarkable functional properties. The use of such few

building blocks together with simple/weak interactions requires limited energy to produce strong materials under benign conditions [3, 26]. One of the best described hierarchical system is the “hierarchy of life” wherein cells, tissues, organs, organisms, species, communities and other such entities are put together in an inclusive hierarchical relationship [25].

Another example is amyloids where hierarchy in biological materials has been exploited to create multifunctional materials. Amyloids have an inherent ability to form hierarchical structures, have been used to make multifunctional materials by exploiting their self-assembling properties (Figure 1.3). Diverse protein sequences can form amyloids once they lose their native configuration, forming fibers with typical hydrogen bonded cross β -sheet structure [27, 28]. This specific feature is universal to all amyloids (even though they have diverse protein sequences) and has been correlated with their higher stiffness, stability, and capacity to provide structural template at various length scales. The diverse sequence features provide different biochemical characteristics particular to a specific set of amyloids. Thus controlling various levels of hierarchy in amyloids can create multifunctional materials. At the basal level, protein sequences can be manipulated to target specific biochemical features. Moreover, chemical functionalization can lead to hierarchical structures that cannot be formed naturally. Assembly of amyloid fibrils in materials can be controlled at higher levels of hierarchy, by altering plasticizing agents to tune mechanical properties [28, 29].

Formation of hierarchy at varied length scales provides remarkable properties such as strength, support, robustness, and adaptability of natural materials (like bones, shells, tooth, sponges, silk and many others), which is intimately connected to the biological design paradigm of simple building blocks and bonds. Each level of hierarchy is thus a structural adaptation to maximize the target material properties. The knowledge of interaction of disparate scales in protein materials can be translated to essentially connect different material scales, from the nano- to the meso- to the macro-scale. Hence, these material concepts can be applied up to their full potential for multi scale engineering of biomaterials [3, 30, 31].

Figure 1.3. Hierarchical self-assembly of amyloids at structural level and in materials. (a) Schematic makeup at different length scales, (b) Experimental procedure to form self-assembled hierarchical materials like the films and the fibers, as shown, (c) Fluorescence microscopy image of amyloid protein fibrils casted into (i) non-functionalized and (ii) functionalized thin films aligned and stacked in the plane of films to form strong material. As shown, the optical properties of the films can be tuned by varying the alignment of amyloid fibrils at the meso-scale. Adapted from Ref. [32, 33].

1.4 Biological Protein Materials

Over billion of years, Nature has evolved a wide range of “engineering” designs that sometimes surpasses human technology. The abundance of smart, hierarchical biological materials, built from proteins that display self-healing and self-assemble into diverse forms autonomously using universal building blocks is one such example. In the past decades, scientists have worked to understand the design principles in specific proteins so as to replicate (or even surpass) the performance of the natural biological materials. The following sections describe in detail the underlying principles and mechanisms that have been unveiled for selected structures, which represent guiding models for the development of novel biomaterials.

1.4.1 Silk

General perspective

Silk is an exceptionally strong protein based polymer that is produced as fibers by several species of the Arthropods phylum, for a variety of specific tasks. For instance, silkworm cocoon silk fibers are commonly used in textile industry and for wound dressing, as they are strong and stiff [34, 35]. Of all the silks, the dragline silk is regarded as nature's high performance fiber as it has an exceptional combination of toughness, high tensile strength, and great extensibility. The viscid spider silk on the other hand features low Young's modulus with high extensibility [36, 37]. Hence, various types of silks are produced that exhibit versatile properties with different functions [37, 38].

Composition

Based on their specific natural function, silk proteins may vary in their composition, structure, and properties. For example in case of *B. mori*, the silk fibers are composed of two classes of proteins, namely sericin and fibroin. Sericins are hydrophilic proteins that conglutinate the silk fibroin fibers. Each fibroin is about 10-25 μm assembled by a hydrophobic heavy chain (~350 kDa) and a relatively hydrophilic light chain (~25 kDa). However, the dragline silk has a moderately different architecture composed of two major ampullate spidroin proteins i.e., MaSp1 and MaSp2 that are heterogeneously distributed to form the core filament which is further coated with glycoproteins and lipids [39].

Structural characterizations

Over the years different structural techniques, including X-ray, solution NMR, IR and Raman spectroscopy have been applied to study the characteristics and the architecture of this class of structural proteins in order to comprehend its remarkable mechanical properties and variations under different conditions [40-42]. Silk fiber proteins from spiders and silkworm are composed of repetitive residues (mainly Gly and Ala) with collaborative repeat lengths. The core primary sequence of silkworm is characterized by a hexapeptide GAGAGS, while the dragline silk has poly-Ala repeats flanked with GGX repeats (X: Tyr, Glu, Leu) [43-45]. The systematic conformation of hydrophobic blocks at the primary level leads to extensive hydrogen bonding and hydrophobic

interactions that further conform into crystalline β -sheet domains at the secondary structure level. Studies show that these highly conserved repeats form β -sheet nanocrystals impart strength and toughness to this biomaterial [6, 46-50]. These ordered antiparallel β -sheet regions are embedded between the semi-amorphous protein domains. The amorphous protein matrix introduces flexibility, specifically under high strain rate, as it allows the crystalline domains to reorient thereby increasing the strength of the material. The β -sheet domains constitutes $\sim 15\%$ of the silk, however with the extended less-ordered amorphous regions, the semi-crystalline domains can extend up to 50% of its volume. The nano-crystals assemble to form fibrils that further packed to form a silk fiber [51, 52] (Figure 1.4). The terminal domains in silk are non-repetitive and are highly conserved over different species. The terminal domains do not provide any technical function, but rather assist in the self-assembly process [53, 54]. Spacer sequences that have charged residues are also believed to be critical in the storing and assembly process [53, 55].

Figure 1.4. Formation of geometrically confined structures in silk. (a) Strength of β -strands under shear loading by varying the H-bond clusters in protein domains. (b) Comparison of H-bond clusters in various protein structures. (c) Hierarchical structure of silk and relevance of H-bond crystals in formation of silk-based materials. Adapted from Ref. [32, 51].

Computational studies suggest that smaller β -sheet nano-crystals (below 3 nm dimensions) require much higher energies to initiate failure providing effective cross-links for high tensile strength and toughness of silk, while larger β -sheets tend to be brittle. This unique structural hierarchy of silk at the molecular level and utilization of (weak) hydrogen bonds under nanoscale confinement of β -sheet crystals plays a fundamental role in achieving great strength, fracture toughness and elasticity of silk fibers (Figure 1.7). Experimental structural studies have shown that silk fibers are composed of bundled fibrils wherein the diameter of each fibril is 20-150 nm. Computational network model studies demonstrate that diameter of 50 ± 30 nm provides a homogeneous deformation state in fibrils, such that the nanoscale properties (or the strength) of the repetitive blocks can be maximized up to the higher level of hierarchy in silk fiber, even in the presence of large flaw [19, 51].

Biomimetic applications

Understanding the structural aspects of the silk proteins has been helpful in designing bio-inspired materials from silk. Owing to their abundant constituents, mechanical strength, stability at high temperature and humidity, biocompatibility with different cell types and controlled degradability; this protein based materials has been used for several biomedical and technological applications [56]. Multi-block copolymers of β -sheet enriched structures flanked with oligopeptides like the elastin-mimetic domains are beneficial to produce the protein with higher yield and it represents a useful method to control the formation of β -sheets by concentration from aqueous solutions [57]. Over the years, continued interest has been observed in investigating methodologies towards development of novel biomaterials, specifically for tissue engineering [37, 58, 59]. The understanding and the ability to control the formation of β -sheets and engineering silk-based multi block composites also enable the fabrication of materials with predictable properties and processing.

1.4.2 Collagen

General perspective

Collagen is among the most abundant protein in mammals. It accounts for 20-30% of the total protein content and is a major component in the extracellular matrix. It is a stiff load bearing protein that confers mechanical stability, toughness, and strength to various organs and tissues ranging from tendons/ligaments to skin, bone and cornea. The versatility of collagen can be seen as it fulfills the diverse mechanical requirements like elasticity, stiffness, and toughness or in storing mechanical energy [60, 61].

Composition

This fibrous structural protein comprises of 28 different types with 40 distinct polypeptide chains and a narrow molecular weight range (100 kDa for α chains and 200 kDa for β chains) [13]. Among the different types of collagen, 80-90% are fibril forming collagen (a) type I which is strong in tensile and forms the main component of the organic part of bone and is majorly found in skin and bone; (b) type II is the chief collagenous component in cartilage; (c) type III is an integral component of reticular fibers, usually comes along with type I and is found in skin and blood vessels. These types of fibrillar collagen maintain the integrity of tissues [62].

The hallmark of collagen is its structural motif that is composed of three polypeptide chains (~1000 amino acids each), which are characterized by repetitive Gly-X-O sequence wherein X is Proline and O is the post-translationally modified Proline 4-hydroxyproline. The triplet motif allows the α chains to form a unique right-handed triple helical configuration known as the tropo-collagen [58]. High Gly content is crucial for the stabilization of tropo-collagen super helix. Gly being smaller in size and its recurrence at every third position allows close association of the fibrils. It is buried within the core and is involved in hydrogen bonding and intermolecular cross-linking, thus stabilizing the triple helical structure. The larger side chains of the other two amino acids are exposed and facilitate the stabilization of the triple helix [63]. The quaternary fibrillar subunits of tropo-collagen further assemble into a right-handed super-helix to form a larger staggered array of collagen microfibrils (1 μ m in diameter).

These fibrils are further stabilized by covalent cross linking forming well organized aggregates of collagen fibers (10 μm) [61, 64] (Figure 1.5).

Figure 1.5. Structure of collagenous tissues. (a) The standard fibrillar collagen molecule is characterized by N- and C-terminal peptide sequences, which flank a series of Gly-X-Y repeats (where X and Y represent any amino acids but are frequently proline and hydroxyproline) (step 1). These form the central triple helical structure of procollagen and collagen. Three α -chains are intracellularly assembled into the triple helix following initiation of this process by the C-terminal domain (step 2). Procollagen is secreted by cells into the extracellular space (step 3) and converted into collagen by the removal of the N- and C-propeptides via metalloproteinase enzymes (step 4). (b) Collagen structure from nano to macro-scale. Adapted from Ref. [31, 65].

Structural characterization

Given the level of hierarchy used by collagen to self-assemble and function, various methods have been used to understand the close interplay between sequence-structure-properties. Various computational models (including atomistic and coarse grain models) have been used to investigate the structural organization and deformation mechanics in native collagen [61, 66, 67]. Simulations of peptides suggest the effect of specific amino acids motifs on the evolution of H-bonding, the elastic behavior and also the stability of tropo-collagen. For instance, while stretching, the number of h-bonds (which can be related to the stability of the molecule) in $(\text{GPA})_3$ and $(\text{GEQ})_2(\text{GEK})$ motif is five times greater than in $(\text{GPO})_3$ and $(\text{GAO})_3$ motifs. The

model suggests that variations in amino acid motifs leads to heterogeneous mechanical properties of tropo-collagen which thereby causes in-homogenous distribution of deformation in collagen [68]. Other models explain the interplay between the length of tropo-collagen molecules and strength of intermolecular interactions wherein the mechanical strength reaches a plateau beyond 200 nm molecular lengths. This is consistent with the molecular lengths in collagen tissues (200 - 400 nm) which are highly conserved and show universal occurrence [66].

Biomimetic Applications

Collagen has been widely investigated for biomedical applications leading to the development of bio-films, composites, or tissue engineering scaffolds that can be applied in reconstructive medicine or pharmaceuticals. Collagen serves as a biocompatible matrix for cell attachment, differentiation, or proliferation. Following different approaches, collagen has been blended with other natural or synthetic polymers as well as inorganic materials for potential applications in drug delivery devices, tissue regeneration, wound dressing and even as biodegradable packaging materials. However a major challenge is to recapitulate the hierarchical self-assembly of collagen fibrils, including the triple helix and the staggered arrangement, a feat that has not been achieved yet.

1.4.3 Elastin

General Perspective

Elastin is one the most versatile class of structural proteins that provide an outstanding combination of strength and flexibility. Elastin proteins are largely found in the connective tissues providing extensive elasticity and resilience for the tissues in the extracellular matrix that undergo repetitive reversible deformations for many required physiological functions [69, 70]. For instance, lungs, elastic ligaments, blood vessels and arteries are subjected to constant distention and physical stress more than a billion times during their lifetime without damage or permanent deformation. As a result, elastin has been characterized as a rubber-like elastomeric protein owing to its diverse biological roles and exceptional mechanical properties [70-72] (Figure 1.6).

Composition

The remarkable properties of elastin are encoded in its protein sequence. The elastin protein is assembled from its monomeric precursor tropoelastin, which is composed of alternating hydrophobic and hydrophilic domains [72, 73]. The hydrophilic domains are rich in alanine and lysine rich, are directly involved in cross-linking by enzymatic oxidation of lysine with the assistance of lysyl oxidase. The alternating covalent cross-linking imparts stability and strength to the matrix. The hydrophobic domain have low complexity, pseudo-periodic sequences with repeated tetra-, penta- and hexa-peptide motifs of glycine, valine and proline (VPGG, VPGVG and APGVGV) [74]. This domain confers elastic properties like extensibility allowing high mobility of the protein backbone and propensity for self-aggregation [69, 70, 73] via a coacervation process (Figure 1.6).

Structural Characterization

Flory and Gosline postulated a random rubber-like structure for elastin [75-79]. Consistent with these findings, macroscopic measurements by solid state NMR showed a high degree of dynamic disorder with the absence of β -sheets and α -helices [80]. Moreover the ^{13}C NMR spectra on the hydrated elastin demonstrates significant mobility which decreases with reduced hydration [81]. With these thermo-elasticity and NMR studies, it is evident that the protein chains are mobile and possess high configurational entropy. While this configurational entropy is minimized on stretching, providing restorative force upon unloading, elastin can therefore resist more than 200% strain before the protein strands are straightened completely [76, 82]. Apart from the supporting data on the random network model, CD and FTIR studies also provide limited structural insight indicating measurable propensity for β -turns and PPII helix along with high degree of conformational flexibility [83, 84].

Figure 1.6. Schematic representation of the formation of elastin fibers. (a) The secreted tropoelastin is tethered to the cell surface creating sites for further aggregation. Discreet packages of tropoelastin are released for further microfibril directed assembly, and undergo lysine oxidation that facilitates cross-linking. Tropoelastin is incorporated into growing elastic fibers. **(b)** Elastin-like-polypeptides derived from tropoelastin with a consensus repeat unit VPGXG promotes high conformational flexibility. Adapted from Ref. [10, 73, 85]

Biomimetic Applications

Elastin also displays low critical solution temperature (LCST) behavior and unlike silk or the collagens, elastin derived peptides can be easily solubilized at low temperatures for various applications. At low temperature, elastin adopts a random coil structure; however upon raising the temperature the protein has an ordered extended state to a collapsed aggregated conformation. [37, 86, 87].

Using all atom MD simulations for elastin like peptides e.g., GVPGV₇ gives a heterogeneous conformational ensemble in which the peptide retains a highly hydrated, disordered structure in both monomeric as well as aggregated form [88, 89]. They lack β -sheet structure and during the conformational exchange predominant presence of PPII and hydrogen bonded turns is observed. Consistent with the spectroscopic studies, elastin-like peptides in solution are disordered, but not random. In the aggregated state, elastin can retain both static and dynamic disorder, similar to the IDPs [90, 91]. Simulations on (VPGVG)₁₈ polypeptide constrained with β -spiral dihedral angles estimated instability of the β -spiral parameters in solution [92, 93]. Therefore based on the MD, solid state NMR and CD spectroscopy a consensus structural model of elastin suggests presence of water-swollen hydrophobic domains

which aggregate into an ensemble of conformations devoid of any regular secondary structure. The polypeptides exhibit local structural form of PPII and turns. These conformations are transient and interconvert rapidly to disordered conformation [91, 94].

The intrinsic ability of elastin to self-organize into a remarkably durable material, its elasticity, biological stability, and activity makes it an ideal model for designing. Elastin-like polypeptides are being used for developing generation of elastin-mimetic protein materials such as injectable hydrogels, micelles, sheets, isotropic networks and electro-spun finer scaffolds [95, 96] which find various biomedical applications like tissue engineering [97] and drug delivery [98].

1.4.4 Mussel Byssus

General perspective

Another artful culmination of structural materials found in nature is the mussel's byssus. These sessile organisms anchor themselves to underwater surfaces with an extracellular collagenous structure that can resist extreme hydrodynamic forces. The secretions from the foot of the mussels are called the byssus thread, which is composed of three mechanically distinct parts: (a) the root which embedded at the base of the mussel foot and attaches the entire system to the 12 reactor muscles of the byssus, (b) the stem which protrudes from the root and gives support to each thread, and (c) the byssal threads which consist of bundles radially distributed in many directions to attach themselves to various foreign substrates [99, 100] (Figure 1.7). The groove of the mussel foot makes the byssal thread by a process similar to polymer injection molding and approximately requires 3-10 mins per thread. Threads themselves can be structurally divided into three distinct parts: (a) the adhesive plaque which is attached to the substrate with a coat of collagen like fibers; (b) the proximal portion which is a corrugated sheath of loosely arranged coiled fibrils; and (c) the distal portion which is much smoother and is twice longer in length than the proximal portion and is composed of bundles of densely packed fibers [101].

Composition

Similar to the silk fibroin, byssal threads are also proteins based, devoid of any living cells. This so-called dovetail fibrous protein is predominantly a graded distribution of copolymers that provide strength and structure to the threads. The Pre-polymerized collagens (PreCOL) family of proteins self-assemble to form the fibrous core and constitutes a 46kDa collagen domain in the center characterized by G-X₂-X₃ repeats. The central collagen block is flanked with domains that define the threads stiffness. In the proximal portion the collagen block is flanked with an elastin domain (Col-P) with the pentablock X-Gly-X-Pro-Gly (X: Gly or a nonpolar amino acid) repeats. The distal portion has collagen with silk-like β -sheet domains (Col-D). While the proximal region is more flexible than the distal domain due to elastin like domains, the latter becomes much rigid due to the distinct β -sheet domains making it the strongest of the three regions. His-rich domains present in the N- and C-terminal allow cooperative metal binding in order to assemble various peptide structures of preCOLs into tenacious fibers [26, 59, 99, 102].

Structural characterization

Extensive studies have shown that mussel byssus contains extra-ordinary protein glues that allow them to stay attached even in saline aqueous media. The 5 unique proteins present within the plaque are known as the mussel foot proteins (mfp). Analysis of these five adhesive proteins (mfp-2, -3, -4, -5 and -6) before they assemble into networks show that their primary sequences are rich in tyrosine and lysine residues [102] (Figure 1.7). The mfps have higher isoelectric points and undergo post-translational modification of tyrosine to Di-Hydroxyl-Phenyl-Alanine (DOPA). The localization of proteins such as mfp-3 and -5 that are enriched in DOPA, act as primers for adhesion. The DOPA units enable adhesion by non-specifically interacting with a wide range of substrates. Depending on the type of surfaces, DOPA can promote adhesion by hydrophobic interactions or from coordination bonds on oxide surfaces by inducing cross-linking via radical processes. Mfp-6 which acts like an antioxidant preserves DOPA in the interfacial surface for adsorption. However, reports also suggest that once its reducing capacity has been depleted it can crosslink forming Cysteiny-DOPA adducts [102].

Studies have revealed the key role of redox interactions between the proteins and DOPA, contribution of hydro-phobic/phillic interactions towards formation of adhesive interactions, strong intrinsic protein-protein binding interactions or the local concentration of these proteins for adhesive activity and plaque formation [103-105]. Structural hierarchical organization of the stiff yet flexible outer core, with a gradient porous core and the fiber assembly of the byssus collagen threads defines their architecture.

Figure 1.7. Schematic view of protein organization in mussel byssus. (a) The enlargement depicts the pentablock copolymer of *M. eludis*, **(b)** Approximate distribution of known plaque proteins. Dopa-containing mfp-3 (fast and slow), mfp-5 and mfp-6 prime the interface for adhesion in *M. eludis*. **(c)** Mussel wet adhesion is dictated by time-regulated secretion of adhesive proteins in *P. viridis*. **(d)** Proposed model for surface adhesion by Pvfp-5β-Dopa in *P. viridis*. Adapted from Ref. [37, 106, 107].

Extending beyond the understanding of the DOPA paradigm, the design principle of mussel adhesion in *Perna viridis* has been explained based on the precise time-regulated secretion and the structural conformation of the adhesive proteins. Combined experimental and computational procedures have been applied to study the structure-property mechanism underlying the adhesive strength of these mussel proteins. By

triggering adhesive secretion with a saline injection into the mussel foot, 3 proteins are secreted in succession (pvfp-5 β followed by pvfp-3 and finally pvfp-6 and not as mixture of proteins). Pvfp-5 β initiates interaction with the substrate as it displaces interfacial water and acts as a primer for the assembly of other adhesive proteins (Figure 1.7). Structural models of these proteins shows that Pvfp-5 β adopts an elongated conformation, mainly composed of β -sheets connected with disordered segments. This extended surface is characterized by an asymmetric distribution of modified tyrosines (DOPA) and positively charged Lysine residues. The lysines likely help to dehydrate the wet surface thus enabling the DOPA to form strong adhesive bonds to the hydrophilic surfaces. A model is proposed in which the inherent flexibility of the protein combined with the spatial organization of tyrosine modified Dopa and lysine residues, easily and rapidly enables strong interactions with large rough surfaces such as those that characterize rocks in the sea, resulting in the assembly of the plaque [107, 108] (Figure 1.7).

Biomimetic Applications

Taking an inspiration from the versatility of mussel adhesive proteins, many mussel adhesive protein mimetic polymers have been explored and synthesized for applications such as tissue repair, (anti-) adhesive coatings or for nanoparticle stabilization [102]. Currently, DOPA chemistry is focused on the development of mussel inspired synthetic polymers, which can substitute as wet resistant sealants, adhesives, or coatings. However investigations of wet adhesion of mussels do not end with the discovery and understanding of DOPA. It is crucial to understand the intrinsic molecular basis of interactions as well as the structure of the adhesive proteins for expanding the compositional space of biomimetic systems inspired by mussels.

1.5 From Natural materials to tailored materials

In the preceding sections the structure-property-process relationships for selected natural materials have been discussed and the key sequence-structure-property-process relationships are summarized in figure 1.8 below. These materials provide key functionalities and balance multiple yet incompatible properties like strength and robustness [109].

Figure 1.8. Sequence-Structure-Property-Process relationship of silk, collagen and elastin. First panel shows the sequence and structure of the natural proteins. Second panel depicts the characteristic mechanical properties. Third panel illustrates the various forms these proteins can be processed into.

The discussed studies show that in order to accelerate the design process for a particular functional material, different configurations and various chemical compositions can be explored in a systematic manner by integrating experimental and computational modeling. The improved understanding of the structure and their assembly obtained using these overlapping concepts can be useful to develop sustainable functional materials with targeted mechanical properties.

In this context we explore yet another intriguing natural biomaterial, the squid's sucker ring teeth (SRT), by integrating experimental and computational methods to investigate the structure-property relationships towards the establishment of principles which can help catalyze fabrication of novel materials with targeted properties. The following sections describe in detail the characteristic features that SRT shares with other existing biomaterials and its own unique features that together can be explored and exploited to design novel biomimetics.

2 Introduction to Squid’s Natural Bio-tool: The Sucker Ring Teeth

---

“You could look at nature as being like a catalog of products, and all of those have benefited from a 3.8 billion year research and development period. And given that level of investment, it makes sense to use it.”

-Michael Pawlyn

Among the myriad of natural, eco-friendly green, protein based materials like the silk, elastin, collagen, keratin and resilin, another intriguing natural material is the sucker ring teeth (SRT) of the pelagic Humboldt squid that inspires this research. Since form and function are intimately intertwined, the focus of this study is to best understand the structural architecture of the building blocks of this natural protein-based bio-tool. To reach this goal, various structural studies have been applied to understand nature’s intricate structural concepts that can be used to effectively and efficiently facilitate in designing sustainable bio-inspired materials with predictable or targeted outcomes.

2.1 *Dosidicus gigas*

The elusive species of Humboldt squid or Jumbo squid (*Dosidicus gigas*) are fierce sea creatures found in the Eastern Pacific Ocean [110] (Figure 2.1). This large aggressive member of the flying squid family Ommastrephidae, hunt in big shoals to capture their prey. Unlike other cephalopods like the octopus (Order: Octopoda), which typically have eight “sucker cups”, squids (Order: Teuthidia) have eight long arms with hook like sucker cups extensions also known as sucker ring teeth and two specialized tentacles lined with razor sharp suckers with which they grasp and drag the prey towards their razor sharp beak, eventually ripping apart the prey. The two hard un-mineralized tissues of this squid (the sucker ring teeth and the beak) used during predation strategy have triggered recent interest in the biomimetic community and one of them is also the focus of our study.

Figure 2.1. *Dosidicus gigas* and its geographical distribution. (a) Elusive and cannibalistic Jumbo Squid with flaring arms lined with suckers. (b) Distribution map of *D. gigas*; red: common historical distribution; purple: recent expansion to south has been more frequent than in past to north. Adapted from Ref. [111-113].

2.2 *Un-mineralized Extracellular Hard Tissues in Squid*

The Jumbo squid has more than one hard tissue for bio-inspiration. The squid's beak is one of the hardest, stiffest and toughest, wear-resistant, organic material in existence used to lacerate tissues and subdue preys [114]. This combination of properties is correlated with a chemical gradient composed of mixtures of hydrated chitin, Gly-rich and His-rich proteins that are extensively stabilized by histidyl-dopa cross links [115]. The ramification of an opposing compositional gradation of building blocks, the inter-protein and protein-chitin crosslinks is the mechanically and functionally graded beak [113, 116]. Understanding and reproducing this graded property will open new avenues for designing functional gradient biomaterials.

In addition, this opportunistic predator has another intriguing extracellular hard tissue *i.e.*, the sucker ring teeth (SRT). The serrated suckers line up in a double series on the tentacles and the arms to grip on to their prey under mechanically high pressure. The suckers provide additional gripping power as they attach to the prey by suction and restrict sliding during prey capture, while the series of formidable triangular dentitions attached to the basal ring of the sucker penetrate into the flesh of the prey for a firm hold (Figure 2.2) [63, 117]. The following sections describe the features and characteristics of SRT that inspired this study.

Figure 2.2. Representations of *D. gigas* sucker ring teeth. (a) Squid's arm lined with suckers; **(b)** SEM image of multiple sucker ring teeth; **(c)** Single SRT image.

2.3 Unique Features of Jumbo Squid's SRT

Just less than a decade ago, suckers from different species of squids were described as chitin-based materials. However, recent investigations have shown that SRT is entirely made of a mixture of proteins, with no minerals, transition metals or polysaccharides. Yet they feature mechanical properties matching those of strong synthetic polymers. They are not as robust as the squid's beak, however they exhibit some unique mechanical and physio-chemical properties[117] as discussed below.

2.3.1 Micro-anatomy of suckers

Micro-structural SEM images show that SRT is composed of series of parallel tubular elements (Fig 2b and 2c). These channels are oriented parallel to the dentitions, spreading throughout the base. Gradual reduction in the channel diameter is observed from the center (~100 nm) to the periphery (~260 nm) of the tooth cross-section. Additionally increase in channel spacing also contributes to the formation of a highly controlled nano-porous architecture (Figure 2.3). The porous architecture is common in the design of a diverse range of natural materials like trabecular bones, woods, and plants. This porosity gradient renders graded hardness from core to the periphery which is a micro-mechanical feature that has also been observed in many biological materials that exhibit high wear tolerance [117].

Figure 2.3. Microstructural features of *D. gigas* SRT: (a) Illustration of SRT cross-section showing channel organization from tooth tip into the base. SEM images (b) and (c) with highly parallel nature of nano-tubular elements. (d) Pore diameter as a function of distance from tooth's periphery to the core. Adapted from Ref. [117]

2.3.2 Mechanical Properties of SRT

The microstructure of SRT with nano-sized parallel channels influence its mechanical properties [117]. Nano-mechanical characterization of SRT under dehydrated conditions result in a distinctive gradient with decrease in both elastic modulus (E) and hardness (H) from periphery to the core. Nano-indentation maps of E and H generated in dry state results in values as high as 8-10 GPa and 0.7 GPa respectively near the tooth periphery, which is similar to some strong synthetic polymers such as polyamides, PEEK or PMMA [118]. As the channel diameter decreases, gradual reduction in E and H is noted with a minimum of 4.5 and 0.4 GPa respectively in the SRT core (Figure 2.4). Under hydrated conditions similar trends are observed, with a 2.5-3 fold decrease in values of E and H compared to the corresponding region in the dry state [117].

Figure 2.4. Micro-mechanical properties measured on *D. gigas* SRT cross-section (a) Illustration of individual tooth microtome cutting plane. (b) Nanoindentation maps of modulus and hardness in dry and wet state. Adapted from Ref. [117]

2.3.3 Biochemical Composition of SRT

Elemental analysis of SRT indicates uniform distribution of Cl and S apart from C, N and O, with absence of metals or mineralization throughout the SRT (Figure 2.5), which is a often observed in Nature to form hard tissues [117]. Chitin, another common natural structural polysaccharide was also absent in SRT [114]. This is intriguing since squid's beak is comprised of chitin by 20 wt.%, which is considered as a common biological structural polysaccharide and is a key block responsible for the mechanical properties of the beak [115]. Initially suckers were often described as a chitinous material; however recent studies have proved it to be a protein based hard tissue. The SRT proteins or the “suckerins” show compositional bias towards glycine (Gly, 37 mol. %), tyrosine (Tyr, 14 mol. %) and histidine (His, 13 mol. %) residues, in addition to abundance of other hydrophobic residues [117, 119, 120].

In general, Natural and synthetic polymers are mechanically stabilized or reinforced majorly by three different physiochemical strategies: (a) chain entanglement as seen in classic amorphous thermoplastic polymers [120]; (b) addition of a much stiffer reinforcement phase like the fine mineral phase in bones or other bio-minerals [121]; in the engineered nano-composites [122]; the less common approach of using transition metal coordination bonding (as seen in arthropod predatory appendages or the marine worm jaws) (c) dense covalent inter-chain cross-linking. The later

framework is commonly found in synthetic polymers like thermoset resins or natural biopolymers like insect exoskeleton [123] or the squid's beak itself [115]. However, strikingly there is complete absence of covalent chain cross-links, mineralization or metals in SRT (Figure 1.15) [119, 120]. This characteristic combination of mechanical properties and its composition makes SRT a distinct model system to explore structure-property-function relationships and in enhancing development of robust structural polymers.

Figure 2.5. Chemical and elemental decomposition of *D. gigas* SRT: (a) EDX spectrum revealing presence of both chlorine and sulfur with no metals detected. (b) Amino acid analysis spectrum showing absence of chitin peak in comparison to glucosamine standard (pink curve). (c) SDS-PAGE analysis sucker ring teeth proteins. Adapted from Ref. [117, 120].

2.3.4 The Suckerin Protein Family

Established proteomics tools combined with transcriptome databases generated by RNA-seq were used to identify the primary sequences of proteins that from the suckerins [108]. The transcriptome analysis of SRT from three distantly related cephalopods *i.e.* the Jumbo squid *Dosidicus gigas* (Order Oegeopsida), the golden cuttle fish *Sepio esculanta* (Order Sepiida) and the bigfin reef squid *Sepioteuthis lessoniana* (Order Myopsida) have shown that SRT is assembled by a repertoire of proteins (M.W. ~ 5 to 60 kDa and pI between 7-10) (Figure 2.6). Although *D. gigas* exhibits larger array of proteins, all the SRT proteins (known as the suckerins) have similar amino acid composition and display extreme modular architecture (Figure 2.6) with ~20-90% sequence similarity [120]. Moreover, exhaustive NCBI or Uniprot search do not yield any statistically significant matches in the protein databases, suggesting that suckerins can be identified as a distinct class of structural proteins [120].

Figure 2.6. Primary structure organization in suckerin protein family. Modular primary sequence architecture of suckerin proteins. **(a)** *D. gigas* **(b)** *S. lessoniana*, and **(c)** *S. esculenta*. Adapted from Ref. [45].

The amino acid sequences of suckerins are composed of two different modular repeats. The M1 modules that range from 10-15 residues, rich in Ala, His, Val, Ser and Thr residues with the archetypal presence of Pro residues flanking these modules and the M2 modules that range from 20 to 40 residues, featuring Gly-rich domains with high amount of Tyr and Leu which occurs in small repetitive peptides motifs of “GGY”, “GGL” and “GGLY” [108]. These two types of modules are ubiquitous in suckerins and each module is speculated to confer distinct characteristics to the protein. The M1 modules which are enriched in hydrophobic Ala and Val residues resemble the poly-Ala domains found in spider silk [45], and are considered to be reminiscent of GGVVIA peptide sequence, which is reinforced to form cross- β structure in β -amyloids [48]. On the other hand, the M2 module rich in Gly, Tyr, and Leu residues are considered to be reminiscent of the amorphous-forming sequence similar to the one found in dragline silk proteins.

Both Ala and Gly-rich motifs resemble peptide sequences from the oyster shell matrix proteins, chicken claw keratins, and migratory locusts. The His-modules prevalent within M1 modules exhibit similarity to His-rich proteins found in the jaw of the marine polychaetes like *Nereis virens* and *Glycera dibranchiata*. However, these hard tissues in the aforementioned species employ metal-coordinate chemistry for the reinforcement of the materials via His-Zn (*Nereis virens*) or His-Cu (*Glycera dibranchiata*) coordination bonds and dense covalent cross-linking. The GGY tripeptide is generally found in many structural proteins like the silk, insect cuticle, shell matrix proteins or the β -keratin protein crocodile skin, suggesting a convergent evolutionary origin of this motif in suckerins [45, 108, 124, 125].

On a large-scale the modular design of Pro-[M1]-Pro-[M2] units is observed to be conserved across the three species of cephalopods (Figure 2.6). However, the length and the frequency of Gly-rich module [M2] varies among different suckerins, leading to different mechanical behavior and self-assembly of SRT in the three species. Phylogenetic analysis suggests that the ancient suckerin gene family arose during the Devonian period and diverged into 6 different clades that are conserved across these three species [120]. Different clades reveal that suckerins vary in size, the number of tandem repeats and in the presence or absence of modules in the sequence [120]. Based on the dataset from the three species, it is noted that genetic divergence for suckerin proteins includes gene duplication, segmental expansion or deletions. These insights illustrate the molecular evolutionary process behind SRT's functional mechanics. For instance, the effect of difference in protein composition (based on the molecular weight) in correlation to the mechanical properties (E) of SRT can be observed among these cephalopods. *D. gigas* have stiffer and stronger SRT (*S. lessoniana* 6 GPa and *D. gigas* 10 GPa) as it has evolved a larger repertoire of proteins and has abundant presence of higher molecular weight proteins, which can allow better inter-chain interactions owing to the longer polymer chains [120].

2.3.5 Molecular Architecture of SRT

Comparative structural analysis on the ultra-microtome cross-section of SRT reveals presence of β -sheets, similar to what has been found in *Bombyx mori* (*B. mori*) silk (Figure 2.7) [126]. Compared to the spider silk and the *B. mori* silk, the Raman spectra

in the cross-polarized mode suggest that the β -sheet domains in the native SRT are oriented randomly [108]. This random orientation of crystalline β -sheet domains in SRT has also been confirmed by Wide Angle X-ray Scattering (WAXS) with equal intensities observed at all azimuths and reflection positions. Integration across all azimuthal angles revealed peaks attributed to reflections from (120), (200), (002) and (100) planes, which are consistent with spider silk [127] and silkworm [128] (Figure 2.7). A detailed examination at the nano-scale level reveal that estimated β -sheet dimensions: 2.4-2.6 nm in hydrogen bond direction and 3-3.5 nm in peptide bond direction. The deconvolution of this pattern can be corresponded to the 5 β -strands (in width) with $\sim$ 8-10 residues in each strand, respectively (Figure 2.7) [120]. These results were also consistent with the presence of a highly conserved Pro residue (which acts a β -strand disruptor). Pro residues consistently flank the M1 modules thereby constraining the β -strand composed of $\sim$ 10-11 residues within a module in SRT (Figure 2.7). This is also in agreement to the reported β -strand length of $\sim$ 3 nm in native SRT using WAXS [120].

Computational modeling studies have shown that nano-confined β -sheets with similar dimensions confer enhanced stiffness, strength and toughness by cooperative hydrogen bonds [48]. Use of weak hydrogen bonds in formation of strong materials requires considerably less energy, facilitating their self-assembly at moderate temperatures. It also imparts the built-in ability for self-healing of material since hydrogen bonds that are ruptured can be reformed [31, 48, 109]. The M2 Gly-rich modules that follow the repeating unit of Pro-flanking M1 modules are thought to be mostly amorphous (with flexible disordered or β -turn conformation). Based on this, the overall structure of SRT is proposed to be similar to a semi-crystalline synthetic polymer; wherein the nano-confined β -sheet domains [M1] (flanked with Pro residues to restrict the β -sheet dimensions) are embedded within the longer [M2] amorphous domains that provide flexibility and elasticity. (Figure 2.7) [120]. Thus, a semi-crystalline molecular design arising from the repetitive peptide building blocks provide a suitable combination of rigidity, strength, and elasticity conferred by the SRT supra-molecular network.

The nanoscale features of the entire tooth investigated by Small Angle X-ray Scattering (SAXS) mapping revealed strong anisotropy which could be strongly related to the existence of elongated nano-fibrils positioned approximately parallel to the tooth contour (Figure 2.8). Further analysis based on the radial integration of pattern led to reflections suggesting hexagonal packing of the elongated nano-fibrils (Figure 2.8) [129]. Based on this X-ray data, the proposed model of native SRT suggests, the isotropic distribution of β -sheets, which stabilize silk like protein polymer networks, strengthen the amorphous protein network and are responsible for load bearing predatory function of SRT which must resist complex compressive shear and torsional load regimes [108]. These nano-fibrillar units assemble into a hexagonal packed lattice that predominantly runs parallel to the tooth contour and provide high bending rigidity during prey capture [129].

Figure 2.7. Characterization of protein structure in *D. gigas* SRT by Raman spectroscopy and WAXS. (a) Raman spectrum of a microtomed SRT section compared against a *B. mori* silk standard, showing matching of the amide I band centered at 1,666 cm⁻¹, which is attributed to β -sheet conformation. (b) Raman spectra in cross-polarized modes suggest that the β -sheet domains in SRT are randomly oriented. (c) WAXS pattern of SRT (tip region) and (d) integrated spectrum (all azimuthal angles), indicating reflection positions consistent with β -sheets (inset: estimation of β -sheet size using the Scherrer analysis). Deconvolution indicates a broad amorphous halo. (e) Semicrystalline structure of SRT, consisting of randomly oriented β -sheets within an amorphous matrix. Source Ref. [108, 120]

Figure 2.8. Characterization of native SRT by WAXS and SAXS. (a) Multiple 2D WAXS patterns with steps of $100\ \mu\text{m}$ covering the entire tip of a single sucker ring tooth. A representative 2D WAXS pattern (b) and the radial integration pattern of the intensity (c) as a function of the scattering vector q . The observed ring patterns confirm the random orientation of the β -sheet crystals. (d) Diffraction pattern mesh highlighting the SAXS region from a, clearly showing structural anisotropy at the nanoscale. (e,f) A representative 2D SAXS pattern (e) and the radial integration pattern of the intensity (f). The pattern reveals the fibrillar organization of the protein network and the positions of equatorial reflections suggest a hexagonal packing of the fibrils with an inter-fibrillar distance of $12.5\ \text{nm}$. (g) Color-coded image depicting the mean fibrillar orientations (black lines) superimposed on an X-ray transmission data set. (h) Schematic model based on the X-ray scattering observations. Source Ref. [129]

2.3.6 Contribution of β -sheets towards SRT's mechanical properties

The key role of β -sheets in contributing towards the mechanics of SRT material has been shown by disrupting the hydrogen bond interactions dominantly present in SRT. Spectroscopic analysis combined with nanoindentation of urea (hydrogen bond disruptor) incubated SRT showed a decrease in β -sheet content and dramatic reduction in elastic modulus (2-3 GPa in water vs. 100-200 MPa in 2M Urea). However ethanol treated sample of SRT did not show any major change, suggesting minor contribution of hydrophobic interactions in mechanics of SRT (Figure 2.9). Therefore hydrogen bonds localized in β -sheets play a critical role in stabilizing and reinforcing SRT structure. With this basic understanding of the structure-property relationship, biomimetic materials can be fabricated with tractable mechanical stiffness by the incorporation of these β -sheet structures in a controlled manner.

Figure 2.9. Correlation between the mechanical properties and secondary structure of *D. gigas* SRT. (a) E (elastic modulus) as a function of time measured by nanoindentation. (b) Micro-Raman spectroscopy of SRT incubated in 0.2 M Urea. Dotted arrow lines correspond to β -sheet structure and its disruption with exposure to urea. (c) Schematic representation of H-bond disruption accompanied with β -sheet dissolution when in contact with urea. Source Ref. [130]

2.3.7 Thermoplastic properties of SRT

Most thermoplastic materials have been derived from petro-chemicals. Materials with this property provide multitude of applications due to their ease of processing and forming. From sustainability perspective use of renewable biopolymers would be highly favorable. However, often protein-based materials do not exhibit thermoplastic behavior due to irreversible degradation of their macroscopic physical properties. For example, collagen exhibits limited solubility due to abundant inter-molecular crosslinks, while the collagen extracted from animal sources often requires acidic, alkaline or other enzymatic treatments that result in irreversible degradation. Moreover hybrid bio-composites using spider fangs or insect cuticle display high mechanical properties, but need harsh denaturing conditions for solubilizing and further processing. The robust SRT is an intriguing exception among thermoplastic natural biopolymers. Because of the weak hydrogen bonds that stabilize the structure of SRT, it can be melted in aqueous solution and can be molded multiple times like polyethylene, providing insights into eco-friendly processing of biomaterials [108, 129].

SRT is primarily a biphasic structure stabilized through hydrogen bonds localized in β -sheet nano-crystals which are further flanked by the amorphous domains [120]. Each phase of this semi-crystalline biopolymer not only exhibits a specific thermal stability, but also shows characteristic affinity towards water and other plasticizing agents. SRT contains a modest ~13 wt.% of water, given that the amorphous domains have lower percentage of hydrophilic residues. This water is bound to the protein chains as opposed to the free water molecules. In addition as shown by thermo-mechanical experiments it plays a critical role as a plasticizing agent. At high temperature and under hydrated conditions, decrease in the modulus of SRT (3 GPa to 200 MPa) is regarded to be due to increased flexibility of amorphous domains in water. However in dry conditions, the modulus increases with temperature (6 GPa to ~12 GPa apparently a very high value for non-crosslinked thermoplastic polymers). This is attributed to the loss in residual water from the structure that ensues loss in chain mobility and thus lattice compaction, leading to a stiffer polymer network. The stiffening of the polymer chain in corroboration with the temperature-dependent spectroscopic measurements reveals melting of β -sheets, leading to disruption of its semi-crystalline structure and

loss in an order of magnitude in the modulus (Figure 2.10) [129]. Thus SRT serves as an interesting model for molecular design of biomimetic proteins with modular architecture that can be fabricated or re-processed into thermoplastic biopolymers. Such bio-inspired eco-friendly materials (substrates) can find potential applications in biomedical devices and even 3D printing.

Figure 2.10. Influence of temperature on SRT structure and mechanical properties (a) Thermal gravimetric analysis (TGA) from room temperature (RT) up to 900 °C. Following water evaporation, SRT remain stable until 250 °C. **(b)** DSC curves for hydrated (blue) and dry (red) samples. When hydrated, the heat flow was measured from RT to 60 °C. Under dry conditions, at first cycle (not shown) the temperature was ramped up to 120 °C, to remove residual water. Subsequent cycles were ramped up to 120 °C (cycle 2), 170 °C (cycle 3) and 250 °C (cycle 4). Exothermic peaks at 137 °C and 51 °C were reproducible. The exothermic peak at 220 °C is attributed to β-sheet melting. **(c)** Reduced elastic modulus (E_r) measured by high-temperature indentation using a temperature-ramping stage under both hydrated (blue) and dry (red) conditions. Adapted from Ref. [129]

2.4 Bio-materials from SRT and Suckerin

Given the unique thermo-mechanical properties of this block co-polymer-like, SRT proteins have been explored for potentially functional biomimetic applications. Native SRT powder can be processed into a liquid film that can be nano-patterned or can be readily spun into fibers. The secondary structure of these solvent free suckerin materials (films and fibers) remains unaffected even after melting and patterning. Processed SRT maintains silk like supra-molecular organization with conserved β -sheet signature [108]. Another biochemical feature exploited for biomimetic applications is the large Tyr content (7-16 mole %) in the suckerin primary sequence (exclusively found in the amorphous domains). Cross-linking these Tyr to form di-Tyr crosslinks in one of the highly expressed suckerin recombinant protein (suckerin-19) has been achieved to fabricate soft hydrogels with storage modulus of 40 to 500 Pa, to stiff cross-linked films with Young's moduli in GPa range. The cross-link density and the β -sheet content of these materials can be tuned to obtain broad range of moduli in order to acquire targeted mechanical response [130].

Cell culture studies with various cell lines on the cross-linked films demonstrate negligible cytotoxicity as the cells displayed high viability, exhibiting biocompatibility and cell growth. Being biocompatible, these materials with tunable mechanical properties can be further explored to target applications in tissue engineering and restorative medicine such as duplicating the mechanical gradients of connective tissues, which is currently a major challenge in restorative surgery.

2.5 *Objective and Scope of the Project*

SRT illustrates remarkable mechanical properties, an intriguing microstructure and a protein dominated structural framework. This exclusive protein composition is enriched in β -sheet reminiscent of silk. Additionally, the modular “block copolymer” primary sequences of the suckerin family are similar to the repetitive spider silk sequences. The recombinant suckerin proteins self-assemble *via* β -sheet formation, similar to the process observed in native SRT. Additionally, their conformational conversion from solution can be tuned efficiently to engineer a wide range of biomaterials like fibers, films, etc., with an integrative approach. These derived materials show potential in engineering and biomedical applications.

For any given biological material the fundamental understanding of the structure (at any level of hierarchy) can potentially enhance the interpretation of their mechanical properties. Thus, it is critical to understand this complex relationship or the influence of the sequence on structural organization and processing. Based on this knowledge and to better comprehend the structure-property relationships in SRT, we propose to elucidate the atomistic structural features of suckerin proteins at the fundamental level of hierarchy.

Since SRT is an assembled block copolymer composed of similar proteins the aim is to structurally characterize one of the dominant proteins, named suckerin-19. This protein can serve as a prototype for the suckerin protein family and has been chosen in this study for the following reason (i) it is the most abundant protein in the *D.gigas* SRT; (ii) The molecular weight of this protein (39 kDa) is in midrange between the smallest (10 kDa) and the largest suckerin (57 kDa); and (iii) it exhibits the classical Pro-[M1]-Pro-[M2] modular architecture, conserved within most other suckerin genes (iv) this recombinant protein has been processed into bio-materials like cross-linked films and gels; thus useful in bridging structure-property-process relationships.

Using established molecular biology tools we intend to obtain the refolded protein for basic spectroscopic characterizations. Since the protein is rich in Gly residues, it is likely that inherent structural flexibility (disorder) or its intrinsic property to self-organize may limit the progress of this stream of research. Therefore, different constructs with varied modular compositions have been designed (maintaining the

modular sequence design of the suckerin proteins) for structural characterization. Given the time frame and the availability of the state-of-the-art structural characterization techniques, various soluble forms of this suckerin protein were explored by integrating experimental methods and computational modeling. These methods can be applied to understand the residue or domain specific (the sequence) influence on the local conformation and subsequently the global structural fold of the protein. The ultimate goal is to better understand the role of these simple repetitive building blocks (sequence) on the tertiary structural fold and eventually on inter-protein interactions. Investigating the intra- and inter- molecular interactions that stabilize the structure is imperative to interpret the source of SRT's mechanical properties.

Taken together, this study outlines recombinant production, purification, and renaturation of suckerin-19 protein and its shorter construct. The structural fold and the stability of proteins will be assessed using various spectroscopic methods including CD spectroscopy, ATR-FTIR and TEM. The tertiary fold will be determined by integrating information from experimental and computational methods. We envisage that these structural insights will be beneficial for establishing molecular-scale design principles that can be used for *de novo* development of protein-based materials towards biomimetic research.

3 Methods Applied To Characterize The Structure of Suckerins

3.1 X-ray diffraction on native SRT fibers

SAXS and WAXS are two closely related techniques often used to identify overall structural details of proteins. While the SAXS data provide information in the nano-scale range (between 10-100 nm), WAXS can be used to investigate secondary structural elements. Therefore simultaneous collection of these two data sets allows significant increase in the resolution. In the experimental set up for SAXS and WAXS, the only difference is in the scattering angle that can be controlled by changing the distance of sample from the detector. For the WAXS, the detector is placed closer to the sample to collect higher angle scattering data, corresponding to smaller-scale (molecular) features of the structure.

The SAXS/WAXS measurements were conducted on the drawn fibers from heated SRT. The measurements were performed at microfocus beamline at BESSY II (Helmholtz Centrum, Berlin, Germany). X-ray scattering patterns were recorded with 2D CCD (charge-coupled device) detector (MarMosaic 225, Rayonix, Inc., Evanston, USA) with a pixel size of 73 μm and an array of $3,072 \times 3,072$ pixels. The energy of the incident beam was 15 keV. The beam was further defined by a toroidal mirror together with MoBC-multilayer monochromator with a pinhole size of 100 μm close to the sample. For sample mapping 60 μm step size was used in scanning mode. The calibration and data evaluation were performed with DPDAK (DESY/MPiKG, Gunthard Benecke, v.1.0.0) and OriginPro 8.6 was used for data plotting.

Fibers made of native suckerins were prepared by melt spinning. 10 mg of SRT were crushed in liquid N_2 . The dry powder was re-suspended in 10 mL deionized water and was heated in a microwave (900 W) for 25 seconds, leading to a viscous melt. Fibers were immediately drawn from the melt, yielding microfibrils of about 1 cm long and 10 μm in diameter. Fibres were allowed to dry on glass slides and about 5 fibres were bundled together and glued on the cardboard frame (1.5 cm $\times$ 1 cm). The scattering data was acquired on these bundled fibres.

3.2 Cloning, expression and purification of suckerin proteins

3.2.1 Suckerin-19

Cloning

The suckerin-19 clone was provided by Dr. Shawn Hoon (A-star, MEL). The gene for suckerin-19 was cloned by RT-PCR and the sequence was confirmed by Sanger sequencing. The gene encoding SRT-39 was cloned into pET23a expression vector. The ligated recombinant product was verified by DNA sequencing. The plasmid was then transformed into *E. coli* BL21 (DE3) cells by electroporation. The transformant was spread on Ampicillin resistant plates. Single colonies were picked and were grown overnight in 3 mL Luria Broth (LB) with ampicillin (100 µg/mL) at 37 °C. The cells with glycerol were aliquot into 1.5 mL sterile screw cap tubes and stored at -80 °C.

Expression

Prior to large scale expression, transformed cells stored in -80 °C were inoculated overnight at 37 °C in 10 mL LB, supplemented with ampicillin (100 µg/mL). The cells were grown in 1 liter of LB medium at 37 °C containing 100 µg/mL ampicillin until an OD₆₀₀ of 0.4-0.6 was reached. A final concentration of 0.5 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to induce the expression at 37 °C for another 4 hours with rigorous shaking. The cells were then harvested at 15000×g at 4 °C for 15 mins and were washed twice with 20 mM Tris pH 8. The cell pellet was stored at -80 °C until next use.

Protein Extraction and Purification

The cell pellet was re-suspended typically in 50 mL lysis buffer (50 mM Tris pH 7.4, 200 mM NaCl, 1 mM PMSF, pH 7.8). The buffered suspension was homogenized by vortexing and was cooled on ice. The cells were lysed using a laboratory scale Microfluidizer high-pressure homogenizer (Model: M110P) at 25000 psi for up to six passes. The resulting suspension was centrifuged at 43000×g for 1 hour at 4 °C to separate the soluble fraction. The soluble fraction was stored separately for analysis of constituent soluble proteins. The pelleted inclusion bodies were washed twice 25 mL Wash Buffer I (100 mM Tris pH 7.4, 5 mM EDTA, 5 mM DTT, 2 M Urea and 2% (v/v) Triton X-100, pH 7.8). The solution was vortexed and centrifuged at 5000×g,

4°C for 15 mins. To remove Triton X-100 and urea the inclusion bodies were further cleaned two more times with 25 ml Wash Buffer II (100 mM Tris pH 7.4, 5 mM EDTA, 5 mM DTT, pH 7.8).

The proteins were extracted by re-solubilizing the inclusion bodies in 5% acetic acid followed by 30 minute centrifugation at 43000×g, 4 °C. The supernatant was separated and the isolated pellet was dissolved using 8 M Urea + 5% acetic acid to extract proteins under denaturing conditions and centrifuged (43000×g, 4 °C, 30 mins). Second step purification of suckerin-19 solubilized in 5% acetic acid was purification by HPLC (1260 Infinity, Agilent, California, USA) using reverse phase chromatography. The protein was separated on C8 S-300 column (Agilent) by decreasing the mobile phase polarity. The gradual increase in the Acetonitrile (ACN) concentration is known to causes de-sorption of the proteins based on their hydrophobicity. Since suckerin-19 is rich in hydrophobic residues this method of purification is employed. The column was initially equilibrated with 5% ACN. Subsequently purification was carried out by gradually decreasing the ACN concentration from 5% to 50% in 50 minutes. The fractions from the HPLC purification were desalted by dialysis against 5% acetic acid and freeze dried.

3.2.2 Sepio Monomer

A shorter *de novo* suckerin construct was designed based on the sequence suckerin-1 from the bigfin reef squid *Sepioteuthis lessoniana* (*S. lessoniana*) [120]. This specific suckerin was selected because it harbors the Pro-M1-Pro-M2 (M1: module 1; M2: module 2) sequence design common to most suckerins, with close homology to *S.l.* suckerin-1 (12 kDa) and *D.g.* suckerin-19 (39 kDa). The generated construct shares 81% sequence identity to *S.l.* Suckerin-1 (length coverage 70%), 72% sequence identity *D.g.* suckerin-19 (length coverage 52%) and 40-70% sequence similarity with other suckerin proteins. However it has a lower molecular weight (MW) of 14.6 kDa, making it more convenient to express and manipulate in solution. The construct is designed with an aim to improve the intrinsic folding yield, stability, and solubility for structural determination. This suckerin construct is called “Sepio suckerin-1 monomer” or “sepio monomer” from here on. Additionally suckerin constructs from this monomer were designed by repeating the monomeric unit multiple times to achieve a MW similar to that of suckerin-19. Specifically a trimeric construct

constituting of three copies of sepia suckerin-1 monomer was cloned, and expressed (which is called “sepia suckerin-1 trimer” or “sepia trimer” in short from here on) Materials derived (recombinant protein hydrogels and films) from the sepia suckerin constructs have shown to exhibit similar mechanical properties to suckerin-19 [unpublished study]. Therefore this construct is expressed to further characterize its structural features and use the insights to understand the structural organization of the suckerin protein family in general.

Cloning

The codon optimized gene encoding double His₆-tagged sepia monomer was purchased from DNA 2.0 (California). The full-length open reading frame (ORF) was amplified by RT-PCR using KOD polymerase. The primers used for amplification were: forward primer 5'-GCTAGCCTGCTGCCGGGTCTGATGGGT-3' and reverse primer 5'-CTCGAGCAGAACGTAACCAGAACCGTA-3. The PCR product was purified using QIAquick Gel extraction Kit (Qiagen). The DNA fragment was subsequently subcloned into NheI/XhoI digested pET28a (Novagen) expression vector. The recombinant protein contained His₆-tag sequence (MGSSHHHHHH-SSGLVPRGSHMAS) located at the N-terminus of the complete amino acid sequence of sepia monomer, which could be used for His-tag cleavage. 6XHis-tag present at the C-terminus of the protein sequence cannot be digested and was required for the stability of the clone. The purified recombinant plasmid sepia monomer construct was transformed into *E.coli* BL21 (DE3) cells. The transformed cells were streaked onto LB agar plates with Kanamycin resistance (50 µg/mL). Sanger sequencing confirmed the nucleotide sequence of the resulting recombinant plasmid. Isolated colonies were picked and grown overnight in LB supplemented with Kanamycin (50 µg/mL) at 37 °C. The overnight bacterial growth was resuspended in 50% glycerol (1:1) and was aliquoted into 1.5 mL sterile screw cap tubes and was stored at -80 °C.

Protein Expression

Unlabeled Protein Expression:

Single colonies from transformed cells were inoculated into selective LB media with Kanamycin (50 µg/mL) and was grown overnight at 37 °C, 250 rpm. For strain activation the cell cultures were diluted to 100-fold into fresh LB media supplemented with Kanamycin (50 µg/mL). The cells were grown at 37 °C, with rigorous shaking to

OD₆₀₀ = 0.6-0.8, following which the cell culture was induced with 0.5 mM IPTG for 20 hours at 37°C. The overnight cell culture was pelleted by high-speed centrifugation (15000×g) at 4 °C for 15 mins and washed twice with 20 mM Tris, pH 8. The cell pellet was stored at -20 °C for lysis.

¹⁵N-¹³C labeled Protein Expression:

For isotope labeling a minimal medium (MM) containing 1 g/L [¹⁵N] ammonium chloride, 2 g/l [¹³C] glucose, and a vitamin cocktail (5 mg/L thiamine, 1 mg/L d-biotin, 1 mg/L calcium chloride, 1 mg/L magnesium sulfate) was used. Expression for MM was induced at OD₆₀₀ = 0.8. The cells were harvested for 20 hours after induction with 0.5 M IPTG at 37 °C. Cells were subsequently centrifuged, re-suspended in 20 mM Tris, pH 8. The harvested cells were stored in -20 °C for lysis.

Protein Extraction and purification

The harvested cells from each liter of culture were lysed in 20 mL lysis buffer (50 mM Tris, 200 mM NaCl pH 7.8) and were vortexed to obtain a homogenized mixture for sonication (Amplitude: 50%; Pulse: 01 sec + 01 sec pause; Time: 10 mins). The lysate was clarified by centrifugation (15000×g, 4 °C, 15 min). The inclusion bodies and the soluble fraction obtained were separated. The proteins from the pellets were extracted in 5% acetic acid followed by 30 minute centrifugation at 43000×g, 4 °C. The supernatant was separated and the isolated pellet was re-dissolved using 8 M Urea + 5% acetic acid to extract proteins under denaturing conditions and centrifuged (43000×g, 4 °C, 30 mins). Size exclusion chromatography (SEC) was performed using AKTA-purifier FPLC system (GE Healthcare) with UV detection at 280 nm. The sepio monomer was purified using Superdex 75 gel filtration column at a flow rate of 1 mL/min with 20 mM sodium acetate, pH 5.0 as a running buffer. The purification run involved the injection of 5 mL acid refolded protein obtained after dialysis in 20 mM sodium acetate pH 5.0. (14 × 1) mL fractions were collected with an in-line fraction collector module.

3.3 Identification of Recombinant Proteins

3.3.1 Polyacrylamide Gel Electrophoresis

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) is an analytical method used to separate components of a protein mixture based on their size. This method is popularly used to identify proteins, estimate their molecular weight, and gauge the purity. For the identification of suckerin proteins the gels were prepared using Bis-Tris buffer at pH 6.4 to enhance gel stability. The buffer compositions for preparation of 8% resolving and 4% stacking gels are mentioned in the table below. 3 μ L NuPAGE® LDS Sample Buffer (4X) was mixed with 10 μ L protein to prepare the samples for denaturing gel electrophoresis. To further ensure complete protein denaturation, the samples were heated at 90 °C for 10 mins, centrifuged and subsequently loaded in the respective wells in the gel. The gels were run in MES running buffer (50 mM MES, 50 mM Tris Base, 0.1% SDS, 1 mM EDTA, pH 7.3) for the separation of the suckerin proteins. This running buffer is mostly used for separating small- to mid-sized proteins, thus it is useful in separating suckerin proteins with molecular weight ranging between 14-40 kDa. As a reference for identification of the molecular weight of proteins, Precision Plus Protein™ Unstained Protein Standards (Bio-Rad) was used. To obtain optimum resolution of the targeted proteins, the gel electrophoresis was performed at 150 V for 50 mins.

Table 3-1. Buffers and their respective volumes used for gel casting.

	8% Resolving Gel	4% Stacking Gel
dH ₂ O	1420 μ L	955 μ L
30% Acrylamide/Bis Solution (Bio-Rad)	970 μ L	240 μ L
3X 1.07 M Bis-Tris pH 6.4	1210 μ L	605 μ L
10% APS	30 μ L	15 μ L
TEMED (Bio-Rad)	3 μ L	1.5 μ L

For visualization of the protein separation pattern, the proteins were initially fixed in the sensitization buffer containing 30% v/v ethanol, 10% v/v methanol and 10% v/v acetic acid for at least 30 minutes. Subsequently the gel was stained in Blue silver staining solution overnight. The gel was destained briefly in 70% v/v ethanol until the desired background transparency is acquired.

3.3.2 MALDI-ToF

Matrix-Assisted Laser Desorption and Ionization Time-of-Flight (MALDI-ToF) mass spectrometry is sensitive technique used for determination of mass of proteins and peptides and hence their respective identification. In theory proteins and peptides are fragile and tend to fragment when ionized. MALDI is a soft ionization technique, which uses short laser pulse to ionize the molecules [131]. MALDI is attached to the ToF analyzer which measures the time it takes for molecules to travel a fixed distance [132]. The sample is uniformly mixed with a matrix on the plate. In vacuum environment the mixture tends to crystallize and is subsequently irradiated with a short pulse. As the sample and matrix enters the gas phase, they are released from the plate. The ions then accelerate in ToF analyzer because they are subject to an equal electric field. The ions travel in a straight and linear direction to the detector and the mass to charge ration (m/z) of the sample ions is recorded and displayed in the spectrum [133]

The precise molecular weight for all suckerin proteins and constructs was verified by MALDI-ToF mass spectrometry using an AXIMA-ToF2 (Shimadzu) equipped with an N_2 laser (337 nm, 4 ns pulse width). Sinapic acidic (Sigma) was used as a matrix for photon excitation and to ionize the target protein since. The matrix was prepared in 50% Acetonitrile with 0.1% TFA in milli-Q purified water. For sample and matrix deposition sandwich method was employed where in 1 μ L of matrix was initially deposited on the MALDI-ToF steel plate and was allowed to dry. Following which 2 μ L of protein extracted in 5% acetic acid was mixed with equal volume of sinapic acid (10 mg/mL) was prepared. The suspension was pipetted onto the same spot and then again the matrix was added on top of it as the final layer. The samples were air dried for 30 mins. A 20 kV accelerating voltage was used and spectra (over mass/charge (m/z) ratio) were acquired in linear mode averaging 100 laser shots for each spot at a power setting of 120 system units.

3.4 Determination of Protein Concentration

Bradford Assay

The Bradford protein assay is based on the observation that the absorbance maximum for an acidic solution of Coomassie Brilliant Blue G-250 at 595 nm on protein binding. Both hydrophobic and ionic interactions stabilize the anionic form of the dye, causing a visible color change. Within the linear range of the assay, the more protein present, the more Coomassie binds. The protein concentration of a test sample is determined by comparison to that of a series of protein standards known to reproducibly exhibit a linear absorbance profile in this assay [134].

Concentration of suckerin proteins was measured by the Bradford method since they either lack or have low molar concentration of aromatic residues. A suitable amount of protein standards were prepared as mentioned in the protocol (Thermo Scientific™ Coomassie Plus™ (Bradford) Assay Kit) using 1 mg/ml stock solution of BSA. After the preparation of diluted BSA standard, microplate protocol was followed and the samples were measured 10 minutes after addition of Bradford reagent. The concentration was determined by measuring absorbance at 595 nm with spectrophotometer. Using the BSA standard curve, the equivalent protein concentration was thus determined.

3.5 Refolding of Suckerin Proteins

The concentration of the denatured proteins was adjusted to 0.1 mg/mL for all the refolding tests. The denatured proteins were dialyzed against an array of solutions such that the concentration of the denaturant decreases as it is buffer-exchanged. The components of the buffer were varied with pH, ionic strength, and additives like stabilizers, detergents, and chelators as mentioned. The buffers screened for renaturation of suckerin-19 protein and sepia suckerin-1 are mentioned in table 3.2 and 3.3 respectively. The unfolded protein sample with high denaturant concentration was placed in dialysis bag with molecular weight cut off: 6 to 8 kDa. The dialysis bag was then transferred into the buffer. The sample was then dialyzed with the buffer in 1:1000 ratios with overnight stirring at room temperature.

Table 3-2. Annotated list of buffers used to obtain soluble suckerin-19 protein

$\text{C}_2\text{H}_3\text{NaO}_2$: Sodium acetate

MES: (2-[N-Morpholino]ethanesulfonic acid)

HEPES: N-[2-Hydroxyethyl]piperazine- N' -[2-ethanesulphonic acid]

CAPS (3-[Cyclohexylamino]-1- propanesulfonic acid

NaCl: Sodium Chloride

pH	Buffer Composition						
4	0.05 M NaH_2PO_4	0.05 M NaH_2PO_4 0.1 M NaCl	0.05 M NaH_2PO_4 0.15 M NaCl	0.05 M NaH_2PO_4 0.2 M NaCl	0.05 M NaH_2PO_4 0.1 M L-Arg	0.05 M NaH_2PO_4 1% Glycerol	
4	0.05 M $\text{C}_2\text{H}_3\text{NaO}_2$	0.05 M $\text{C}_2\text{H}_3\text{NaO}_2$ 0.1 M NaCl	0.05 M $\text{C}_2\text{H}_3\text{NaO}_2$ 0.15 M NaCl	0.05 M $\text{C}_2\text{H}_3\text{NaO}_2$ 0.2 M NaCl	0.05 M $\text{C}_2\text{H}_3\text{NaO}_2$ 0.1 M L-Arg	0.05 M $\text{C}_2\text{H}_3\text{NaO}_2$ 0.05 M NaCl 1% Glycerol	0.05 M $\text{C}_2\text{H}_3\text{NaO}_2$ 1% Triton- X-100
4.5							
5							
5.5	0.05 M MES	0.05 M MES 0.1 M NaCl	0.05 M MES 0.15 M NaCl	0.05 M MES 0.2 M NaCl	0.05 M MES 0.1 M L-Arg	0.05 M MES 0.05 M NaCl 1% Glycerol	0.05 M MES 1% Triton-X-100
6							
6.5							
7	0.05 M Tris	0.05 M Tris 0.1 M NaCl	0.05 M Tris 0.15 M NaCl	0.05 M Tris 0.2 M NaCl	0.05 M Tris 0.1 M L-Arg	0.05 M Tris 0.05 M NaCl 1% Glycerol	0.05 M Tris 1% Triton-X-100
7	0.05 M HEPES	0.05 M HEPES 0.1 M NaCl	0.05 M HEPES 0.15 M NaCl	0.05 M HEPES 0.2 M NaCl	0.05 M HEPES 0.1 M L-Arg	0.05 M HEPES 0.05 M NaCl 1% Glycerol	0.05 M HEPES 1% Triton-X-100
10	0.05 M CAPS	0.05 M CAPS 0.1 M NaCl	0.05 M CAPS 0.15 M NaCl	0.05 M CAPS 0.2 M NaCl	0.05 M CAPS 0.1 M L-Arg	0.05 M CAPS 0.05 M NaCl 1% Glycerol	0.05 M CAPS 1% Triton-X-100

Table 3-3. Annotated list of buffers used to obtain soluble sepio monomer

$\text{C}_2\text{H}_3\text{NaO}_2$: Sodium acetate

MES: (2-[N-Morpholino]ethanesulfonic acid)

HEPES: N-[2-Hydroxyethyl]piperazine- N' -[2-ethanesulphonic acid]

CAPS (3-[Cyclohexylamino]-1- propanesulfonic acid

PBS: Phosphate buffer saline

NaCl: Sodium Chloride

pH	Buffer Composition		
4	0.02 M $\text{C}_2\text{H}_3\text{NaO}_2$		
4.5			
5			
5.5			
5.5	0.02 M MES		
6			
6.5			
7	0.02 M Tris	0.02 M HEPES	0.02 M PBS
7.5			
10	0.02 M CAPS		

The following day the dialysate was clarified by high speed centrifugation (14,000 rpm at 4 °C for 15 mins.) to remove insoluble protein fraction and the supernatant was stored in the separate tube. The pellet obtained was solubilized in 5% acetic acid, while the protein in the supernatant was concentrated using Vivaspin sample concentrator (GE Healthcare). The concentrated soluble fraction and the pellets dissolved in acetic acid were loaded on SDS-PAGE to infer the solubility/ insolubility. The recombinant suckerin protein concentration in the soluble supernatant fractions were quantified using Bradford assay described above.

3.6 DLS Measurements

Dynamic light scattering (DLS) analysis is routinely used in structural biology to detect aggregates in macromolecular solutions or to determine the size of proteins, nucleic acids, and complexes and also to monitor the binding of ligands. The sample is illuminated by a laser beam and the fluctuations of the scattered light are detected at a known scattering angle θ by a fast photon detector [135-137]. Simple DLS instruments that measure at a fixed angle can determine the mean particle size in a limited size range. From a microscopic point of view the particles scatter the light and thereby imprint information about their motion. Analysis of the fluctuation of the scattered light thus yields information about the particles.

Particle size dimensions of suckerin proteins were thus determined by DLS. The samples were analyzed with a 90Plus particle size analyzer (Brookhaven Instruments) equipped with a 658.0 nm monochromatic laser. To minimize the reflection effect, all measurements were taken at a scattering angle of 90°. Suckerin-19 was re-suspended in 5% acetic acid and water at 1 and 6 mg/mL or in 0.05 M MES, pH 5.5 at 0.64 mg/mL. Sepio monomer was analyzed at 1.5 mg/mL concentration in 5% acetic acid, water and 0.02 M sodium acetate, pH 5.0. The number-weighted histogram profiles of the proteins in solution are reported.

3.7 *CD Spectroscopy*

To examine folding, especially secondary structure of proteins, circular dichroism (CD) spectroscopy is an important technique in structural biology. CD refers to differential absorption of left and right circularly polarized light. When the chromophores of the amides of the polypeptide backbone of proteins (absorption below 240nm, far UV region) are aligned in arrays, their optical transitions are shifted or split into multiple transitions as a result of 'excitation' interactions. [138]. There is a weak but broad $n \rightarrow \pi^*$ transition centered around 220 nm and a more intense $\pi \rightarrow \pi^*$ transition around 190 nm [139]. Therefore, different types of regular secondary structures give rise to characteristic CD spectra in this far UV region. Proteins with β -sheet structure show one minimum at 217 nm and one maximum at 198 nm, while random coils only give one minimum at 198 nm [139].

Far-UV CD spectra of soluble proteins were collected using a Chirascan spectropolarimeter (Model 420, AVIV Biomedical Inc.). Quartz supracil cells (0.2 mm path length; Hellma Analytics) were used for all measurements. Measurements were made at 25°C, down to wavelengths between 190 nm to 260 nm, using a 1nm step size and 1 nm bandwidth, with instrument dynode voltage less than 600. In each case, baselines were buffer solutions in which the respective proteins were dissolved. Three scans were obtained for each sample and its respective baseline. The averaged baseline spectrum was subtracted from the averaged protein sample spectrum.

In the temperature dependent CD experiments, the sample was loaded into a quartz UV cuvette (0.2 mm) and was sealed by parafilm after the sample was loaded. The sample was cooled down in the CD spectrometer to 10 °C. Spectra were acquired from 260 nm to 190 nm at 1 nm resolution with an averaging time of 1 sec and repeated twice. The spectra were acquired from 10 °C to 90 °C with a 5 °C increment. The sample was equilibrated for 5 min at each temperature before the measurement. After heated to 90 °C, the sample was cooled down to 10 °C for 30 min and the spectrum was acquired again. Buffer background spectra were recorded and subtracted from the sample spectra. Spectra were smoothed using the Savitzky-Golay method [140] with a second order polynomial.

3.8 FTIR Spectroscopy

Fourier transform infrared spectroscopy (FTIR) spectroscopy measures the wavelength and intensity of the absorption of infrared (IR) radiation by a sample [141]. The absorption of IR radiation excites vibrational transitions in molecules. The vibrational frequency and probability of absorption depend on the strength and polarity of the vibrating bonds, which are influenced by intra-molecular and inter-molecular effects [142].

Nine characteristic group frequencies arise from the polypeptide repeat unit, and they have been identified as follows: amide A ($\sim 3,300\text{ cm}^{-1}$), amide B ($\sim 3,100\text{ cm}^{-1}$), amide I ($\sim 1,650\text{ cm}^{-1}$), amide II ($\sim 1,550\text{ cm}^{-1}$), amide III ($\sim 1,300\text{ cm}^{-1}$), amide IV ($\sim 735\text{ cm}^{-1}$), amide V ($\sim 635\text{ cm}^{-1}$), amide VI ($\sim 600\text{ cm}^{-1}$) and amide VII ($\sim 200\text{ cm}^{-1}$) [143]. The differential pattern in H-bonding and geometric orientations of amide bonds in α -helices, β -sheets, β -turn and random coil structures allow the different vibration frequencies to be associated with individual secondary structural folding [144]. The amide I and amide II bands are the two major bands in the protein IR spectrum. The most sensitive spectral region to protein secondary structure compositions is amide I, which originates from the C=O stretching vibration of the amide group coupled with the in-phase bending of the N-H bond and stretching of the C-N bond [143, 145]. These vibrations are found between $1,700$ and $1,600\text{ cm}^{-1}$, and they are directly related to the backbone conformation, each frequency corresponding to a particular secondary structure and conformational change in the protein.

The deconvolution of the amide I band was done using the second-derivative method to determine the number of bands and their positions. Based on this information the spectra were resolved into its components, by a commercial curve-fitting process. The fitting process is controlled by the characteristics of the individual bands, and by the position of the baseline. The band area for each component peak was used to calculate the relative contribution of component to a particular protein secondary structure. In water based solutions, bands between $1,650 - 1,658\text{ cm}^{-1}$ are expected for α -helical structures. The bands between $1,612 - 1,640\text{ cm}^{-1}$ and $1,670 - 1,690\text{ cm}^{-1}$ are assigned to antiparallel β -sheet components, while $1,626 - 1,640\text{ cm}^{-1}$ are assigned to parallel β -sheets. The bands located between $1,655 - 1,675\text{ cm}^{-1}$ and $1,680 - 1,696\text{ cm}^{-1}$ are

associated with β -turns. Non-ordered conformation, usually refer to as random coils is associated with an IR band between 1,640 - 1,651 cm^{-1} [142, 143, 146].

FTIR measurements were performed with a Bruker Vertex 70 infrared spectrometer equipped with liquid- N_2 cooled mercury cadmium telluride (MCT) detector. The instrument was continuously purged with liquid- N_2 . For all soluble proteins, spectrum was recorded with 4 mins scans in ATR (attenuated total reflection) mode with a diamond crystal as substrate in the region of 4,000 - 400 cm^{-1} at room temperature. The respective solvents were used as the controls. In order to estimate the percentage of different secondary structures, the amide I band (1600-1700 cm^{-1}) was deconvoluted using OPUS 6.5 software (Bruker). The positions of the fitting peaks were fixed based on peaks identified in the second derivative of amide I. In this case, the widths of the fitted peaks were approximately the same.

3.9 TEM

Transmission Electron Microscopy (TEM) in high resolution and positive/negative staining is useful to study conformations of proteins, biological polymers, and small molecules. It can be used to interpret complex aggregation phenomena and to increase our understanding insoluble protein aggregates consisting of tightly packed β -sheets, which are not amenable to conventional structural biology techniques.

TEM uses electron beams to produce images in an analogous fashion to photon beams in a light microscope. TEM beams pass through the sample interacting with the specimen to produce images or diffraction patterns. TEMs are capable of imaging at a much higher resolution than light microscopes, owing to the small de Broglie wavelength of electrons, enabling examination of fine details. While the amount of radiation required to collect an image from the specimen can potentially damage the delicate structure of proteins, additional high vacuum can also create harsh environment for the sample. The later issue can be resolved using negative stains, however even after chemical staining and fixing of the protein samples structural collapse upon dehydration can be expected [147]. It is thus critical to keep these points into consideration while analyzing the data.

Aliquots of 4 μL of various measured concentrations of recombinant suckerin-19 were applied to glow-discharged carbon coated grids (Ultrathin Carbon Film on Holey Carbon Support Film, Ted Pella, California, USA) adsorbed for 30 seconds and then fixed with 2% Glutaraldehyde (Grade-I, sigma) for 30 seconds. Excess solution was removed using filter paper and the samples were stained with 4 μL 2% Phosphotungstic acid (PTA) for 45 seconds. The excess liquid was removed by filter paper and the grids were subsequently air dried at room temperature for an hour. TEM examination was carried out with an energy filtered Carl Zeiss LIBRA 120 Plus TEM microscope equipped with an in-column Omega spectrometer, at an acceleration voltage of 120 kV. Different acquired images were analyzed using ImageJ (NIH) to measure the mean diameter of conformations observed.

3.10 Theoretical Suckerin Protein Structure Prediction

Using various computational methods diverse questions can be addressed efficiently to understand the structural design at the fundamental levels of hierarchy of biological materials. Based on the questions to be addressed, a suitable model can be opted which can be defined at the sub-atomic, molecular, or higher level of hierarchy. Using computational methods the tertiary structure of a protein can be predicted from its linear amino acid sequence (primary structure.) Before predicting the tertiary fold, secondary structure prediction is a natural intermediate step that is applied to uncover the local conformations. The following sections discuss in brief the methods applied for predicting computational model of suckerin-19.

3.10.1 Secondary Structure Prediction

Secondary structure predictions can be useful to discover the local conformations of regions in the given protein sequence. It precedes any informed guesses of the target sequence's 3D fold. Secondary structure prediction is also significantly easier because it essentially predicts 1D structural data from sequence (assignment of local state to each residue), as opposed to tertiary structure prediction, wherein 3D coordinates are predicted from the sequence.

There are some best performing methods implemented as fully automated servers that can be used for secondary structure predictions. The performance of such servers are assessed in the Critical Assessment of protein Structure Prediction (CASP)

experiments [148]. CASP assesses based on large number of blind predictions of the protein structures. The secondary structure predictions were performed using PSIPred [149] and JPred [150] that have been evaluated as best servers in CASP experiments. To obtain an unbiased consensus a third server was used, namely GOR IV, which was selected on the basis of its high Q3 score of (another common accuracy measure). The Q3 score [151, 152] is the average of each Q_i (i = helix, sheet, loop), where Q_i is defined as the percentage of correctly predicted residues in state i to the total number of experimentally observed residues in state i . A consensus representation of the secondary structure prediction from the three servers was built for further analysis.

3.10.2 Tertiary Structure Prediction

In structural bioinformatics, the main challenge is to understand how the information encoded in the linear amino acid sequence is translated into a 3D structure. Based on the acquired knowledge computational methodologies and algorithms have been proposed, tested and analyzed that can correctly predict the native structural fold [153]. Broadly, 3D structure prediction method can be divided into two extremes: homology modeling [154, 155] and *ab-initio* methods [156-158].

Homology Modeling

Homology modeling is also known as comparative modeling. It is based on the observation that the proteins with similar sequence have similar tertiary structure [159, 160]. As a result, the tertiary structure of a protein can be built using the templates of the known structures that shares significant similarity with the target sequence. If the similarity is more than 30% then the three-dimensional structure is assumed to be the same [161]. In order to search for a structural template to model suckerin-19, an extensive homology search was done using the BlastP server [162] against the Protein Data Bank [163] (PDB). Since no matched reference structures were obtained, *ab initio* modeling method was subsequently applied for tertiary structure prediction.

Ab initio modeling

Ab initio prediction method seeks to predict the tertiary structure of a protein from its amino acid sequence alone, without the knowledge of similar folds. Also known as *de novo* modeling [164], this approach predicts the tertiary fold using physical chemistry principles as one of its goals. It is mainly based on the thermodynamic hypothesis

[165] which states that the tertiary structure of the protein is the conformation with the lowest free energy. It thus relies on the structure construction strategy and the energy functions used to discover the native structure [166]. However, *ab initio* modeling is challenging.

The tertiary fold of the full-length suckerin-19 protein was predicted using *ab initio* modeling methods that are implemented as fully automated servers. Regardless of the methodology implemented, all the developed 3D structure prediction methods are tested for their ability to predict new protein structures. The CASP experiment [167, 168] provides an independent assessment of the state-of-the-art protein structure modeling. Among the fully automated servers benchmarked in CASP experiments [148], three methods have been selected for *de novo* structure prediction of suckerin-19. Robetta [169], I-Tasser [170] and RaptorX [171] have been listed among the best CASP-certified structure prediction servers and have been used in this study.

Robetta server [172] provides both *ab initio* and comparative models of protein domains. Domains without a detectable PDB homolog are modeled with the Rosetta *de novo* protocol [173]. In the Rosetta algorithm [174], the distribution of conformations observed for each short sequence segment in known protein structures is taken as an approximation of the set of local conformations that the sequence segment would sample during folding. The program then searches for the combination of these local conformations that has the lowest overall energy.

RaptorX server [171, 175] has been specifically designed to address cases where no close template exists. It takes into consideration the number of non-redundant homologs available for the target sequence and a template structure to assess the quality of information content in sequence profiles. The server uses conditional neural fields (CNF), to integrate a variety of context-specific biological signals in a nonlinear probabilistic scoring function [176]. Finally, it implements a multiple-template threading (MTT) procedure [177], enabling the use of multiple templates to model a single-target sequence.

I-Tasser suite [178-180] is another successful free modeling approach that uses a knowledge-based energy to construct 3D models. The used energy terms include information about predicted secondary structure, backbone hydrogen bonds, consensus predicted side chain contacts, a short-range correlation, and hydrophobic interactions. In this model, the authors used both threading, to search for possible folds first, and *ab initio* modeling to reassemble full-length models, and build the unaligned regions. The quality of the predicted models can be evaluated based on the confidence score (C-score) [178] provided for each model. It is calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations. C-score is typically in the range of [-5,2], where a C-score of higher value signifies a model with a high confidence and vice-versa.

3.10.3 Model refinement using MD Simulations

A molecular dynamics (MD) simulation solves the Newton's equation of motion at each time step of atom movement and is considered to be a more realistic method of depicting what is occurring in the protein atomistically [181, 182]. Due to long simulation time required to study protein folding pathways, currently no all atom MD simulations are performed for protein structure prediction starting from either an extended or random initial structure [183]. However, MD simulations can be used for structure refinement when a low resolution model is available, as conformational changes are assumed to be small [164]. Therefore suckerin-19 *ab initio* models predicted were further refined using MD simulations to illustrate and characterize a more native-like conformation of the dynamic protein. The following sections provide a brief overview on the theory of MD simulations.

Equation of Motion

MD simulations basically record time dependent behavior of molecular system where the fluctuations and conformational changes of biomolecules are observed. It numerically solves Newton's second law of motion i.e.,

$$F = Ma = M \left(\frac{dv}{dt} \right) = M \left(\frac{d^2r}{dt^2} \right) \quad \dots(3.1)$$

Where “ F ” is the force exerted on particle of mass “ M ” and “ a ” is acceleration; “ v ” is the velocity and “ r ” is position of the particle to observe structural fluctuations with respect to time “ t ”.

For small time interval dt , the above equation can be simplified as,

$$dv = \left(\frac{F}{M}\right) dt \rightarrow v_2 = v_1 + \left(\frac{F}{M}\right) dt \quad \dots(3.2)$$

$$r_2 = r_1 + v_2 dt = r_1 + v_1 dt + \left(\frac{F}{M}\right) dt^2 \quad \dots(3.3)$$

From equation 3.2 and 3.3, it can be deduced that if the a particle is at position “ r_1 ” and has attained acceleration “ a ”, due to the applied force “ F ”, the particle will acquire new position “ r_2 ” and hence the velocity “ v_2 ” of the particle at time “ $t+dt$ ” can be predicted using these equations. Therefore the method can determine the futuristic state of the system based on the current state.

As this example is valid for independent particles, however in case of proteins, a specific atom at any given time is in interaction with other atoms through either covalent or non-covalent interactions. Thus, the force experienced by the atom is due to its interaction with other atoms and so on. Therefore in such model systems, potential energy of the system is first calculated which is subsequently differentiated to obtain the forces experienced between the particles or atoms using the equation mentioned below.

$$\frac{-dE}{dr_i} = F_i \quad \dots(3.4)$$

Wherein “ E ” is the potential energy function and “ F_i ” is the force experienced by the “ i^{th} ” particle or atom in the system.

Potential Energy Function and Force fields

In classical a MD simulation that involves system of interacting particles, the molecular mechanics force field defines the force of between the atoms and the potential energy. This potential energy (E) function can be fractionated into bonded and non-bonded energy components as;

$$E_{total} = E_{covalent} + E_{non-covalent} \quad (3.5)$$

The covalent and non-covalent contributions can further be decomposed into different contributions:

$$E_{covalent} = E_{bond} + E_{angle} + E_{dihedral\ term} \quad (3.6)$$

$$E_{non-covalent} = E_{electrostatic} + E_{van\ der\ Waals}$$

Factors like bonds, bond angles, bond dihedrals, non-bonded interactions and electrostatic interactions sum up to define the empirical potential energy function [184]. The given equation shows the functional form of a typical force field:

$$V(r^N) = \sum_{bonds} \frac{k_i}{2} (l_i - l_{i,0})^2 + \sum_{angles} \frac{k_i}{2} (\theta_i - \theta_{i,0})^2 + \sum_{torsions} \frac{v_n}{2} (1 + \cos(nw - \gamma)) \\ + \sum_{i=1}^N \sum_{j=i+1}^N \left(4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \right) + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \quad (3.7)$$

Where k_i is the force constants, l is the bond length, l_o is the equilibrium bond length, θ is bond angle, θ_o is equilibrium bond angle, v_n is n^{th} order force constant for dihedrals, w is dihedral angle, γ is phase angle, σ_{ij} is the distance at which $E_{vanderWaals} = 0$, q is atomic charge and r_{ij} is interatomic distance.

For studying MD simulations, some of the classical force fields include AMBER [185], CHARMM [186], GROMOS [187] and OPLS [188] have been used. In this work, simulations have been performed using the AMBER package.

Integration of equations of motion

In a bio-macromolecular system, the motion of each individual atom is coupled to other atoms as they experience respective inter-atomic interactions. This gives rise to continuous potential and many body problems that cannot be solved analytically. Therefore to generate a MD trajectory with continuous potential models finite difference techniques are used and are assumed to be pairwise additive. The basic strategy thus applied in MD simulations can be summarized as follows:

- Compute the potential energy of the system using the potential energy function as mentioned above.
- From potential energy, compute the force experienced by the individual particles/atoms (due to all interactions with other atoms) in the system (Using equation 3.4).
- Integrate the equation of motion to predict the position and velocity of the atoms at a later time “ $t + \delta t$ ” (Using equation 3.2 and 3.3)
- Repeat the above steps “ n ” number of times to generate information on the position as well as velocity of the system within the given time step. The resulting set of (discrete) coordinates (trajectory) for each atom (particle) is considered as an approximation to the “real” path the atom takes in time. This is known as a MD trajectory [189] as shown below.

Several algorithms are used for integrating the equation of motion using finite difference method (mentioned above). In all the algorithms it is assumed that the positions and dynamics properties like velocities and accelerations etc. can be approximated as the Taylor series expansion:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2} \delta t^2 \mathbf{a}(t) + \frac{1}{6} \delta t^3 \mathbf{b}(t) + \frac{1}{24} \delta t^4 \mathbf{c}(t) + \dots \quad (3.8)$$

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \delta t \mathbf{a}(t) + \frac{1}{2} \delta t^2 \mathbf{b}(t) + \frac{1}{6} \delta t^3 \mathbf{c}(t) + \dots \quad (3.9)$$

$$\mathbf{a}(t + \delta t) = \mathbf{a}(t) + \delta t \mathbf{b}(t) + \frac{1}{2} \delta t^2 \mathbf{c}(t) + \dots \quad (3.10)$$

$$\mathbf{b}(t + \delta t) = \mathbf{b}(t) + \delta t \mathbf{c}(t) + \dots \quad (3.11)$$

Where $\mathbf{r}$ is the position, $\mathbf{v}$ is the velocity (the first derivative of the positions with respect to time), $\mathbf{a}$ is the acceleration (the second derivative with respect to time), $\mathbf{b}$ is the third derivative and so on.

Several integrators which are currently used include *Verlet algorithm* and its variations, the *leap frog algorithm* and the *velocity Verlet* [190].

The *Verlet algorithm* uses positions and accelerations at time t and the positions from time $t - \delta t$ to calculate new positions at time $t + \delta t$. The relationship between these quantities and the velocities at time t can be written to derive the *Verlet algorithm* as follows:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2} \delta t^2 \mathbf{a}(t) + \dots \quad (3.12)$$

$$\mathbf{r}(t - \delta t) = \mathbf{r}(t) - \delta t \mathbf{v}(t) + \frac{1}{2} \delta t^2 \mathbf{a}(t) - \dots \quad (3.13)$$

These two equations can be added to give the following:

$$\mathbf{r}(t + \delta t) = 2\mathbf{r}(t) - \mathbf{r}(t - \delta t) + \delta t^2 \mathbf{a}(t) \quad (3.14)$$

The advantages of the *Verlet algorithm* are, (a) it is straightforward, and (b) the storage requirements are modest. The disadvantage is that the algorithm is of moderate precision as the positions $\mathbf{r}(t + \delta t)$ are obtained by adding a small term ($\delta t^2 \mathbf{a}(t)$) to the difference of two much larger terms, $2\mathbf{r}(t)$ and $\mathbf{r}(t - \delta t)$. The *Verlet algorithm* uses no explicit velocities and therefore it is difficult to obtain the velocities, which are not available until the positions are computed.

The *leap-frog algorithm*, uses the following relationship:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}\left(t + \frac{1}{2} \delta t\right) \quad (3.15)$$

$$\mathbf{v}\left(t + \frac{1}{2} \delta t\right) = \mathbf{v}\left(t - \frac{1}{2} \delta t\right) + \delta t \mathbf{a}(t) \quad (3.16)$$

Herein, the velocities $\mathbf{v}(t + \frac{1}{2} \delta t)$ are first calculated from velocities at time $(t - \frac{1}{2} \delta t)$ and acceleration at time t . These are used to calculate the positions, $\mathbf{r}$, at time $(t + \delta t)$ using equation 3.15. The velocities at time t can be calculated from

$$\mathbf{v}(t) = \frac{1}{2} \left[\mathbf{v}\left(t + \frac{1}{2} \delta t\right) + \mathbf{v}\left(t - \frac{1}{2} \delta t\right) \right] \quad (3.17)$$

In this way, the velocities ‘leap-frog’ over the positions, then the positions leap over the velocities. The advantage of this algorithm is that the velocities are explicitly calculated and it does not require the calculation of the differences of large numbers. However, the disadvantage is that the velocities are not calculated at the same time as the positions.

The *velocity Verlet method* yields positions, velocities, and acceleration at time t and there is no compromise on precision:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2} \delta t^2 \mathbf{a}(t) \quad (3.18)$$

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \frac{1}{2} \delta t [\mathbf{a}(t) + \mathbf{a}(t + \delta t)] \quad (3.19)$$

It is implemented as a three-stage procedure (as seen in equation 3.19) wherein new velocities are calculated based on acceleration at both t and $t + \delta t$. Hence, in the first step the positions at $t + \delta t$ are calculated based on equation 3.18 using velocities and acceleration at time t . The velocities at time $t + \frac{1}{2} \delta t$ are then determined using:

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \frac{1}{2} \delta t \mathbf{a}(t) \quad (3.20)$$

Then new forces are derived from the current positions giving $\mathbf{a}(t + \delta t)$. In the final step, velocities at time $t + \delta t$ are calculated using:

$$\mathbf{v}(t + \delta t) = \mathbf{v}\left(t + \frac{1}{2} \delta t\right) + \frac{1}{2} \delta t \mathbf{a}(t + \delta t) \quad (3.21)$$

MD simulation parameters for Suckerin-19

MD simulations were performed using the AMBER12 [191] ff99SB force field [192]. For simulations of flanked proteins, the partial charges and the force-field parameters were generated using the *Antechamber* module in AMBER. The *Xleap* module of AMBER was used to prepare the system by adding hydrogen atoms. The terminal amide and carboxyl functional groups were neutralized using “ACE” and “NME” capping. The atomic motions implemented in AMBER are followed using the velocity Verlet algorithm [193].

Biomolecule are generally simulated in the environment in which they perform their functions. Thus to mimic realistic conditions water models were developed which use the approximations of molecular mechanics. In our simulations, the structures were solvated using a box of TIP3P [194] water molecules. (However, for simulations on peptides, in which the structures were initially in the extended conformation, we first performed implicit solvent simulations for 100 ns and the subsequently the conformation obtained at the end of the simulation was used for further explicit solvent simulations following the protocol mentioned below) An octahedral box was constructed such that a minimum distance of 10 Å was ensured between any atom of

the protein and the boundary of the box. An appropriate number of Na^+ and Cl^- counter ions were further added to each system, to ensure net charge neutrality. All short-range, non-bonded van der Waals interactions were truncated at 9 Å and the particle mesh Ewald method [195] was used to model the long-range electrostatic interactions. The SHAKE algorithm [196] was used to constrain bond vibrations involving hydrogen atoms. Solvent water molecules and counter ions were initially relaxed by energy minimization, keeping the protein atoms restrained. Relaxing the entire system using energy minimization to relieve any steric clashes followed this step. In order to accurately simulate the experimental conditions the physical conditions like pressure, temperature etc., all the simulations in performed at 300 K (NVT canonical ensemble [197]). Careful energy minimisation and finite temperature equilibration for all structures was performed before the final production run.

MD simulations were thus performed using the following protocol: an equilibration phase where harmonic restraints were placed on the $\text{C}\alpha$ atom with a force constant of $25 \text{ kcal mol}^{-1} \text{ \AA}^2$ and the system gradually heated from 50 to 300 K for a period of 250 ps. This was followed by a gradual reduction of the restraints over the next 250 ps until the restraints were reduced to 0; the system was allowed to evolve at 300 K over the next 2,000 ps without restraints. Finally, the full-length equilibrated structures were subjected to three independent (different initial velocities) 1 μs molecular dynamics simulations (totaling 3 μs for each system). The flanked peptides simulations were performed for 100 ns. The conformations were saved every 10 ps.

Analyses of the simulations were carried out using the *ptraj* module. Before the analysis, the conformations generated in molecular dynamics simulations were centred and root mean square (r.m.s.) fitted to remove overall translation/rotations. The conformations sampled during the simulations for all the three models were stable as judged by the temporal evolution of r.m.s. deviation of $\text{C}\alpha$ atoms. The simulations reached an equilibrium state within ~40 ns in all the simulations, and therefore the analysis was carried out over the equilibrated trajectory. The SASA, over the trajectory for each amino acid was computed by assigning each atom an average radius of 2 Å and by summing up the total surface area. The trajectories were viewed in VMD [198] and the images were also created using VMD.

3.11 X-ray Crystallography

Many advances in the understanding of biological systems at the molecular level have been made possible through knowledge of the detailed structures of proteins. X-ray crystallography is one the most popular and reliable experimental method used currently for tertiary structural determination [199-201]. A crucial requirement for this work is the production of macromolecular crystals of suitable size and perfection for X-ray. Growing crystals is often the major bottleneck in structure determination [202]. The section below describes the crystallization trials for sepio monomer construct. These were undertaken with an aim to obtain crystals that could provide structural information at atomic-resolution.

3.11.1 Crystal Screening

The crystallization process involves three steps [202]. The first step is nucleation in which the protein molecules associate together providing a foundation for the growth of a crystal. The next step is the growth of the crystal itself. The final step is when the crystal ceases to grow usually as a result of lack of free protein molecules within the solution or the presence of contaminants that hampers the growth of the crystal.

Commercial crystallization screens are especially designed to help identify initial crystallization conditions for any target protein [203]. Multiple crystallization conditions were screened using available commercial crystallization screens (Morpheus® HT-96, Molecular Dimensions; Crystal Screen #1 & #2 and Index HT, Hampton Research). Automated crystal screening was performed with a Mosquito (TTP Labtech) nanolitre crystallization robot in a vapor hanging-drop format using Nunc™ 96 well plates. An equal volume of purified protein was mixed with a crystallization reagent solution (100 + 100 nL drops) with 90 µL buffer in the well. Following set-up, the plates were incubated at the room temperature and later analyzed for crystals using a light microscope. The identification of conditions suitable for crystal growth takes place using a trial and error approach and represents one of the biggest obstacles in X-ray crystallography. Unfortunately, no precipitation was observed in any screen over a period of 4 weeks.

3.12 NMR spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy is a unique method among the available tertiary structure determination methods, since NMR data can be acquired in solution. It can provide structural insights into partially folded polypeptide chains (which are difficult to crystallize) [88, 204-206]. The ability of NMR to capture structural information from the dynamic regions is critical in our case, as the Gly-rich modules in suckerin proteins are proposed to predominantly conform into an amorphous matrix (disordered regions).

3.12.1 One dimensional NMR spectroscopy

Similar to all forms of spectroscopy, NMR spectra can also be considered to arise from transitions made by atomic nuclei between different energy states. Additionally for macromolecules, nuclei in different chemical environments (e.g., the different ^1H nuclei) will resonate at different frequencies. A plot of the intensity against these resonance frequencies is known as the 1D NMR spectrum. These response frequencies are often reported as chemical shifts as well. The 1D ^1H -NMR spectrum is generally the first to be recorded which simply shows the signals for each proton (hydrogen atom) in a biomolecule [207]. This preliminary measurement can provide valuable information about the folded state or the aggregated state of the protein. Unfolded proteins show poor dispersion, since all the amide protons will be in the similar chemical environment (i.e., exposed to solvent). Also, self-association or oligomerization also broadens all signals. Acquiring 1D spectra on a sample over a period of time can also be useful to determine the stability of the sample in the given solution [207, 208]. To assess the presence of these structural conformations, 1D NMR was recorded for suckerin-19 and sepio monomer in various conditions and their respective solution behaviors were reported.

The standard 1D NMR experiment has only one frequency dimension. However for larger molecules, like proteins, this method can result in a very complicated spectrum. The poor dispersion of signal results severe spectral overlap which is difficult to analyze [209]. To overcome this problem, 2D NMR and 3D NMR experiments are performed which have second and third frequency dimensions respectively. Thus, the spectra acquired have increased resolution in which overlapping resonances can be

separated and distinguished. While the resolution of the acquired spectrum can be increased by an additional spectral dimension, homonuclear (signals of the same isotope) experiments can still be challenging for larger proteins. Homonuclear correlation spectrum of macromolecules larger than M.W. 10 kDa exhibit poor signal to noise ratio and dizzying number of cross peaks with decreased sensitivity, thereby limiting the analysis [209].

3.12.2 Heteronuclear Correlation Spectroscopy

Apart from the protons, proteins also contain other magnetic active nuclei such as the ^{13}C and ^{15}N . Using these hetero nuclei allows additional features to be detected in NMR, which facilitate the structure determination, especially for larger proteins (> 10 kDa). The natural abundance of ^{13}C and ^{15}N is very low (1.1% and 0.4%, respectively) and their gyromagnetic ratio is markedly lower than the protons. Therefore for increasing the sensitivity of these nuclei two strategies are used. Firstly, the recombinant protein is expressed in a defined minimal medium isotopically enriched for these nuclei (e.g. ^{13}C -glucose and $^{15}\text{NH}_4\text{Cl}$) during. Secondly, the signal to noise ratio is enhanced using inverse NMR spectra in which magnetization is transferred from proton to the hetero nucleus [210].

3.12.2.1 Two Dimensional ^{15}N -HSQC

The ^{15}N -HSQC (^{15}N -Heteronuclear Single-Quantum Coherence) is one of the most frequently used experiments in NMR. It provides correlation between the amide proton and its covalently bonded nitrogen [211]. In general, each residue produces an observable peak in the spectra corresponding to a particular amide nitrogen/hydrogen pair, with exception of Pro, which lacks an amide proton. Additionally, side chain (NH groups) can also be detected with the terminal amine group present in glutamine and asparagine residues displaying two additional peaks in spectra with same nitrogen shifts [209]. The ^{15}N -HSQC is basically like the fingerprint of proteins and is used in this study for characterization of dynamic suckerin protein.

3.12.2.2 Three Dimensional Heteronuclear NMR

Triple resonance experiments are used to assign resonances of larger proteins (>10 kDa). In 3D experiments, each peak is designated by three coordinates relating to the chemical shifts of three nuclei (for an isotopically enriched protein, a combination of ^1H , ^{13}C and ^{15}N). Each experiment exploits different path of magnetization transfer and thus provides different structural information. Details of the pulse sequence are not discussed here; rather the focus is on the information each experiment provides. The following sections describe 3D NMR experiments recorded to assign sequential backbone chemical shifts (HNCACB, CBCA(CO)NH and HNCA), side-chains (^{15}N -TOCSY-HSQC and HCCH-TOCSY) and to obtain restraints for structure calculations (^{15}N -NOESY-HSQC) [212].

HNCACB

The HNCACB experiment [213] correlates the $^{13}\text{C}^\alpha$ and $^{13}\text{C}^\beta$ resonances with the amide ^1H , ^{15}N resonances of the same residue and the preceding residue (i.e., i and $i-1$) [209]. Thus, $^{13}\text{C}^\alpha$ and $^{13}\text{C}^\beta$ shifts provide information on the amino acid identity and the sequential connectivities are also indicated in the spectrum. Interpretation of the spectrum is further aided by the fact that the $^{13}\text{C}^\alpha$ resonances are 180° out of phase when compared to $^{13}\text{C}^\beta$ resonances, allowing them to be distinguished easily [213]. Similar to other 3D spectrums, the HNCACB spectra comprises of a 3D cube of data. However, in practice, so called “strips” of the spectrum are created to analyze with each corresponding cross-peak from the ^{15}N HSQC experiment. For this, the ^{15}N HSQC is edited, with the projection along the ^{13}C dimension being a ^{15}N HSQC. Each strip in the spectrum contains four peaks, one for each of the $^{13}\text{C}^\alpha$ and $^{13}\text{C}^\beta$ shifts of residue i and $^{13}\text{C}^{\alpha-1}$ and $^{13}\text{C}^{\beta-1}$ for $i-1$ residue (where applicable). The Gly residues can be identified easily since they lack $^{13}\text{C}^\beta$. Ala, Ser and Thr residues can also be readily identified easily due to their characteristic $^{13}\text{C}^\alpha$ and $^{13}\text{C}^\beta$ chemical shifts [209].

CBCA(CO)NH

The CBCA(CO)NH experiment [214] is similar to the previous experiment and together they are the two most vital experiments for backbone assignment. The data obtained from this experiment again comprises of ^1H and ^{15}N correlations with the $^{13}\text{C}^\alpha$ and $^{13}\text{C}^\beta$ of the preceding residue only. These inter-residue correlations are obtained

using ^{13}C spin as a means of coherence transfer [209]. Therefore, each strip obtained from this experiment only contains two peaks, which correlate the amide ^1H and ^{15}N of i residue with the $^{13}\text{C}^\alpha$ and $^{13}\text{C}^\beta$ of the preceding residue.

HNCA

The HNCA experiment [215] was performed to distinguish possible ambiguities on which resonances originate from the i and $i-1$ residue, subsequently assisting in the backbone assignment with higher confidence. It provides correlation between ^1H and ^{15}N nuclei of the backbone amide group with the $^{13}\text{C}^\alpha$ chemical shifts of the i residue and the $i-1$ residue (basically $^{13}\text{C}^\alpha$ and $^{13}\text{C}^{\alpha-1}$). In the strip, the coupling to the directly bonded $^{13}\text{C}^\alpha$ is stronger and thus these peaks appear with greater intensity than those from the preceding $^{13}\text{C}^\alpha$. Generally using the strips from both HNCACB and CBCA(CO)NH spectra in concert allow the assignment of most of backbone amides resonances. Inter-residue connectivities in the HNCACB spectrum can be easily identified by overlaying the CBCA(CO)NH data as the same inter-residue peak will appear in both spectra. Following sequential backbone assignments using triple resonance experiments, the next step in structure determination pipeline is to assign the side chain assignments. This is done using ^{15}N -TOCSY-HSQC spectrum [216].

^{15}N -TOCSY-HSQC

The ^{15}N -TOCSY-HSQC (^1H - ^{15}N total correlation spectroscopy-HSQC) [217] allows intra-residue correlations between side chain protons to be observed. An isotropic mixing step allows magnetization to be transferred between the ^1H spins. Then the magnetization is transferred to the neighboring ^{15}N nuclei and back to ^1H for detection. The resulting strip shows side-chain proton resonances on the y-axis and amide (HSQC) resonances on the x-axis. The resonances are from the same residue as the amide. This spectrum can be used in conjunction with HCCH-TOCSY for an unambiguous side-chain hydrogen assignment.

HCCH-TOCSY

The HCCH-TOCSY (^1H - ^{13}C - ^{13}C - ^1H total correlation spectroscopy) [218] experiment is based on the 2D total correlation spectroscopy (TOCSY) experiment (in which x and y axis are ^1H) that is extended into a third dimension (^{13}C). This spectrum is used for side chain assignment. The magnetization is transferred from the side chain or the

backbone proton to the directly attached carbon, followed by the isotropic ^{13}C mixing and is finally transferred back to their attached proton for detection. The resulting strip has all side chain hydrogen resonances visible at each carbon frequency. ^1H detected dimension is displayed along the y - axis, therefore peaks belonging to one side-chain appear in a vertical strip.

^{15}N -NOESY-HSQC

Nuclear Overhauser effect spectroscopy (NOESY) experiments can be utilized for assignment of amino acid residues in a polypeptide and provide distance restraints for structure calculation. It identifies sequential through-space connectivities [212]. The spectrum obtained ^{15}N -NOESY-HSQC [216, 217] for consists of diagonal peaks and cross peaks; however the cross peaks connect resonances from nuclei that are spatially close rather than those that are chemically connected to each other. The relative intensity (I) of a NOESY cross-peaks is related to the inter-atomic distance (r) as $I \propto r^{-6}$. As a result, the NOE intensities dramatically decrease when the residues (dipolar coupled spins) are further apart. Therefore, the maximum observable spatial operation between the spins is approximately 5Å [211, 219].

3.12.3 NMR acquisition and processing

All NMR spectra were recorded on Bruker Avance 600 MHz NMR spectrometer, equipped with a cryo-probe and pulse field gradients. Data acquisition and processing were performed with Topspin software (Bruker) running on a Linux workstation. Spectra were calibrated with respect to internal DSS (2,2-dimethyl-2-silapentane 5-sulfonate sodium salt) (0 ppm) standard. All the experiments were performed at 298 K. WATERGATE (4) was applied for suppression of the water resonance in all the acquired NMR experiments. NMR data processing and analysis were carried out using Topspin (Bruker) and Sparky (T.D.Goddard and D.G.Kneller, University of California, San Francisco, CA) programs respectively.

3.12.4 Structure calculations

The information rich data obtained from the NMR assignment allow structural conclusions to be drawn for the target protein. The data analyzed in structure determination is discussed in the sections below.

3.12.4.1 Automated structure calculation by CYANA-FLYA

NMR structures were calculated using automated structure calculations by CYANA-FLYA program [220]. All of the backbone HNCA, HNCACB and CACBCONH spectra and side chain ^{15}N -TOCSY and ^{15}N -NOESY spectra were processed and calibrated according to root ^{15}N -HSQC spectra. Peaks from all of these spectra were picked using CARA (Computed Aided Resonance Assignment) [221] or SPARKY [222] program by setting appropriate tolerance level for each spectrum. Picked peaks were then converted into peaklists with relevant headers required for Cyana formatting.

CYANA-FLYA has inbuilt macros like ASSIGN, TALOS and RUNFLYA which will aid the structure calculation. ASSIGN macro assigns the peaks that were provided in the peaklists and TALOS generate the angle constraints required for structure calculation. All of the information obtained from these macros will be used in RUNFLYA to finally run the automated structure calculation. The Ramachandran plot [223] generated after the structure calculation was used as validation tool for further refinement of the structures thus obtained.

3.12.4.2 Secondary structure chemical shifts

In NMR experiments, there are strong correlations between the chemical shifts, which are influenced by the primary structure and the molecular conformation or the local structure. As individual nuclei will have different chemical shifts due to their dissimilar chemical microenvironment [211, 224], this highly sensitive information can be used to predict regions of regular secondary structure and the dihedral ϕ and ψ angles from chemical shifts.

^1H , ^{13}C and ^{15}N random coil chemical shift values have been determined for the twenty amino acids with no defined secondary structure [225]. The difference between the observed chemical shift and the equivalent random coil resonance is termed as the secondary structure shift ($\Delta\delta$). A positive $\Delta\delta$ indicates a downfield shift from the random coil values and vice versa. The $\Delta\delta$ plot against the amino acid sequence, can depict the regions of regular secondary structure by observing the pattern of $\Delta\delta$ [226].

It has been shown that β -sheets display $^1\text{H}\alpha$ and NH resonances which are downfield ($\Delta\delta > 0$) ($^1\text{H}\alpha$ shifts in the order of 0.37 ppm) and $^{13}\text{C}\alpha$ resonances up field ($\Delta\delta < 0$) from random coil [226]. On the other hand, α -helical structures display the converse, whereas flexible or unstructured regions exhibit little deviation from the random coil shifts ($\Delta\delta \approx 0$) [226].

3.12.5 Hydrogen/Deuterium exchange strategy

Analyzing the kinetics of hydrogen exchange in biological macromolecules can provide rich structural and dynamic structural information. Amide hydrogens on the protein backbone chain continuously and reversibly interchange with hydrogens in the surrounding solvent. For particular amide hydrogen, local factors like a steric blocking caused by neighboring amino acids and solvent accessibility can affect the exchange rate [227]. In a folded protein, hydrogens presents in the unstructured regions and lacking hydrogen-bonding will exchange rapidly; while those that are buried deep within the hydrophobic core or involved in hydrogen bonding will exhibit longer exchange rate [228]. Using high-resolution NMR site-specific hydrogen/deuterium exchange rates, which is a function of both the protein structure and solvent accessibility, can be measured [229]. The acquired data can be used to interpret protein structure and the dynamics at residue resolution.

4 Biophysical Characterization of SRT derived fiber and Suckerin protein.

4.1 Introduction

Nature applies the generic paradigm of using a limited number of universal building blocks, networks, motifs or modules to create robust structural materials such as silk, mussel's byssus, elastin or collagen with controlled and yet varying degrees of complexity. As silk science continues to evolve, it also forms a benchmark to assess newly discovered load bearing proteins like the suckerins.

As discussed in chapter 2 suckerins are also similar to silk in their mechanical properties and molecular design. They can be engineered for a wide range of biomedical applications and can be readily modified chemically or functionalized by genetic engineering. It is evident that in order to understand the mechanical, functional and other macroscopic properties of SRT, it is critical to gain a fundamental understanding of its nanoscale morphology. Towards this, using bottom up approach we aim to:

- (a) study the conformational propensity of the modular repeats that form the primary structure of suckerins and to predict their atomic structure by modeling and simulations approach,
- (b) experimentally characterize the structural morphology of a recombinant suckerin protein;
- (c) illustrate the molecular features of suckerin assembly in one of the processed materials at the nanoscale length scale.

A plausible model (Figure 4.1) suggesting the nanoscale molecular features of native SRT was proposed based on the X-ray data. It was reported that in native SRT, the β -sheets are randomly oriented and are embedded within an amorphous protein network exhibiting isotropic structural stability, which further assembled into nanofibrils. It was also noted that the nano-fibrillar lattice exhibits structural re-arrangement between 100 - 150 °C and remains stable up to 220 °C. Melting of β -sheets within the nano-fibrils was reported beyond this temperature range inducing the collapse of the nano-

fibrillar lattice. With this basic understanding of the molecular structural features and thermal stability of β -sheet at the nanoscale level, we compare the molecular organization of fibers re-processed from the SRT with the native SRT as discussed in the following section.

Figure 4.1. Structural features of SRT as obtained by *in situ* X-ray scattering (a) WAXS pattern acquired from the native SRT tip. **(b)** Scattering intensity plots at different temperatures derived by radial integration of WAXS patterns. Structural transformations can be observed at 80 °C (due to water removal), at 230 °C (due to β -sheets melting) and at 300 °C (due to polymer degradation). **(c)** SAXS pattern acquired from native SRT tip **(d)** Scattering intensity plots at different temperatures by radial integration (highlighted in red box in figure c) of SAXS patterns. **(e)** Schematic representation of the thermally induced structural transformation observed in SRT at multiple hierarchical levels (molecular to nanoscale). The structural transformations observed are reversible up to 220 °C; at this temperature the β -sheets melt (an irreversible transformation) and the nano-fibrillar lattice collapses. Source Ref. [129]

To further study the molecular features of the suckerins, a representative protein from the array of 21 similar proteins that assemble to form *D. gigas* SRT was selected. The rationale for the selection of suckerin-19 for structural characterization was based on the following reasons: suckerin-19 is one of the most abundant and highly expressed proteins in the suckerin tissue. Studies also show that recombinant suckerin-19 exhibits high solubility in mild acidic and aqueous buffers, which has been applied to establish relevant parameters to engineer suckerin-based materials by green processing. It has been shown that in 5% acetic acid and aqueous buffer suckerin-19 forms nanometer scale colloids that are stabilized by inter-molecular β -sheet interactions specifically at higher concentrations. These sub-micron structures which are similar to micro-spheres or colloids developed for silk and for generating β -sheet reinforced materials, can find potential applications in biomedical applications [119].

Figure 4.2. Biophysical characterization of suckerin-19 in solution (a) CD spectra at different concentrations of suckerin-19 in 5% acetic acid; (b) CD spectra at different concentrations of suckerin-19 in water; (c) DLS particle size distribution of suckerin-19 colloids in 5% acetic acid and water; (d) Particle size distribution of suckerin-19 colloids solubilized in water and determined by AFM in the dry state. Source Ref. [119]

4.2 Results and discussion

4.2.1 Biophysical characterisation of processed SRT fibers

In order to characterize the molecular organization in melt spun SRT fibers, the X-ray scattering profiles post processing (processing native SRT into fibers) were analyzed. High aspect ratio fibers (10 μm in diameter and few cm in length; refer chapter 2) were drawn from the crushed SRT powder dissolved in water and heated to obtain a viscous solution state. The resulting WAXS pattern of the drawn fibers was similar to the one obtained for the native SRT, suggesting that the β -sheet organization is preserved during mechanical agitation and thermal processing (Figure 4.3) [129]. Further examination of the spectra revealed a minor shift in the main reflections towards higher q -values, indicating small rearrangement of the β -sheets that seem to become more ordered in this case [129].

The SAXS pattern is also in accordance with the morphological conclusions derived from the WAXS data. A highly anisotropic pattern appeared due to the heterogeneous organization of elongated nano fibrils as seen in native SRT (Figure 4.3). This data suggests that the associated β -sheets are consistently maintained in the bi-phasic polymer network. The thermally stable β -sheets embedded in the amorphous lattice that is heat-resistant exhibit viscous flow below the melting point of this crystalline phase [129]. The viscous flow allows significant chain-rearrangement in the more dynamic/flexible amorphous domain and results in ordered elongated nano fibrils enriched with stiff β -sheet crystalline phase. The chain alignment might facilitate intermolecular interactions in the biphasic semi-crystalline SRT (for instance formation of additional H-bonds) which can mechanically reinforce the suckerin fiber

Although the diameter and the protein chain alignment in the SRT drawn fibers is similar to that seen in the native silk fibers, a separate study noted that the mechanical strength of suckerin fibers is lower than that of the post-processed regenerated silk fibers or the native silk fibers [230]. Thus it is essential to understand the organization of the biphasic polymer network at the fundamental level of hierarchy, such that the balance between the amorphous domains and crystalline β -sheet content in suckerins can be modulated to achieve targeted extensibility or tensile strength in the suckerin fibers.

Figure 4.3. Characterization of native SRT fiber using WAXS and SAXS (a) Small-angle X-ray scattering (SAXS) pattern and, (b) wide-angle X-ray scattering (WAXS) pattern (red) of a micro-scale fiber drawn from SRT powder that was initially heated under hydrated conditions in a microwave for 30 seconds. The resulting WAXS (red) pattern is compared with the WAXS pattern of the native SRT (black). Source Ref. [129]

4.2.2 Biophysical Characterization of Suckerin-19

In this stream of research suckerin-19 was characterized, to gain insight into its native conformation and molecular interactions. The understanding can be translated to interpret the mechanical properties of this material and for processing of bio-inspired SRT based materials. The following sections discuss the expression and purification methods for obtaining the recombinant suckerin proteins. The attempted strategies used to obtain soluble refolded proteins were guided by generic secondary structure analysis methods; including CD and FTIR. We also used TEM to identify certain self-assembled states acquired by the protein.

4.2.2.1 Primary Structure Analysis

Preliminary analysis of the amino acid sequence of suckerin-19 (GI: 551702290) without the signal peptide (since it is cleaved, once the protein is secreted) was performed to understand the primary structural features using ProtParam. Suckerin-19 consists of 398 amino acids with a theoretical molecular weight of 39.27 kDa. The calculated isoelectric point (pI) is 8.42, which can be explained with the presence of a few of basic residues. The most abundant residues are Glycine (32%) followed by Tyrosine (15%), Leucine and Alanine (11%) and Histidine (9%) and a complete

absence of negatively charged residues is seen (Figure 4.4a). For downstream applications such as calculation of protein concentration and refolding and quantitative studies, it is worthy to note that Tryptophan and Cysteine residues are absent in the protein sequence. The computed high aliphatic index (63.8) and a low instability index (II) (10.7) classified this protein to be thermally stable. While the thermal stability of the SRT block copolymer has been reported previously [129], it is noteworthy to assess the stability of this protein at higher temperatures.

4.2.2.2 Recombinant Expression and Purification

Suckerin-19 plasmid was transformed into *BL21* (DE3) strain of *E. coli* and was maintained with Ampicillin selection. Initial evaluation of the *E. coli* cell line expression products containing the recombinant suckerin-19 plasmid suggest that following induction the protein is present in IB (Figure 4.4b). Microfluidization has been successfully used for suckerin protein extraction with higher yields [119]. Therefore, this method was used for homogenized cell disruption. To ensure a high degree of purification of the protein, the crude IB preparation was washed with a buffer containing Triton X-100 and 2 M Urea. Addition of detergents and low concentrations of chaotropic compounds in wash buffers results in removal of contaminants such as bacterial membrane proteins that are non-specifically adsorbed onto the hydrophobic inclusion bodies during processing, and can ultimately affect protein refolding. A second wash step ensured proper purification of IB from other cellular components and complete removal of Triton X-100 for further processing. Following microfluidization, the protein was extracted from the IB in 5% acetic acid, followed by treatment with 8 M urea and 5% acetic acid (Figure 4.4b). The extracted protein was subjected to reverse phase purification by HPLC. The chromatogram indicated one major peak that spans an elution volume of 13 to 18 ml and constitutes the target protein as indicated by the SDS-PAGE analysis (Figure 4.4d). The molecular weight and the purity of the extracted protein was validated by MALDI-ToF [231]. An intense peak observed at 39.28 kDa corresponds to the predicted theoretical molecular weight of 39.27 kDa (Figure 4.4c). The total protein yield measured using Bradford assay [232] was ~15 mg/L culture. The purity was estimated to be greater than 90% and was considered suitable for subsequent solubility assays.

Figure 4.4. Primary sequence and recombinant expression of suckerin-19 (a) Large scale modular architecture of the primary sequence. Ala, His, Thr, Val-rich modules are highlighted in red and Gly, Tyr, Leu -rich modules are highlighted in blue. Source Ref. [119]; (b) SDS-PAGE of expression profile. 1, Molecular weight marker; 2, soluble bacterial proteins; 3, IBs expressing with suckerin-19 extracted in 5% acetic acid; 4, IBs expressing with suckerin-19 extracted in 8 M Urea + 5% acetic acid; (c) MALDI-ToF spectra of protein extracted in 5% acetic acid; (d) C8 reverse phase purification by HPLC. The SDS-PAGE of the elution showing pure suckerin-19 is shown in the inset (lane 2).

4.2.2.3 Solubility, Refolding and Secondary Structure Characterization

High-level expression of recombinant protein in *E. coli* either yields bioactive proteins as soluble proteins or often results in aggregation of the expressed protein molecules into inclusion bodies. Since suckerin-19 was expressed in IBs it is essential to develop an efficient native structure recovery procedure for the target protein based on IB refolding [233]. As there is no “universal” refolding buffer, optimal conditions for solubilization of each protein have to be determined [234]. The aim here was to establish favorable micro-environmental conditions for suckerin-19 that would enable it’s folding into the native β -structure as observed in the SRT and to further characterize it.

As a first step towards re-naturation, different screening conditions were tested to assess the protein. Various buffers and pH screening conditions were tested as shown in table 3.1 (in methods section). At physiological pH and in all buffered saline solutions, recombinant suckerin-19 protein aggregated suggesting that these specific conditions were unsuitable for its re-naturation. However following dialysis, in 50 mM sodium acetate and MES buffer at pH 5.0 and 5.5 respectively, recombinant suckerin-19 protein was soluble. Since the refolding analysis was performed at 0.1 mg/mL concentration of the protein, it was necessary to concentrate the sample for further structural characterization. While concentrating, the protein precipitated in sodium acetate (at 0.4 mg/mL), however it was predominantly soluble up to 0.65 mg/mL (=0.02 mM of suckerin-19) in MES buffer at pH 5.5.

Figure 4.5. Preliminary spectroscopic characterization of suckerin-19 in 50 mM MES at pH 5.5. (a) DLS particle size distribution; (b) CD spectra depicting β -sheet signature with a minima at 218 nm; (c) ATR-FTIR analysis, with deconvoluted amide-I. The positions of the fitting peaks were defined by the 2nd derivative of the amide spectra and the width of the peaks were fixed [119]. The protein concentration for all measurements was 0.64 mg/ml.

Under these conditions, the secondary structure of the protein was analyzed by far-UV CD spectroscopy. The spectra response yielded minima at 217-218 nm (Figure 4.5), which is typical of a β -sheet conformation. A maximum at 202 nm was observed in MES buffer, which was absent in the case of sodium acetate buffer. The BeStSel algorithm was used to obtain percentage estimates of the secondary structure. The CD

spectra of the recombinant suckerin-19 in MES buffer exhibited a typical β -sheet secondary structure (estimated β -sheet content: 48.1%), while the protein in sodium acetate has contained a comparatively lower amount of β -sheet structure (estimated β -sheet content: 23.2%). Due to higher solubility and β -sheet content in MES buffer, further characterizations of suckerin-19 were done in this buffer condition.

These results were further verified by ATR-FTIR measurements. Deconvolution of the amide I ($1600\text{--}1700\text{ cm}^{-1}$) band can be corroborated with the backbone conformation, providing the relative contribution of a proteins' secondary structure semi-quantitatively. The spectra were characterized by strong absorbance between $\sim 1623\text{--}41\text{ cm}^{-1}$ in the amide-I region that can be attributed to the β -sheet conformation. However, the proportion of secondary structures for recombinant suckerin-19 following standard peak deconvolution in MES buffer exhibited a relatively higher content of β -structure (β -sheet: 42.7% and β -turns: 18.0%) and a less abundant α -helix or random coil (Figure 4.5) when compared to the protein in sodium acetate (β -sheet: 25.7% and β -turns: 17.4%). DLS measurements indicated that the protein tends to oligomerize at low concentrations even in MES buffer ($D_H = 50\text{ nm}$) (Figure 4.5) and displayed a similar hydrodynamic radius as observed in aqueous buffer or 5% acetic acid at high protein concentration (Figure 4.2) [119].

The conformation stability of recombinant suckerin-19 in three different buffer conditions (50 mM MES, 5% acetic acid, aqueous buffer) was assessed using thermal CD scans. The CD spectra for suckerin-19 in MES buffer displayed the consistent presence of β -sheet structure across the tested temperature range of 10 to 90 °C, suggesting a high thermal stability of the β -sheets within the suckerin. The β -sheet structure was maintained even when the sample was cooled back to 25 °C, post heat ramping (Figure 4.6a). For protein in the mild acidic conditions and under similar thermal scan parameters, it was found that on cooling the sample to 10 °C, the minima shifted from 218 nm to 210 nm and the acquired conformation was retained as the temperature was raised. While minima at 210 nm cannot be corroborated as a random coil structure, it is suggested that at low temperatures in mild acidic conditions the protein assumes certain structural re-arrangement of β -structures that is irreversible in this case (Figure 4.6c). However, when the temperature was raised from 25 to 90 °C

the β -sheet signature was retained throughout and was conserved even when the sample was cooled back to 25 °C (Figure 4.6d). Additionally, the effect of temperature on the conformation of suckerin-19 in aqueous buffer depicts differences. The sample displayed β -sheet signature up to 60 °C, following which a further rise in temperature initiated structural re-arrangements as the spectra showed a significant shift in the minima towards 230 nm (Figure 4.6b).

DLS measurements previously suggested that at high concentrations the protein tends to oligomerize or form colloids in water (Figure 4.2) [119]. It is likely that as the temperature increases, the protein-bound water evaporates from the network causing stabilization of inter-chain hydrogen bonds that further induces formation of higher order oligomers, featuring minima at 230 nm. Thus, the observations noted are a consequence of decrease in water-protein interactions due to dehydration at higher temperatures. Here again, this effect seems to be reversible as the signal recorded at 25 °C (post temperature ramping measurements) from the heat resistant β -sheets in the bi-phasic protein network appeared with a minima at 218 nm. This thermal conformational change occurs through a β -dominant structural change. Similar observations were also reported during thermo-mechanical measurements (recorded between room temperature and 120 °C) of native SRT, wherein the semi-crystalline fibrils begin to re-arrange as the bound water evaporates from the protein network at higher temperatures [129].

Figure 4.6. Temperature dependent CD spectrum for suckerin-19 in solution. (a) 50 mM MES, pH 5.5, protein concentration 0.64 mg/ml; **(b)** water, protein concentration 5 mg/ml; **(c,d)** 5% acetic acid, protein concentration 5 mg/ml.

Based on these observations and the aforementioned studies undertaken to characterize the structural behavior of suckerin-19 in mild acidic and aqueous conditions, we proposed a model [119] suggesting oligomerization of the (rec)-suckerin-19 which is facilitated by the formation of inter- and intra-chain heat resistant β -sheets (Figure 4.7). The model illustrates that suckerin-19 adopts oligomeric conformations in tested buffers. This is enabled by inter-molecular β -sheet interactions induced at high protein concentrations (Figure 4.7). In 5% acetic acid suckerin-19 at 1mg/ml exists in a monomeric conformation due to the effect of low pH. The high concentration of hydrogen ions at low pH will result in more groups being protonated, considering the high percentage of His residues in M1 domains and because of the excess like charges on the chains/domains the molecules will repel each other and will exist as monomers in the acidic solution.

Figure 4.7. Schematic oligomerization state and structure of suckerin-19 in 5% acetic acid, water and 50 mM MES, pH 5.5, at different concentrations.

To further structural characterization of the protein, solubility is critical, as structural investigations often require very high concentrations of a homogenous protein sample. It is also known that solubility is influenced by extrinsic factors such as pH, temperature, ionic strength, and additives in the buffer. Here factors such as pH and various additives were tested and the protein was not soluble. The ionic strength of the MES buffer was thus varied to increase protein solubility. Decreasing the salt concentration can increase protein solubility as more water molecules can interact with the charged residues (because of the high abundance of His residues which are protonated at pH 5.5, suckerin-19 bears a net positive charge) to stabilize the protein conformation. We observed that the protein was soluble when the concentration of MES buffer is lowered to 2 mM, however reducing the ionic strength still could not enhance the protein solubility. In various concentrations of MES buffer, the distribution of hydrodynamic radius ranged from 20 nm to 60 nm suggesting that as the ionic strength of the buffer is increased in the range 5-10 mM, higher order protein oligomers were formed due to charge screening. As the ionic strength is further increased, the protein formed smaller oligomers with hydrodynamic radii of 25-30 nm (Figure 4.8). Since the secondary β -structure organization seems to be conserved at high temperatures in 50 mM MES and as the hydrodynamic radius of the protein is similar to those in other MES concentrations, the sample in this condition was used for further characterization.

Figure 4.8. Secondary structure characterization and particle size distribution of suckerin-19 in various concentrations of MES buffer at pH 5.5. (a) CD spectra of suckerin-19 in 5% acetic acid depicting random coil with minima at 198 nm (dotted black line) and in various concentrations of MES buffered at pH 5.5, showing β -sheet signature; (b) DLS particle size distribution of suckerin-19 in various MES concentrations buffered at pH 5.5.

4.2.2.4 NMR Characterization

It is instructive to compare our findings mentioned above using conventional ^1H - NMR spectra of suckerin-19 in different buffers as a probe to evaluate the protein behavior in solution, including its conformational stability and residual structure in these microenvironments. From the resonances obtained in mild acidic conditions and in 50 mM sodium acetate, poor chemical shift dispersion was seen (Figure 4.9). This is generally observed for unfolded proteins in which all the amide protons are exposed to similar chemical environments (basically the solvent) and there exists no specific structural organization [208]. But in the case of folded proteins, since the residues are exposed (or buried) in wide variety of microenvironments, based on their placements in the secondary-tertiary fold, the chemical shifts observed are well dispersed. Thus in these two conditions (50 mM sodium acetate and 5% acetic acid) the protein is suggested to be in an unfolded conformation (Figure 4.9).

Figure 4.9. 1D NMR spectrum of Suckerin-19 in 50 mM MES, pH 5.5 (blue), 20 mM sodium acetate, pH 5.0 (green) and 5% acetic acid (red).

However, in 50 mM MES buffer, a better dispersion of resonances was observed, indicating a partially folded conformation (Figure 4.9, top panel). The acquired ^{15}N -HSQC spectra in the same condition displayed a small percentage of peaks with severe overlap only in the amide region between 8.8-8.0 ppm (Figure 4.10). If the protein is well folded, the peaks are usually well dispersed with resonances that appear between 10-7 ppm (in the amide region), and most of the individual peaks can be distinguished. Large cluster of acutely overlapping peaks, observed only in the middle of the spectra, indicate the presence of a partially folded or an aggregated conformation. Since most peaks were overlapping in the acquired spectra, the resonances could not be assigned.

As mentioned in the previous sections, the CD spectra of suckerin-19 exhibited a β -sheet signature in 50 mM MES, pH 5.5 and DLS measurements revealed a large hydrodynamic radii of the protein (50 nm) [119]. Unexpectedly, it has been reported that, such partially folded proteins often exhibit substantial secondary structure as observed from CD spectroscopy and poor NMR dispersion indicates the existence of a

“molten globule” state [235]. This state is considered to be an intermediate conformational state, in which the protein has native-like secondary structure content, with little detectable tertiary structure and an increased solvent exposed hydrophobic surface area relative to the native state wherein the hydrophobic core is buried [236]. It is hypothesized that suckerin-19 in 50 mM MES, pH 5.5 displays similar features; it conserves a native-like structure but without a tightly packed protein interior, and hence not a well-defined tertiary fold. Due to this reason, the protein tends to exhibit self-association stabilized by inter-molecular β -sheet interactions. However, in this buffer condition it was not feasible to proceed further with detailed structural calculations of suckerin-19.

Figure 4.10. ^{15}N -HSQC spectra of suckerin-19 in 50 mM MES, pH 5.5 with poor peak dispersion from 8.8-8.0 ppm (in the amide region) suggesting severe overlapping of amino acids.

4.2.2.5 TEM characterization

Low solubility and the oligomeric state of the protein in the tested buffers proved to be challenging in order to obtain the (soluble) native state of suckerin-19 for experimental structural determination by NMR or X-ray crystallography. However, the morphology of suckerin-19 in mild acidic and MES buffers was subsequently explored using electron microscopy. Previous investigations have already identified formation of colloids in water by AFM studies [119]. TEM observations in mild acidic (5% acetic acid) conditions and at lowest concentration of MES buffer (2 mM) indicated predominant formation of spherical structures at low protein concentration (Figure 4.11b and 4.11c). Increasing the ionic strength of the buffer to 40-50 mM MES (Figure 4.11d-f) or increasing the protein concentration in mild acidic condition (Figure 4.11a) drove the formation of chain-like proto-fibrils that appear to be composed of spherical sub-units.

It is observed that in 5% acetic acid at higher protein concentration (5 mg/ml) the diameter of the particles (29.3 ± 2.1 nm; $n=10$) is almost double in magnitude to the diameter of particles (15.4 ± 1.6 nm; $n=10$) at low protein concentration (1 mg/ml) (Figure 4.11a and b). In case of MES buffer, we speculate that with higher buffer concentration the particles form higher order oligomers due to charge screening. As observed, the diameter of particles in 2 mM MES (Figure 4.11c) is 26.7 ± 3.1 nm ($n=10$) and as the buffer concentration is increased to 40 mM (Figure 4.11d), the diameter of particle size increases to 34.36 ± 3.7 nm ($n=10$). In case of 50 mM MES, the dimensions of two different conformations were observed (Figure 4.11e and f). The diameter of spherical or globular conformation of protein is 32.1 ± 2.1 nm ($n=10$) and diameter of the fibers is noted as 6.5 ± 1.3 nm ($n=10$). It is likely that in this buffer condition the protein tends to be in conformational equilibrium between an oligomeric and an amyloid-like fibrillar state; as suggested from CD and FTIR analysis the protein is enriched in β -sheets.

However it should be noted that TEM does not provide a full accurate representation of the self-assembled sample, firstly because the protein is in the dehydrated form on a specific substrate, and secondly because the chemicals used for fixing and staining on the substrate can potentially influence the structural morphology and assembly.

Nevertheless, we note that suckerin-19 in these conditions assembled into regular morphologies with a relatively monodisperse particle size. These results can be further validated for an accurate identification of the various conformations in tested buffers using Cryo-TEM and AFM studies.

Figure 4.11. TEM micrographs of suckerin-19 in 5% acetic acid at **(a)** 5mg/ml protein concentration; **(b)** 1 mg/ml protein concentration; in **(c)** 2 mM MES, **(d)** 40 mM MES, **(e,f)** 50 mM MES (different scale). The MES was buffered at pH 5.5 and the protein concentration was kept constant in this buffer *i.e.*, 0.2 mg/ml.

4.2.3 Design and characterization of Suckerin-19 constructs

Suckerin-19 in different buffers exhibited an oligomeric conformation, which is enriched in β -sheets. It can be hypothesized that this state is the energetically favorable conformation in the examined buffers. Apart from the partially folded conformation hypothesis, it is likely that the monomeric form of the protein is energetically not favorable. This could be due to long Gly-rich regions adopting disordered conformations leading to self-association, during which the intermolecular β -sheet interactions stabilize the structure. The inherent flexibility of certain domains and structural interactions leading to assembly into high-order conformations has strongly limited protein structure investigations. However in such cases, alternative methods including fragmentation or truncating the predicted disordered/flexible regions from

the primary structure of the proteins have been employed, to successfully determine the structure of such truncated proteins or constructs [237-239]. While the structural understanding is certainly limited in such truncated proteins, fundamental information regarding the fold and certain specific interactions that stabilize the structure remain extremely useful to infer the full-length protein structure.

Since most suckerin proteins share a high sequence similarity and are composed of repetitive modular domains, we attempted to truncate the suckerin-19 sequence into shorter constructs by varying the length of the Gly-rich fragment, as well as the overall length of the designed construct. As the Gly-rich regions are predicted to adopt a flexible conformation, thus truncating these regions should reduce structural disorder and will enhance the solubility of the protein during expression or during refolding. While designing the sequences, the molecular weight of the constructs were targeted to be in a favorable range for structural determination using NMR, in case crystallization trials are unable to give any positive results. Specific sequences from the suckerin protein family were identified which are smaller in size and could potentially be expressed in soluble form rather than IBs. Table 4.1 below lists the constructs and proteins that were cloned and expressed in *E. coli* BL21 (DE3).

Table 4-1. Constructs derived from suckerin-19 and other suckerin proteins cloned and expressed.

Name	Protein M.W.	Expression	Refolding
S4 (Construct)	5.1 kDa	No expression	N/A
S6 (Construct)	7.2 kDa	No expression	N/A
S8 (Construct)	9.3 kDa	IB	Unsuccessful
S15a (Construct)	14.8 kDa	IB	Unsuccessful
S15b (Construct)	14.8 kDa	No expression	N/A
Suckerin-19-truncated	14.5 kDa	IB	Soluble in 20 mM sodium acetate, pH 5.0 but shows disordered conformation
Sepio monomer	14.4 kDa	IB	20 mM sodium acetate, pH 5.0.
<i>D.g.</i> Suckerin-7	16.5 kDa	IB	Not attempted
<i>S.l.</i> Suckerin-1	12.09 kDa	IB	Not attempted

While few short constructs failed to express, the rest were expressed as IBs. Two suckerin constructs were used for renaturation. While one of them exhibited random coil conformation (called suckerin-19 truncated), the other (called sepio monomer) displayed a promising β -sheet enriched structure and was further characterized using NMR. The characterization of this construct is discussed in detail in Chapter 6.

4.3 Conclusion

Suckerin proteins, the building blocks of squid's SRT, displaying a unique combination of biophysical and mechanical properties that can be exploited to fabricate materials with targeted properties. To enable this, we investigated and compared the molecular organization in a material derived from native SRT (i.e., the SRT fiber). The WAXS data indicated a well-preserved β -sheet arrangement similar to that seen in native SRT, with slight molecular rearrangement. In the fiber, the β -sheets displayed a highly anisotropic distribution, while the general organization of the elongated nano-fibrils was well preserved (as noted from the SAXS patterns). Together, these data suggest a conservation of thermally stable β -sheets embedded in the amorphous network that defines the native structure assembly of suckerins.

Preliminary biophysical investigations were performed for one of the most abundant suckerin proteins in *D. gigas* SRT (suckerin-19) to obtain insight into molecular features of suckerins. Several re-naturation screens were performed to refold the denatured protein. The protein was soluble in 50 mM MES at pH 5.5 (post refolding) and was characterized using various orthogonal methods. Although the protein exhibited β -sheet secondary structure (as confirmed using CD and ATR-FTIR), which is stable at higher temperatures, oligomerization (as supported from the DLS and NMR experiments) of suckerin-19 in this buffer hampered further detailed tertiary structural characterizations. The NMR data also indicated that the protein is partially folded in this buffer, which is likely not similar to that of its native conformation. The TEM analysis suggested the existence of two conformations of the protein in MES i.e., an extended fiber-like conformation and a globular-like conformation. Further experimental structural characterization of the suckerin-19 was hampered due to aggregation. Therefore, based on the above experimental observation for suckerin-19 and basic structural information about the native SRT, computational studies were carried out (discussed in chapter 5) with a view to obtain insights into local

conformations and provide the tertiary fold adopted by suckerin-19. In addition for structural studies, various short constructs of suckerin-19 were designed to reduce the disorder in the structure arising from the long repetitive Gly-rich regions. Two small MW suckerin proteins were also expressed. Among the 8 different clones, sepia monomer was successfully re-folded from IB and was used for structural characterization (discussed in chapter 6).

5 Theoretical Structure Prediction of Suckerin-19 protein

5.1 Introduction

As discussed in detail previously, SRT is a notable exception, as it carries out its functional role by exhibiting mechanical properties that can surpass or match those of existing synthetic polymers [117]. Understanding the key nanostructure of SRT has been useful to comprehend the basis of its mechanical features which are related to its core structural organization, in particular the interactions within the building blocks that assemble to form SRT [120]. To understand the structure-property links, it is thus crucial to further develop the atomic detailed models of these protein materials to rationally design engineered proteins and materials with superior properties.

In SRT, the evolutionary conserved domains (Pro-M1-Pro-M2), organize themselves into semi-crystalline β -sheet networks embedded in the less orderly amorphous matrix. Thus it can be said that the semi-amorphous molecular structure is the source of outstanding mechanical properties exhibited by the SRT. Moreover, the existence of thermally resistant β -sheets in the SRT and the suckerin proteins have been clearly shown in our experimental efforts to stabilize the SRT [129] and the suckerin protein structure (chapter 4), however no atomic level model could be identified.

Initially systematic experiments were performed to identify the structure of suckerin-19 using two widely used traditional methods *i.e.*, X-ray crystallography and NMR which are by far the most successful approaches for structure determination of macromolecules in general. Since, biophysical characterization of the proteins by X-ray crystallography or NMR often requires highly concentrated monodisperse homogenous soluble proteins [240, 241], suckerin-19 aggregation at higher concentrations and oligomeric conformations in tested buffers turned out to be a limiting factor for its structural determination by these experimental methods. Preliminary spectroscopic measurements showed that the oligomeric conformation adopted by the protein in different solutions was stabilized by intermolecular and intramolecular β -sheet interactions [119]. However, under these conditions, the atomic structure could not be further characterized by either method. While aggregation limits

the experimental identification of the atomistic structure of proteins, it is likely that in its native state this abundant suckerin protein would be in an energetically favorable stable oligomerized conformation either with itself or with other suckerin proteins that assemble to form the SRT.

Since tertiary structure details are indispensable for the understanding of material properties, in cases, where the structure determination of proteins are impeded due to experimental limitations, computational methods for protein structure prediction can be applied [242]. A popular method adopted for theoretical protein structure modeling is homology based modeling [161, 243, 244]. The fundamental principle of this approach is based on the fact that most proteins are evolutionarily related and can be classified into different families. Proteins in the same families share functional similarities and thus three-dimensional structural folds [245-248]. Therefore, with the knowledge of the structure of single member of the protein family, the structure of all proteins that belong to that family can be in principle reliably predicted using homology modeling. However, the accuracy of this method is highly dependent on the sequence identity between the target sequence and the sequence of the available structural template. While a sequence identity of more than 40% generally leads to reliable models, identity range between 30-40% can lead to errors which can probably be resolved by further structural refinement procedures. However, sequence identity below 30% or lower sequence coverage can lead to inaccurate protein models [157, 247, 249-251].

Therefore predicting the structure of the suckerin protein based on computational methods can provide useful links between the structural features and the mechanical properties of the suckerin building blocks. Thus, in this Chapter computational efforts to predict the conformation of the modular building blocks that are evolutionarily conserved in the suckerin protein family are outlined. An attempt is made to predict a theoretical model of the full-length suckerin-19 to propose the possible conformation and the organization of the crystalline β -sheets at the atomic level.

5.2 Results and Discussion

5.2.1 Secondary Structure Prediction

Experimental secondary structure determination using CD spectroscopy and FTIR, as discussed in the previous Chapter revealed that similar to the native SRT nanostructure, the suckerin-19 protein is enriched with β -sheets [119]. While these experimental methods are significantly useful to estimate the overall secondary structure content, no residue specific information can be obtained [252].

However, there are various algorithms available for residue specific secondary structure prediction of a protein, which are based on machine learning or neural networks. Among the available methods, PSIPred [253, 254] and JPred4 [255] have been assessed as best prediction secondary structure prediction methods in the CASP experiments (see Chapter 3, Section 3.10.2). To obtain an unbiased consensus, another prediction server, namely GOR IV [256] was used (see methods section, chapter 3). While JPred itself is a consensus-based method, PSIPred makes use of knowledge-based potentials, with neural networks that learn from input evolutionary datasets. On the other hand, GOR IV is based on information theory that derives its probability parameters from empirical analysis of known tertiary structures solved by X-ray crystallography. The predictions from the three methods indicate that suckerin-19 is predominantly β -sheets, with no α -helical regions (Figure 5.1). The PSIPred server predicted that the Ala- and His-rich domains, exclusively adopt β -sheets. The other two servers predicted that certain Gly-rich motifs in the amorphous domains also have structural propensities towards β -sheet formation. A consensus prediction derived from the three methods depicted that Ala- and His-rich motifs frequently exhibit β -strand formation along with the certain GGX motifs.

Secondary structure prediction results are a valuable guide, specifically when they are in agreement with experimental data (CD or FTIR) and provide an accurate starting point for tertiary structure predictions [257]. However, to understand the fundamental fold geometry and the β -sheet arrangement, it is crucial to identify the tertiary fold. Thus, this knowledge of the local conformations acquired is used next to predict the tertiary fold of the suckerin-19 using computational modeling.

<u>MAAIFTLLAVLAISNYASAIL</u>		
PAATSVSR	TTHRTGYGY	49
PATTAVSHT	TTHHAPAALGVY	81
PGAATVSH	TTHHAPYGY	124
PAATAVSHT	TTHHAPYGYGLGGY	173
PAATAVSHT	TTHHAPLGYGLAGYGGLY	211
-GAAVSH	TTHHAPLGYGLGGYGGLY	252
PAATAVSHT	TTHHAPLGYGLGGYGGLY	274
GAAVS	-HTTHHAPLGF	319
PAAATVSH	TTHHAPYGY	343
- - -AVSHT	TTHHAPLGF	365
---GAVSH	TTHHAPYGF	417

Figure 5.1. Consensus secondary structure predictions (using JPred, PSIPred and GOR IV web-servers) **for suckerin-19**. The residues highlighted in yellow are predicted to adopt β -strand conformations, while the rest are predicted to be disordered regions.

5.2.2 Peptide Simulations

Homology modeling is a reliable method for the structure prediction of target proteins given, that a suitable structurally homologous template is identified with high sequence similarity [161, 245]. As mentioned in Chapter 2, suckerins have been characterized as a novel protein family. Even though it has similar modular architecture such as the Gly-rich motifs found in spider silk protein or the shell matrix proteins, unfortunately attempts to find structural templates with full or even partial sequence identity using web-based search engine BlastP [162] were unsuccessful. Certainly, in the absence of a homologous template, the structure could not be predicted for suckerin-19 protein or any other suckerin protein using comparative modeling.

Therefore, an alternative *de novo* prediction method (*ab initio* modeling) was applied to predict the structure of the full-length suckerin-19. In *de novo* modeling, the conformational search is conducted under the guidance of various scoring and designed force field functions [157, 158, 242, 258]. Practically, the starting point for *de novo* modeling is the use of established all-atom physics based force fields (like AMBER [184, 185, 259, 260], CHARMM [186, 261, 262], GROMOS [187] or OPLS [188, 263]) with large numbers of selected atoms types. The physical and chemical properties of each atom type are similar to the parameters calculated from quantum

mechanics theory [184, 264]. The above-mentioned force fields mainly differ in the selection of their atoms types and interaction parameters. In principle, these classical force fields can be coupled with molecular dynamics simulations to predict the structural fold of proteins. However these methods are limited by protein size, computational resources required and they have yet to be systematically assessed for their accuracy [164].

We first predicted the local conformations adopted by the repetitive modules that assemble to form the suckerin-19 structure. To do so, the suckerin-19 primary sequence was reduced into 8 different (peptides) combinations of modular units with Ala- and His-rich (approximately 12 residues, also called M1) as well as Gly-rich domains (called M2). The strictly conserved Pro residues separate the two domains such that primary modular structure of suckerin-19 can be distinguished as Pro-M1-ProM2. Table 5.1 shows the sequences of truncated suckerin-19 peptides that were simulated to predict their local conformational propensity. Such fragment-based simulations have been performed to study the structural assembly and mechanical properties of various material proteins [19, 91]. For instance, the influence of β -sheet nano-crystals on the properties of silk [48], interplay between hierarchical nature and structural organization in collagen [66] or illustrating conformational ensembles in elastin-like peptides at atomic level [265] have been studies using this approach.

Extended conformations of peptides were simulated using Generalized Born (GB=5 in *Sander*) model. We used high temperature runs to overcome potential energy barriers, induce the peptide collapse or to get a collapsed state of peptide from its initial fully extended conformation. Simulated annealing is another method that could have been applied in this case, however the method requires multiple annealing simulations to be performed for single system in order to obtain convergence. Given the large size of peptides, during annealing simulations high temperatures may completely unfold the peptides and they may not be able to reach the folded state in a short relaxation time period. Therefore we initially used short duration implicit solvent simulations at high temperature to collapse the peptides from a fully extended state and then perform folding simulations in explicit solvent. Similar approaches have been used earlier in protein/peptide folding studies [266].

The final conformation obtained after performing implicit solvent simulations was then used for subsequent production run in explicit solvent and at 298 K. The peptides were minimized, equilibrated in water to obtain an ensemble of structures using MD simulations. The final production run in explicit water was performed as it would provide an accurate description of the environment that drives the self-assembly of these modular building blocks into the structural fold adopted by the protein [31, 267-269]. The representative structures were selected using a clustering algorithm based on mutual similarity according to the RMSD for each structure. Within the largest cluster, the structure closest to the cluster center was chosen as the representative structure. The secondary structure content was calculated using *ptraj* module in AMBER.

Table 5-1. Modular units that assemble to form suckerin-19 on which MD simulations were performed.

Name	Sequence	Length	Density %* of H-Bond
BP1	PAAATSVSRTHRTGYGYGGLGGYGLHY	29	57
BP2	PGAATVSHITTHAPYGYGGLYGGYGLAHGGLYGGYGLGAYGYGY	45	74
BP3	PAATAVSHITTHAPYGYGLGGYGGYGLAHGGLYGGYGLGGYGLHY	50	76
BP4	PAATAVSHITTHAPLGYGLAGYGGLYGGLYGGYGLGGY	39	79
BP5	GAAAVSHITTHAPLGYGLGGYGGLYGGLYGGYGLGGYGLHY	42	63
BP6	PAATAVSHITTHAPLGYGLGGYGGLYGGLYGGYGLGGY	39	62
BP7	GAAVSHITTHAPLGFGLGYGGYGGYGGYGYGYGY	32	69
BP8	PAAATVSHITTHAPYGYGGLYGGYGLAHGGLYGGYGGY	39	64

The conserved Proline residues highlighted in pink separate the Ala- and His-rich M1 domain and the Gly-rich M2 domain. The residues highlighted in yellow participate in β -sheet formation as observed from MD simulations.

*Density % is the percentage of hydrogen bonds conserved for over 50% in the equilibrated MD trajectory.

The secondary structure analysis of the representative models obtained from the simulations depicted high conformational propensity towards β -structures (β -strands or β -turns). The Ala- and His-rich domains flanked by the Pro residues, fold as β -strands embedded within relatively disordered Gly-rich segments (Figure 5.2). However, no α -helical conformations were observed. Additionally, it was noted that the non- β -sheet conformations were dominated with either β -turns or disordered structures, mostly acquired by the Gly-rich domains. During simulations, it was observed that the flexible GGX motif, in the Gly-rich domain, oriented itself to initiate stable H-bond interactions with the Ala- and His-domain forming β -sheets. The *ptraj* secondary structure analysis revealed that apart from adopting β -sheets or β -turn conformation, certain regions in the amorphous domain, have propensity to form 3_{10} -helices (mostly observed in the M2 domain) (Figure 5.3). From these simulations, it can be observed that apart from residues in the M1 domain, the YGGL motif in the M2 domains can also participate in β -sheet formation.

Figure 5.2. Local β -sheet conformations adopted by suckerin-19 (BP1-BP8) peptides, predicted using MD simulations. The structures are representative conformations obtained from the largest cluster for each peptide.

Figure 5.3. Secondary structure conformations of suckerin-19 (BP1-BP8) peptides predicted during the MD simulation and calculated using *ptraj*. 3_{10} helix (black line), β -sheets (red line), β -turns (green line).

Subsequently the long Gly-rich region in the peptides was reduced to 7-residue length (Table 5.2), keeping the YGGL motif intact in the primary structure so as to validate the effect on the conformational propensity. The ensemble of conformations obtained results in similar observations as noted before, where in the flexible YGGL motif orients to form stable β -sheet with the Ala- and His-rich domain (Figure 5.4).

Table 5-2. Modular units with reduced Gly-rich domains on which MD simulations were performed.

Name	Sequence	Length	Density %* of H-Bond
SP1	PATTAVSHTTHHAPLGYGGLY	22	68
SP2	PGAATVSHTTTHHAPLGYGGLY	22	59
SP3	PAATAVSHTTHHAPLGYGGLY	22	63
SP4	PAAATVSHTTHHAPYGYGGLY	22	63

The conserved Proline residues highlighted in pink separate the Ala- and His-rich M1 domain and the Gly-rich M2 domain. The residues highlighted in yellow participate in β -sheet formation as observed from MD simulations.

*Density % is the percentage of hydrogen bonds conserved for over 50% in the equilibrated MD trajectory.

Figure 5.4. Local conformations of suckerin-19 peptides (SP1-SP4) with short Gly-rich domain (sequences shown in table 5.2). The extended peptides conform into β -sheets. The secondary structure of each peptide (SP1-SP4) during the course of the MD simulation was calculated using *ptraj*. 3_{10} -helix (black line), β -sheets (red line), β -turns (green line).

In separate MD simulations, two shorter peptides (SP1 and SP2) were designed as a single peptide (AP1) and simulated to mimic the protein primary structure assembly (Figure 5.3). The ensemble of conformations acquired depicts β -sheets formation with the Gly-rich motif being stabilized by intra-molecular H-bonding. The density of H-bonding is observed to increase (30%) with the assembly of peptides, most likely due to relatively higher ordering or formation of additional β -strands that stabilize the β -sheet arrangement. Further simulations attempted on larger peptide assemblies lead to compact structures, most likely due to hydrophobic collapse resulting from inefficient sampling. Moreover, it was computationally expensive to simulate larger assemblies reaching the full length extended protein.

Figure 5.5. Structure acquired by assembling SP1+SP2 peptide depicting β -sheet conformation. The table shows the sequence of the assembled peptide (AP1) length and the density percentage of H-bonds observed during the 100 ns MD simulation. The residues highlighted in yellow are predicted to participate in β -sheet formation during MD simulations. Density % means percentage of hydrogen bonds conserved for over 50% in the equilibrated MD trajectory.

Simulations on further assembly could not be performed due to large size of the protein. The challenge to perform computations on longer peptides requires enhanced sampling methods (like replica exchange molecular dynamics (REMD)) since such states are otherwise difficult to achieve at ambient temperatures over accessible time scales in MD simulations [270-272]. REMD is an effective method which can provide larger sampling by running independent replicas with similar (but not identical) ensembles and periodically exchanging the replicas between the ensembles [273]. It has been shown that the sampling is closer to the ergodic ideal and is generally statistically correct. Other widely used methods include “umbrella sampling” and “multi canonical sampling”. However, given the current time scale and the computational resources required to perform simulations on multiple peptides, we computed long explicit solvent MD simulations, which have been used in numerous cases to predict the folded state of the protein [274-277].

The MD simulations discussed above have been useful in efficiently predicting the local conformations of modular units that assemble to form suckerin-19. Additionally, the predictions were also in agreement with secondary structure predictions of the full length suckerin-19 obtained using secondary structure prediction web-servers (JPred, PSIPred and GOR IV), which also predict the Ala- and, His-rich and GGX motifs to adopt β -strand conformation; that gives more confidence in these models.

As seen in silk proteins, various secondary structures dictate a diverse range of mechanical behaviors [278]. For instance, the poly Ala/poly GA motifs forms crystalline β -sheets; GXX repeats preferentially adopt 3_{10} -helices; whereas the elastic β -turn like regions are composed of multiple GPGXX motif. Further, the extensibility of silk is largely due to the Gly-rich regions that connect the hydrophobic repeats [45, 278, 279]. The current MD simulations also predicted similar observations for suckerin building blocks, showing different secondary structures adopted by different modular peptide regions. MD simulations on peptides suggest that Ala- and His-rich domains have high structural propensity towards β -sheet and the flexible Gly-rich domains participate to form elastic β -turns or 3_{10} -helices. The MD simulations of peptides thus reveal distinct secondary structure features adopted by the specific building blocks of suckerin-19 protein. The Ala- and His-rich modules in combination with the specific Gly-rich motifs adopt local β -strand conformation and assemble into β -sheet domains, which are likely to provide mechanical integrity to SRT at a higher level of structural organization. In general, the Gly-rich building blocks (other than the specific YGGL motif) lead to disorganized conformations that can be considered as the soft flexible regions of the protein.

5.2.3 Full Length *Ab initio* Modeling

Successful *ab initio* modeling depends on the use of accurate energy function with which the thermodynamically stable state or the low energy states of the proteins. This state in general is associated with their native structure. *Ab initio* method aims to apply an efficient search method that allows quick identification of the low energy states through conformational searches. The energy functions used for *ab initio* modeling prediction can be broadly categorized into physics based and knowledge-based energy functions. In physics-based energy functions the interactions between the atoms are strictly defined by quantum mechanics and coulomb potential. The knowledge-based energy function involves empirical terms that are derived from statistics of solved structures deposited in the PDB and assembles fragments for structure prediction. While *pure ab initio* model building methods are limited to short target sequences (< 120 amino acids), *semi-ab initio* approaches can be applied that constructs models based by threading short structural fragments or use statistical information to spatially restrain the model building process. In past few years, CASP experiments (see methods section) have resulted in many successful *ab initio* modeling methods [148].

Among the fully automated servers benchmarked in CASP experiments, Robetta, I-Tasser and RaptorX have been used in this study. Robetta and RaptorX are based on the threading approach and the so-called composite method i.e. the I-Tasser suite has been used extensively in the last few years. The Robetta predicted models displayed α -helices connected with disordered loops, with no local β -sheets conformation of the protein. This is clearly unexpected, since detailed experimental investigations repeatedly show that suckerin proteins adopt β -sheet structural arrangement. Thus, these models were not used for further structural analyses.

Among the five different models predicted from the I-TASSER suite, only two models (referred in this study as IT1-model and IT2-model) were enriched in β -sheets, without any α -helices. Both models were assessed to be better than the other predicted models based on the high C-Score (IT1-model: -1.9 and IT2-model: -1.6). While the two models were enriched in β -sheets, they were significantly different in their structural topologies (Figure 5.6). IT1-model showed an extended conformation of suckerin-19 with stacks of β -sheets connected with shorter β -turns or disordered loops. The model illustrated the formation of an L-shaped β -helical structure with multiple stacks of β -sheets in the N-terminal region that form the stem of the letter L. This was followed by a slight bend in the structure composed of a long Gly-rich loop (aa 225-252), which can be indicated as conformationally similar to a hinge. The C-terminus conforms into a β -helix domain that forms the bottom of this L-shaped structure.

While searching through the literature, a protein known as Antigen 43 [280] was found (PDB ID: 4KH3). This is a self-associating bacterial protein that also adopts a more extended β -sheet conformation and has a similar structural topology with the predicted IT1-model. Even though the protein shares low sequence similarity with suckerin-19 protein, the self-association or oligomerization function of Antigen 43 [280] is similar to the self-association behavior and the ability to form of large oligomeric assemblies, shown experimentally for suckerin-19 in both mild acidic and aqueous buffer. The oligomeric structure of Antigen 43 is stabilised by inter-strand H-bond network between the two monomers that arrange by twisting the β -helix in a *trans* configuration. Since experimental data on suckerin-19 showed β -sheet enriched oligomerization, this model can be hypothesized as one of the conformations adopted by the protein.

The second IT2-model and the RaptorX model predicted similar globular-like structural fold. In these globular-like models the β -sheets lie between larger disordered segment or flexible loops with no helical conformations observed. In comparison to the aforementioned models, IT1-model has an extended structural fold, however all the models are composed of mainly β -sheets. It is worthy to mention, that this protein can adopt multiple β -sheet conformations like in solution (monomeric or oligomeric), in processed form (cross-linked biofilms and gels or extended fibers) or its native form in SRT wherein the conformation of the protein might vary when in complex with other suckerin during assembly of SRT. Therefore the two different conformations (the extended conformation in IT1-model and globular conformation in IT2-model and RaptorX model) predicted could reflect few of the various conformations of this specific protein. Although these two predicted models are to be further refined, tested for stability using MD simulations, which can be useful to explain some of the structure-property relations; tertiary structure determination using experimental methods (X-ray or NMR) is necessary to validate these models and to understand the structure-property relationships.

The prediction of similar structural folds by two different methods (I-Tasser and RaptorX server) increased the reliability of these 3D-models, however the IT2-model was relatively enriched with β -sheets. Therefore further analysis was only performed on the IT2-model. In IT2-model seven β -sheet domains were connected by longer disordered loops. This conformation is similar to the structure proposed in the native SRT structure, wherein the β -sheets are randomly oriented between the amorphous domains [120]. Further comparisons of the secondary structure components from this model and simulations performed on shorter peptide fragments (previous section) reveal extensive similarities notably in the strong propensity to adopt β -structures. Even in the full-length model the Ala- and His-rich domains exhibit higher propensity towards β -strands and arrange to form β -sheets along with the Gly-rich motifs. MD simulations were carried out for the model refinement, to allow evolution of secondary structure elements and the tertiary fold into an energetically favorable conformation that can be correlated with the native structure.

Figure 5.6. Tertiary structure of suckerin-19 predicted using I-TASSER. (a) IT1-model depicts an extended 3D conformation, **(b)** IT2-model, which conforms into a more globular-fold.

5.2.4 Model refinement with MD

MD simulations coupled with classical force fields were used to determine the protein fold, however from structural prediction perspective, the results were not successful. Despite many remarkable efforts attempted for structure prediction of typical sized protein (100-300 residues), all atom physics based MD simulations are not routinely used for structure prediction. However, the physics-based MD simulations can be useful for structure refinement, provided a low-resolution protein model is initially available. The basic idea of using MD based structure refinement is to draw the predicted low resolution model to a more native-like conformation by refining the global backbone packing and local side chain conformations [164]. The expected conformational changes are relatively small and the simulation time is less. Therefore we further refine the predicted models (IT1-model and IT2-model) and analyze the stability of the β -sheets using explicit solvent MD simulations. The starting conformation of IT1-model was enriched in β -sheets. However, during the trajectory (of 1 μ s), the model evolved towards a more dynamic structure with some β -sheet predicted regions transitioning into flexible loops (Figure 5.7). It is also noted that the N-terminal domain depicted dynamic motions due to the large flexible Gly-rich

regions (139-173 aa). As noted from the primary structure, this stretch corresponds to the relatively longer Gly-rich module (in comparison to other Gly-rich domains). In the peptide simulations the Ala- and His-rich domains are predicted to also involve the Gly-rich domain in β -sheet formation, which is also observed in the full-length predicted IT1-model. However, this specific region (139-173 aa) conforms into a more flexible loop or β -turns due to the remoteness of Ala, His-rich domain. The observations from the ensemble of structures suggest that this region could form the hinge region between the flexible N-terminal domain and a more ordered C-terminal region. In comparison to what was initially observed in the structure, the C-terminal probably forms the stem of this L-shaped structure, whereas the N-terminal forms the base.

Figure 5.7. Structural evolution of IT1-model observed during MD simulation. (Left panel) Conformation of suckerin-19 extended model over 1 μ s simulation. The structure evolves into a more dynamic conformation with less ordered β -sheets. (Right panel) Distribution of the inter-strand distances between the N- and C-terminal β -sheets as observed during MD simulation.

The inter-strand β -sheet distances (excluding the flexible hinge region) were calculated for the IT2-model MD trajectory. The values obtained for the ensemble of structures revealed an average of 0.5 nm (Figure 5.7). This value is similar to the inter-strand distance provided from the experimental WAXS data of native SRT [129]. Integration of intensities across the entire azimuth indicated nano-scale β -sheet domains with an inter-strand distances (0.45 nm) similar to the values found in silk-like β -sheet crystals (0.43 nm) and amyloids (0.46 nm) [129].

MD simulations for IT2-model provided an ensemble of conformations with anisotropic β -sheet arrangement (Figure 5.9), which is consistent with the experimental findings [129]. The Ala- and His-rich domains formed β -strands, which further orient the flexible Gly-rich motifs to form a β -sheet domain. Moreover, the large disordered regions that connect the 7 β -sheet domains provide a higher degree of freedom to these β -sheet domains in the conformational space. These β -sheet domains are stabilized by inter-strand H-bonds, observed to be stable for over 50% of the equilibrated MD trajectory. The spatial organization of Tyr residues and their influence on the structure as well as the material properties of suckerin-19 derived materials was also predicted. The Tyr residues located within the Gly-rich motifs have been successfully exploited to fabricate di-Tyr cross-linked suckerin-19 films for material applications [130]. Similar to silk and native SRT, the Young's modulus of these cross-linked protein based films display variations in their respective β -sheet content. As the cross-linking density increases, the Young's modulus decreases exponentially in hydrated conditions (Figure 5.8). It is expected di-Tyr cross-links would lead to overall structural stabilization of the protein assembled. However, experimental data depicted that increase in di-Tyr cross-linking resulted in decrease in β -sheet content [130] (as obtained from FTIR analysis, shown in Figure 5.8a).

Figure 5.8. Characterization of secondary structure by FTIR and elastic modulus in suckerin-19 films (a) Amide I bands of suckerin-19 films depicting transition from β -sheet to random coil with increase in the concentration of APS (Ammonium persulfate; cross-linker) (b) E (Young's modulus) versus β -sheet content (wet conditions) of cross-linked suckerin-19, illustrating the correlation between the modulus and the β -sheet content. Adapted from Ref. [130].

To examine these observations, the ensemble of conformations generated during MD simulation of the IT2-model was analyzed. It is observed that 52% of the Tyr residues are spatially arranged in the flexible dynamic loops and exposed to the solvent (calculated using solvent accessible surface area of Tyr residues). Of the remaining Tyr residues, around 50% residues participate in β -sheet formation (Figure 5.9).

Figure 5.9. Characterization of Tyrosine residues in IT2-model MD simulations. (a) Solvent accessible surface area (SASA) of all Tyr residues calculated from the ensemble of conformations obtained from MD simulations of the IT2-model. (b) Cartoon representation of IT2-model depicting the spatial arrangement of Tyr residues: (b) exposed Tyr represented as cyan sticks; (c) buried Tyr residues represented as green sticks; (d) buried Tyr residues that are part of β -sheets, represented as yellow sticks.

Therefore based on the previous experimental data and the observation from MD, it can be hypothesized that in weakly cross-linked films (which exhibit higher β -sheet content) the Tyr residues located in the flexible disordered regions are labile for any chemical modification and therefore participate in cross-linking. However, as the percentage of cross-linking increases, these Tyr residues become saturated. Instead the Tyr residues located in the β -sheet are recruited and rearranged structurally, to form

additional di-Tyr crosslinks. This structural re-arrangement frustrates β -sheet formation, leading to a more disordered secondary structure content. Thus, based on the IT2-model, the structural phenomenon observed for the suckerin-19 derived cross-linked films was explained at the atomic level.

5.3 Conclusion

Using computational modeling and simulation we attempted to predict the 3-dimensional structure of Suckerin-19. We first predicted the local β -sheet conformations adopted by the specific Ala- and His-rich domains as well as the GGX motifs. We next predicted the full-length model of suckerin-19 using *ab initio* modeling approaches. We found that the model derived using the I-Tasser server emerges as the most robust and consistent with the available experimental data. The obtained models were enriched in β -sheet domains that are embedded within the disordered flexible Gly-rich regions, and this arrangement is in agreement with the previous experimental data. An alternative conformation of this protein is also predicted by I-Tasser that has more ordered β -sheets than in comparison to the globular-like model predicted. The quality of these models remained moderate, as depicted from the C-score values indicating poor prediction which is possibly due to the large size of the protein (398 amino acids). MD simulations were used to refine the models. The predicted dynamics of the Tyr residues provides a hypothesis to explain how disorder increases upon increasing di-Tyr crosslinking leading to frustration of β -sheet formation.

6 Biophysical Characterization of Sepio Monomer

6.1 Introduction

The sepio monomer construct shares homology as well as block copolymer like structure with *S.l* suckerin-1 and *D. gigas* suckerin-19 protein (sequence identity 81% and 72% respectively) [119] as explained in Section 3.2.2. Further sequence based comparison of the sepio monomer construct showed that it shares 40-70% sequence similarity with all other proteins in the SRT family. The primary sequence of sepio monomer is also composed of repetitive M1: Ala-, His- rich and M2: Gly-, Tyr-rich modules (arranged as Pro-M1-Pro-M2), which is the distinctive feature of SRT proteins (Figure 6.1). However, sepio monomer (14.6 kDa, Figure 6.1) is three times shorter than *D.g.* suckerin-19 (39 kDa) and has a similar MW as *S.l.* suckerin-1 (12 kDa). The primary sequence is composed of maximum number of Gly (24.3%) followed by Tyr (14.4%) and Leu (11.7%) residues. Asp and Glu exhibit minimum frequency (0.9%). Unlike suckerin-19, this construct contains one negatively charged residue *i.e.*, Asn and two positively charged Arg and two Trp. The amino acid composition and the MW of sepio monomer yields structural expectations, wherein the hydrophobic residues impart stability and solubility, which should enable structural characterization of this construct.

To compare the mechanical properties of the sepio construct and the previously tested suckerin-19, a new construct was engineered from the sepio monomer, known as the sepio trimer (Section 3.2.2.). Repeating the sequence of sepio monomer thrice resulted in a 37 kDa construct which has a comparable molecular design and weight to suckerin-19 protein (Figure 6.1). The sepio trimer is expected to exhibit oligomerization at high concentrations as observed in suckerin-19, which can be useful for engineering materials and further compare their properties. A separate study performed (at MEL A*Star laboratory by Dr. Shawn Hoon; unpublished work) showed that both sepio constructs tend to form hydrogels at high concentration (100 mg/ml) when sonicated overnight in water and 5% acetic acid. It was also found that the sepio trimer formed β -sheet-rich structures. Given that no chemicals were used for gel formation, these gels are considered to be “physical” gels whose cross-linking is based on weak hydrogen bond interactions rather than “chemical” gels that are covalently

cross-linked. The gelation is hypothesized to be thermally-activated during sonication and/or triggered by ultrasonic waves, both of which frequently initiates β -sheet structural re-arrangements in proteins [281]. Rheology measurements further demonstrated that gels formed in water exhibit a higher elastic modulus (1.5 times) than the gels made in mild acidic conditions (Figure 6.1). It has also been shown that the seprio trimer can be used to cast homogeneous films with controlled thickness and crystalline content through methanol-based water annealing treatment [282].

Figure 6.1. Primary structure features and material properties of seprio constructs. (a) Schematic representation of modular repeats distribution in seprio monomer and trimer in comparison to *S.l.suckerin-1* and *D.g. suckerin-19* (the Ala- and His-rich regions called M1 domains are shown in red and Gly-rich regions called M2 domains are shown in blue). The molecular weight is indicated on the right, (b) Schematic representation of seprio monomer and its trimer engineered by repeating the monomer sequence thrice (c) Primary sequence of seprio monomer, (d) dynamic measurements of storage modulus (G') and the loss modulus (G'') as a function of strain for seprio trimer gels. $G''/G' > 1$ in the low strain regime indicates a gel-like behavior.

These preliminary data forms the basis to develop materials using the seprio constructs. Furthermore, it is appealing to understand the structure-property relationships for suckerin proteins in general, given their highly similar modular architecture that can be considered representative of the entire suckerin family. Towards this goal, the seprio

monomer construct was cloned for recombinant expression. The solution behavior of the construct in different buffers was tested and the conformational preferences were characterized. Spectroscopic measurements by NMR spectroscopy were carried out in order to experimentally determine the structural fold adopted by seprio monomer.

6.2 Results and Discussion

6.2.1 Cloning, Expression and Purification of Seprio monomer

The codon-optimized gene encoding seprio monomer was purchased from DNA 2.0 (Newark, California). Through the primed extension the PCR product containing this gene was generated and digested with NheI and XhoI restriction enzymes. The amplified cDNA fragment of the monomer gene was purified and subcloned into NheI/XhoI digested pET28a expression vector and N- (with TEV cleavage site) and C-terminal His₆ tags were added to the protein. The transformants were screened by test digestion and DNA sequencing, which confirmed correct and unique DNA products. *E. coli* BL21 (DE3) cells were cultured to expression tagged seprio monomer.

The recombinant double his-tagged seprio monomer was solely found in the cell pellet of *E. coli* suggesting that the protein is predominantly assembled in the IB during cell expression. The amount of IB for the target protein was observed to be higher when induced at 37 °C than at 16 °C for 20 hours under the control of a T7 promoter by 0.5 mM IPTG. IBs appear due to over expression of proteins causing protein misfolding and aggregation [283]. Soluble proteins are always preferred for downstream structural characterization, since they are considered to sustain native-like structures in the soluble form. However, formation of inclusion bodies can yield relatively pure protein of interest from the bacterial cytoplasmic proteins [284].

After sonication and subsequent wash steps the over-expressed recombinant seprio monomer was subsequently extracted in 5% acetic acid from the IBs with a reduced amount of low MW contaminant proteins, as depicted by SDS-PAGE (Figure 6.1). The MW of the expressed double his-tagged protein was determined to be 14.83 kDa by mass spectrometry (expected theoretical MW is 14.46 kDa). This corresponds to the

monomeric state of the protein, whereas the SDS-PAGE shows an apparent molecular mass of 18 kDa (Figure 6.2). The observed discrepancy between the two methods is most likely due to the extended conformation and charge of the protein. Since the His residues are protonated at pH used for electrophoresis, the protein binds less efficiently to SDS and hence displays a lower mobility on the polyacrylamide gel. Nevertheless, the MWs obtained by MALDI-ToF are more accurate and match the predicted value.

Figure 6.2. Recombinant expression of sepio monomer. (a) SDS-PAGE gel of the expressed protein. 1, molecular weight marker; 2, protein expressed in IB extracted in 5% acetic acid and 3, protein refolded in 20 mM sodium acetate, pH 5.0. (b) MALDI-ToF spectrum of protein recorded in presence of 5% acetic acid, using sinapic acid as matrix.

6.2.2 Renaturation of Sepio monomer

Recovery of the recombinant protein from the IB was accomplished by unfolding with chaotropic agents. To obtain the native conformation, it is thus essential to carry out dialysis in appropriate refolding buffer. Hence, multiple conditions were screened to obtain the renaturated sepio monomer. Since suckerin-19 precipitated on addition of salts during solubility screening, sepio monomer was initially tested in buffers without any additives (see Chapter 3, Section 3.5). Among the tested conditions, the denaturant was gradually dialyzed and the protein was found to be soluble in 20 mM sodium acetate buffered at pH 5.0. However the highest solubility concentration obtained was ~1.8 mg/ml with 90% protein recovery after refolding. As a final purification step, size exclusion chromatography was used to obtain the highly purified monomeric protein under native conditions. Based on the chromatogram, single peak containing the

refolded protein was eluted between 60-70 mL with 95% purity estimated from SDS-PAGE (Figure 6.3). The pooled fractions were concentrated to 1.5 mg/ml (= 0.1 mM protein concentration) and subsequently used for further structural characterization.

Figure 6.3. Size exclusion chromatogram of seprio monomer. The inset shows SDS-PAGE analysis of the peak fractions 60-69.

6.2.3 Preliminary Spectroscopic Characterization

To explore conditions that allow structural analysis of seprio monomer in an isotropic, homogenous microenvironment, the solubility behavior and spectroscopic features were examined. From the aforementioned solubility analysis it is noted that the seprio monomer is soluble in three buffers: (i) mildly acidic, (ii) water (similar to suckerin-19), and (iii) 20 mM sodium acetate in which the protein is assumed to have a native-like conformation. In order to investigate the influence of the aqueous-based microenvironment on the morphology and structure of seprio monomer various spectroscopic measurements were performed.

6.2.3.1 Dynamic Light Scattering

At 0.1 mM protein concentration, DLS measurements (Figure 6.4) indicated that in water and mild acidic conditions seprio monomer tends to self-assemble, forming oligomers with mean hydrodynamic diameters of D_H of about ~30 nm and ~55 nm, respectively. But in the case of 20 mM sodium acetate the protein apparently exists in

a monomeric state with a D_H distribution around 6-10 nm. The polydispersity indices measured for the sample in each buffer condition were similar (~ 0.30).

Figure 6.4. Spectroscopic characterization of sepio monomer (a) DLS particle size distribution **(b)** Far-UV CD spectrum in 5% acetic acid (red), water (blue) and 20 mM sodium acetate, pH 5.0 (green).

6.2.3.2 CD Spectroscopy

Far-UV CD spectroscopy (190-250 nm) is frequently applied to experimentally identify the secondary structure composition of the sample. The CD spectrum of solubilized sepio monomer concentrated to 0.1 mM in 5% acetic acid and 20 mM sodium acetate buffer (pH 5.0) features a characteristic random coil protein conformation, with minima at 198 nm (Figure 6.4). However in the case of aqueous buffer the minima was shifted towards 208 nm (Figure 6.4). Previous structural analysis showed that SRT is rich in β -sheets. Similar results have also been reported for recombinant suckerin-19 protein that tends to form intra-molecular and intermolecular β -sheet networks.

Numerous studies have shown that reliable estimation of secondary structure content from CD spectra cannot be made for noncanonical β -strand containing proteins [285]. Variations in length, direction, extent, degree of distortion and twist of β -sheets as well as the orientations of the neighboring β -strands can lead to structural diversity in β -structures [286, 287]. These variations exhibit diverse spectral CD features. Moreover β -rich-structures tend to assemble in a range of sizes and form different morphologies like oligomers, fibrils or amorphous aggregates, which exhibit specific physiological

effects based on the micro-environmental conditions [288, 289]. Consequently, due to the spectral variety and low spectral amplitudes, secondary structure estimation for proteins enriched in β -sheet is considered difficult and biased [285].

So far, structural predictions of such β -sheet enriched proteins has been a contentious issue due to the lack of standardized as well as calibrated CD reference spectra. Thus, for quantitative estimation from the CD data, the recently developed BeStSel [290] algorithm was applied. In comparison to other fitting methods, it provides an improved prediction of secondary structure, specifically for the β -structure-rich proteins by taking into account the twisting angles between the β -strands. Based on the BeStSel fitting, we estimated that the seprio monomer exhibits up to 45% β -sheet with low β -turns and no α -helical content (Table 6.1). The BeStSel fitting were in the acceptable range with RMSD = 0.09 in sodium acetate and 0.1 in water and 5% acetic acid.

Table 6-1. Estimated secondary structure of seprio monomer from BeStSel fitting of experimental CD spectrum.

Estimated Secondary Structure	20 mM Sodium Acetate	Water	5% Acetic Acid
β-sheets	44.1%	48.2%	51.5%
β-turns	13.5%	13.6%	15.2%
Others	42.3%	36.1%	33.2%

6.2.3.3 FTIR spectroscopy

To further verify re-folding of seprio monomer FTIR was used as a complementary tool to examine the secondary structure. FTIR is a useful tool to estimate the secondary structure of proteins, and the amide I band has been widely used to quantify the secondary structural composition of proteins and polypeptides. Fourier transformed plot of intensity vs. wavelength indicated a maximum peak in amide-I region at 1628 cm^{-1} , which is characteristic of β -sheet. Semi-quantitative analysis of the protein secondary structure in obtained by standard peak deconvolution of amide I region. We found that seprio monomer in 20 mM sodium acetate (pH 5.0) is abundant in β -sheet (42.5%) and β -turns (34%) with low amount of disordered regions (Figure 6.5), which is consistent with CD data.

Figure 6.5. ATR-FTIR analysis of seprio monomer. Deconvoluted amide I spectra in (a) 20 mM sodium acetate, pH 5.0; (b) in water; (c) in 5% acetic acid. The positions of the fitting peaks were defined by the 2nd derivative of the amide I spectra and the width of each peak was fixed.

In mild acidic conditions, the FTIR deconvolution data depicts that the structural conformation acquired by seprio monomer is the most abundant in β -structures (β -sheet composition: 61.2%; β -turns: 22.5%) with relatively low disorder (Figure 6.5). In aqueous buffer it exhibits 56.5% β -sheet content and 25.7% β -turns (Figure 6.5). No helical conformation was estimated in any of the buffer conditions.

6.2.3.4 ^1H -NMR spectroscopy

To further evaluate the suitable conditions for characterization of the tertiary fold of seprio monomer in homogenous and isotropic environments we tested the sample by conventional ^1H -NMR. In 5% acetic acid a broad chemical shift, depicting aggregation of the protein was observed (Figure 6.6, bottom panel).

Figure 6.6. 1D NMR spectrum (amide region) of seprio monomer in 20 mM sodium acetate, pH 5.0 (green), water (blue) and 5% acetic acid (red). Full length spectrum is shown in Appendix B, figure B1.

Although the resonances were relatively better dispersed in water, in comparison to 5% acetic acid (Figure 6.6, middle panel), the most promising measurements were obtained in 20 mM sodium acetate at pH 5.0, where the spectra showed homogenous protein preparations, with chemical shift dispersion similar to that seen for globular proteins (Figure 6.6, top panel). In these conditions, three chemical shift peaks at 10 ppm were observed (Figure 6.6) which represent resonances from the indole NH group of two Trp residues. Based on the primary structure of seprio monomer, these Trp residues are present in the Gly-rich amorphous domains. The low chemical shift dispersion of these peaks indicates that the corresponding Trp residues (*i.e.*, the indole rings) could be exposed to solvent.

Based on the overall agreement of preliminary spectroscopic measurements using various methods (1D NMR spectrum, quantitative data from FTIR deconvolution and the DLS measurements), a correlation can be derived between the size of the protein and the β -sheet content. Larger the size of the protein, higher the β -sheet content, and lower the disorder. This suggests that the protein is stabilized by intra-molecular β -sheet interactions in 20 mM sodium acetate (pH 5.0). Owing to its monomeric size

detected by DLS as well as its dispersed chemical shift resonances obtained by ^1H -NMR, it can be hypothesized that the sepio monomer in 20 mM sodium acetate exhibits a well-folded conformation. On the other hand, in aqueous buffer and in mild acidic conditions, the combination of CD, FTIR and DLS measurements suggest that inter-molecular β -sheet interactions stabilize the protein into larger oligomeric particles, similar to suckerin-19 [119]

Together these data indicate that the structural feature of sepio monomer in sodium acetate is in agreement with the characteristic enriched β -sheet conformation of the suckerin proteins. This preliminary data encouraged us to further investigate the tertiary fold in this buffer condition. The following sections discuss the details of crystal screening attempts and using more advanced multi-dimensional NMR spectroscopy in order to obtain the detailed tertiary structure of the folded protein.

6.2.4 Crystal Screening

For crystal screening of sepio monomer in 20 mM sodium acetate, three initial broad range commercially available crystallization screens kit were used namely Morpheus® HT-96 (Molecular Dimensions), Crystal Screen #1 & #2 and Index HT (Hampton Research). The screens were checked periodically for 4 weeks, however no precipitation or crystal formation was observed. For crystal growth, it is crucial to have protein in high purity and in a homogenous solution form. The purity and homogeneity of the protein was assessed previously and was analyzed to be favorable to proceed for further structural determination. It is known that protein crystallization involves two stages namely nucleation and crystal growth. Nucleation typically requires high protein concentration in order to achieve super-saturation (at least 5mg/mL). Unfortunately, for sepio monomer the maximum soluble concentration of the refolded protein was 1.5 mg/mL, at which the crystal screens were set up. While protein concentration plays an important role during crystallization, the optimum being 10-15 mg/ml (optimum for moderately sized proteins between 25-50 kDa); proteins have been crystallized at concentrations as low as 0.75 mg/ml and as high as 300 mg/ml [291]. The general trend being, low protein concentration for high molecular weight proteins and high protein concentration for low molecular weight proteins. Therefore for this specific protein the various factors that could have hindered protein

crystallization propensity could be either the concentration of protein, or the presence of large flexible/amorphous domains in the acquired conformation of the protein due to significantly higher number of Gly residues.

6.2.5 NMR Spectroscopy

We attempted to identify the structural fold adopted by sepia monomer in 20 mM sodium acetate at pH 5.0 using multi-dimensional NMR spectroscopy. Since the MW (14.8 kDa) and the maximum soluble concentration obtained for the refolded protein (0.1 mM) are in the favorable range for detection using NMR, this method of structural characterization was successfully applied.

6.2.5.1 Backbone and side-chain NMR assignment

The ^{15}N -HSQC spectrum, often considered as the fingerprint of the protein fold, showed well-resolved/dispersed peaks that can be correlated with the folded conformation of sepia monomer in this buffer (Figure 6.8). The ^{15}N -HSQC spectra recorded on the 1st and the 7th day, showed no change in the spectra or appearance of any new signals. This ensured stability of the sample, required for long data acquisition period, for NMR structure determination. All the NMR spectra were therefore recorded on ^{15}N - ^{13}C uniformly labeled sepia monomer at 298 K. The calibrated and processed ^{15}N -HSQC spectrum was loaded onto CARA program and the peaks were picked in the corresponding amide region. After peak picking, the peaks were analyzed and the noise peaks were deleted. These peaks (selected) were then converted into a peaklist mentioning the appropriate chemical shifts. Similarly, peaks were picked for all of the 3D backbone NMR spectra (HNCA, HNCACB and CACBCONH). For peak picking of 3D spectra, tolerance was set to high to avoid picking for noise peaks. ^{15}N -HSQC was used as root spectrum to aid in the picking of peaks corresponding to the amide regions. Appropriate peak lists were subsequently generated. In the same way, peak list for 3D ^{15}N -NOESY spectra was also created.

By following the stepwise protocol published for Cyana-FLYA, spectra were assigned and NMR structures were generated. NMR structure refinement was also accomplished by manually checking the assigned peaks. For refinement, the assigned backbone peaklists was loaded by FLYA onto appropriate spectra on SPARKY and the assignments were verified. The overlapping peaks or the peaks with multiple

6.2.5.2 NMR structure calculation and validation

NOESY – Nuclear Overhauser Effect spectrum is considered as one of the facile spectra to derive structural information of the protein. The proton spins are detected through space in this spectrum unlike through bonds in all of the other spectra. Hence NOEs are detected for residues that are in close contact with one another including neighboring residues. The NOEs thus detected can be categorized into sequential, medium and long range NOEs. Medium range NOEs of $i-i+2$ and $i-i+3,4$ are used to identify structural conformation of β -sheet and α -helix respectively. Long range NOEs play an important role in deciphering tertiary structure of the protein and parallel/antiparallel β -turns or hairpins. 3D ^{15}N -NOESY spectra of Sepio monomer was analyzed with automated structure calculation by CYANA. The peaks were picked and the peaklist with appropriate header was fed in CYANA calculations along with corrected peaklists of backbone spectra of Sepio. The resulting assigned NOESY peaklists were then analyzed with ANNEAL macro of CYANA and the long range NOEs were inspected for structural propensity of the protein.

6.2.5.3 Structure of Sepio monomer

Automated NOE assignment and structure calculation by CYANA resulted in the detection of several long range NOEs. For example, as shown in table 6-2 and figure 6.8, there were considerable number of backbone mediated NOEs between residues Leu35, 36 Gly36, Tyr37 and Tyr60, Gly61, Gly62 respectively, indicative of one β -hairpin conformation. Similar NOEs were also detected between residues Leu59, Tyr60 and Gly87, Tyr88, Gly89.

Table 6-2. List of backbone and side-chain NOEs

Backbone NOEs		Side-chain NOEs	
37 TYR H	61 GLY H	35 LEU QB2	61 GLY H
36 GLY H	61 GLY H	35 LEU QB2	60 TYR H
36 GLY H	62 GLY QA2	36 GLY H	60 TYR QB2
36 GLY H	62 GLY H		
72 SER HA	77 GLY H	59 LEU QB2	88 TYR H
73 THR HA	77 GLY H	60 TYR QB2	87 GLY H
60 TYR H	89 GLY H	59 LEU QB2	87 GLY H
60 TYR H	87 GLY H	60 TYR QB2	89 GLY H
59 LEU H	86 LEU H		
59 LEU H	87 GLY H		

Figure 6.8. NOE connectives of sepia monomer in 20 mM sodium acetate pH 5.0
 (a) Bar diagram representing NOE summary, (b) Bar diagram representing medium and long range NOE connectivity.

The structure characteristics validated using Ramachandran plot suggest that 60% residues are in the allowed region and 40% were in the additional and generously allowed region (Table 6-3). None of the residues were found in the disallowed region.

Table 6-3. Structural statistics of sepio monomer in 20 mM sodium acetate pH 5.0.

Distance Restraints		
Short-range	($ i-j \leq 1$)	670
Medium-range	($1 < i-j < 5$)	63
Long-range	($ i-j \geq 5$)	177
Average RMSD		(Å)
Backbone	(N, C $_{\alpha}$, C)	14.36 $\pm$ 2.82
Heavy atoms		14.79 $\pm$ 2.67
Ramachandran Plot* Statistics		Residue (%)
Residues in most favored regions [A,B,L]		47 (60.3)
Residues in additional allowed regions [a,b,l,p]		24 (30.8)
Residues in generously allowed regions [$\sim$ a, $\sim$ b, $\sim$ l, $\sim$ p]		7 (9.0)
Residues in disallowed regions [xx]		0 (0.0)
Number of non-glycine and non-proline residues		78 (100.0)
Number of end-residues (excl. Gly and Pro)		2
Number of Glycine residues (shown as triangles)		27
Number of Proline residues		7
Total number of residues		114

* Ramachandran plot is shown in Figure 6.9

Figure 6.9. Ramachandran plot of sepio monomer. The plot statistics are as evaluated with the PROCHECK program for protein structure validation.

The NMR calculated tertiary structure of sepio monomer in 20 mM sodium acetate (pH 5.0) is shown in Figure 6.9. The average backbone RMSD of ten lowest energy structures calculated was found to be 14.36 Å and the backbone RMSD of strands in core region were found to be in the range of 1.6-2.7 Å, as shown in Figure 6.9. The core region of the protein is composed of four β-strands formed by residues between

35-45 (strand I), 50-57 (strand II), 65-74 (strand III) and 78-85 (strand IV); also shown in Figure 6.11. It is noted that residues from both M1 and M2 domains participate in β -strand formation in the structure (also observed in MD simulations) and on an average the length of each strand is 9-10 residues, which is in agreement with the SAXS and WAXS data acquired for native SRT (discussed in section 2.3.5). The terminal ends are observed to adopt flexible conformations and are speculated to conform into β -strands during self-assembly of SRT or oligomerization due to inter-molecular β -sheet interactions. This is also observed in case of viruses wherein the flexible arms in the terminal ends that extend away from the core surface, loop in and adopt ordered conformations depending on their environment in the nascent assembly [292].

Figure 6.10. Tertiary structure of sepia monomer in 20 mM sodium acetate pH 5.0, in cartoon representation determined using NMR. The four β -strands marked are shown in yellow and the dynamic loops are shown in green.*The average backbone RMSD of the strands has been calculated in MolMol [293].

6.2.5.4 Hydrogen/deuterium exchange

Hydrogen/deuterium exchange (H/D-Ex) is a rapid method to probe the protein structural fold and dynamics [294]. Briefly, the method is based on the strong influence of hydrogen bonds on amide proton exchange rates. NMR spectroscopy is often used to monitor residue specific H/D-Ex rates for protein owing to their sensitivity and power to provide site-specific information [295-297]. For dynamic proteins like suckerins, H/D-Ex can be coupled with solution state NMR as a critical indicator to characterize the residues that form the stabilizing core in the adopted tertiary fold and those that are exposed to the solvent. The amide protons involved in the core hydrogen bond interactions are solvent protected and exchange more slowly than the residues on the protein surface or in the dynamic regions.

The exchange of amide protons was monitored by recording series of ^{15}N -HSQC experiments in deuterated 20 mM sodium acetate, pH 5.0 for a period of 120 mins. Figure 6.9 summarizes the spectra recorded for a series of time steps. The acquired spectra depicted a rapid exchange of residues with intensity loss in the majority of peaks within 30 mins of acquisition (Figure 6.9a and b). Subsequently signals from the solvent protected peaks or those involved in core interactions remain (without much loss in signal strength) for up to 120 mins.

From this data it can be inferred that L4, Y37, Y39, A17, A41, A42, T49, H51, L59, S75, H76, H79, W84, L86, Y88, A94 and Y110 are solvent protected residues in the fold acquired by sepia monomer in this buffer (Figure 6.9e). While certain chemical shifts for residues such as A42, Y37, Y39, A41, V74, H76, Y88 were still visible for up to 60-90 mins (Figure 6.9c and d), a significant decrease in intensity for these residue specific signals was observed after 120 mins.

Figure 6.11. H/D Ex of sepia monomer in deuterated 20 mM sodium acetate, pH 5.0 recorded at 298 K and various time steps: **(a)** 0 minutes, **(b)** 30 minutes, **(c)** 60 minutes, **(d)** 90 minutes, **(e)** 120 minutes. Region representing solvent protected residues are presented here.

Correlating this data with the tertiary structure calculated using NMR, it is noted that these resonances form the core of the structure and are involved in the β -strand formation. Apart from these residues since the overall structure is dynamic rapid exchange of other residues can be observed (*i.e.* within 30 mins). Moreover, the residues from both the modular domains (Ala- and His- rich as well as the Gly-rich domains) are involved in weak hydrogen-bond interactions that disappear due to the dynamic nature of the protein and hence solvent exchange within 90 mins of acquisition. However, the core stabilizing interactions held by relatively stronger hydrogen-bonds are observed to be stable up to 120 mins. It is likely that the overall dynamic structure of this protein affects the chemical microenvironment of the molecule and therefore less number of solvent protected peaks are observed in the H/D-Ex experiment.

6.2.5.5 Secondary Chemical Shifts

The chemical shifts are sensitive to the local physiochemical environment and are influenced by the secondary structure in both folded and disordered conformation of the given protein [298, 299]. Information from these chemical shifts is often exploited to identify the type of secondary structure in the protein. To interpret the secondary structure adopted by sepio monomer we used observed backbone ^{13}C chemical shifts (δ_{Exp}) since they are the most sensitive to the presence of secondary structure [226]. $^{13}\text{C}\alpha$ secondary chemical shifts (derived as $\delta_{\text{SCS}} = \delta_{\text{Exp}} - \delta_{\text{Random coil}}$) that are negative can be interpreted as β -sheets, while $^{13}\text{C}\beta$ follows inverse dependence [298, 300]. The $^{13}\text{C}\alpha$ and $^{13}\text{C}\beta$ secondary chemical shifts plotted for each residue in sepio monomer clearly indicates β -strand formation for the four selected regions with clear separation between the negative $^{13}\text{C}\alpha$ shifts and the positive $^{13}\text{C}\beta$ shifts (Figure 6.10). The data is also in agreement with the H/D-Ex experiment as the resonances from these respective regions were conserved up to 90-120 mins. The data also validates the MD results, wherein apart from Ala- and His-rich motifs, the Gly-rich motifs were also observed to conform into β -strands.

Figure 6.12. Secondary chemical shifts (SCS) plot for residues in sepio monomer derived from the $^{13}\text{C}\alpha$ and $^{13}\text{C}\beta$ chemical shifts. The black bars represent δ_{SCS} derived from $^{13}\text{C}\alpha$ chemical shifts and red bars indicate δ_{SCS} derived from $^{13}\text{C}\beta$ chemical shifts. Each graph represents the peptide domains that form β -strands in the NMR structure calculated.

Based on these data and the previous experimental and computational work on suckerin-19, it can be hypothesized that suckerins in monomeric form adopt dynamic conformations with low β -sheet content and weak hydrogen bonds interactions. However, we speculate that on self-assembly or interaction with other suckerin proteins (with similar sequence design), β -sheets formation quickly ensues allowing weak interactions to transform into highly stable hydrogen bonds that influence the overall structural assembly, and hence the observed mechanical properties.

6.3 Conclusion

Sepio monomer is similar to suckerin proteins in its modular design with sequence homology to *D.g.* suckerin-19 and *S.l.* suckerin-1. The materials engineered from this construct also exhibit mechanical properties that can be further enhanced for material applications. To identify the structural and link the structure-property relationship for suckerins in general, this construct was successfully expressed and refolded in 20 mM sodium acetate at pH 5.0 from inclusion bodies. Preliminary investigations indicated monomeric conformation with β -sheet content. ID-NMR also indicated non-aggregated conformation. Therefore based on these preliminary data, multi-dimensional NMR experiments were performed. The triple resonances were assigned using automated assignment in Cyana. The long-range NOESY interactions were determined to calculate the structure of protein in 20 mM sodium acetate pH 5.0. Four β -strands are observed in the core of the structure and the terminus ends adopt flexible conformations. Residues from both M1 and M2 domains assume β -strand conformations. In support to the aforementioned results, H/D-Ex indicated that seprio monomer adopted a dynamic structural fold, with weak hydrogen bond interactions and few solvent protected interactions that stabilize the structural core to further. We speculate that on self-assembly or oligomerization the terminal ends which are flexible and dynamic in the monomeric form conform also conform into β -strands, due to inter-molecular β -sheet interactions. The transition from a highly dynamic structure into a very stable, solid supramolecular network is likely a key feature underlying SRT self-assembly.

Conclusion and Perspective

Conclusions

Nature provides a myriad of sophisticated biological materials with a rich spectrum of mechanical properties. Biomimetic researchers derive inspiration from these sophisticated biomaterials to engineer or mimic the form and function of their natural designs. Together with phenomenological and descriptive appreciation, biomimetic science also attempts to elucidate structure-property-function relationships in natural biological systems, such that the underlying design principles can be translated into fabrication of new biomimetic materials. The current study has been inspired by the grappling hook like structures found in the pelagic *D. gigas* squid namely the SRT. Commonly used for predatory attacks, the SRT is an unmineralized tissue, with neither chitin nor covalent cross-links. Rather, it is exclusively composed of block-copolymer like structural proteins called the suckerins, which are considered to be reminiscent of the silk proteins. It is thus critical to understand the fundamental structural molecular features of the suckerin proteins in order to create load bearing biomacromolecular materials that can rival existing synthetic polymers. This study used various complimentary methods in an attempt to characterize the structures of suckerin proteins.

It is reported that SRT is an assembly of semi-crystalline suckerins that form a supramolecular network reinforced by heat resistant nanoconfined β -sheet structures. In this project, we initially investigated whether the structural organization of suckerin proteins in processed SRT was maintained. In the drawn melt-spun SRT fibers, the SAXS pattern suggested anisotropy, whereas the WAXS pattern depicted preserved β -sheet structure. Thus the structural organization in the elongated nanofibrils is maintained similar to the bi-phasic semi-crystalline polymer network found in native SRT. Moreover, the heat-resistant nanocrystalline β -sheet domains embedded in the amorphous regions exhibit viscous flow at higher temperature (below the melting point of the β -sheets) due to chain re-arrangement in the amorphous matrix. This suggested that suckerin proteins exhibit excellent high temperature stability arising from conserved β -sheet topologies and thus would enable easy processability into novel materials.

To further understand the atomic structural features of suckerin proteins, suckerin-19, one of the most abundant protein in *D. gigas* SRT, was chosen as a representative and characterized combining experimental and computational methods. Since the protein was expressed in inclusion bodies, several re-naturation screens were performed to refold the denatured protein. Similar to the native SRT, suckerin-19 in 50 mM MES exhibited β -sheet structure which is stable up to 90 °C. The measurements illustrated an oligomeric conformation stabilized by intermolecular β -sheet interactions. The 1D-NMR spectrum depicted a partial fold of this protein in the buffer used. Given that we could not obtain further insights into the overall structure using available experimental methods, we used *ab initio* computational modeling (using web-based tool I-Tasser) to build the 3-dimensional structure of suckerin-19 at an atomic level (structural models based on homology could not be built due to the absence of a homologous template structure).

Molecular dynamics simulations on the repetitive modules of the sequence of suckerin-19 suggested β -sheets are not exclusively restricted to the Ala- and His-rich repeats, but can also be formed the GGX motifs in the Gly-rich modules, which were previously thought to be strictly amorphous. Consistent with the experimental results, the full length I-Tasser models predicted for suckerin-19 were enriched in β -sheets and were in consensus with the local secondary structure conformations predicted using several web-based secondary structure prediction servers and our simulations. Simulation results of the globular-like model of the protein suggested, that the intermolecular β -sheet domains connected *via* long flexible Gly-rich loops stabilize the largely dynamic suckerin structure. Similar to the native SRT complex, suckerin-19 exhibits anisotropy as the β -sheets domains are randomly dispersed within the flexible Gly-rich loops. The understanding of atomistic features, specifically the spatial arrangement of Tyr residues that have been exploited for generating cross-linked suckerin-19 films has been useful to hypothesize the existence of correlations between in crosslinking content and structural disorder in the recombinant films. The model highlighted that Tyr residues in the flexible loops (that are more accessible) initially participate in cross-linking. However as cross-linking proceeds, Tyr residues in the disordered regions saturate and therefore the Tyr residues from the β -sheet domains were recruited for di-Tyr cross-linking, leading to structural re-arrangement. This

structural re-arrangement frustrates β -sheet formation, in turn explaining the more disordered structure experimentally measured in these highly cross-linked films.

A short *de novo* suckerin construct called the seprio monomer was derived based on its homology with *D.g.* suckerin-19 and *S.l.* suckerin-1. The construct harbors similar modular design (Pro-M1-Pro-M2) characteristic of suckerin protein family. Due to its lower MW and protein composition, the construct was readily refolded in 20 mM sodium acetate, pH 5.0 up to a solubility limit of 0.1 mM (favorable for NMR structure determination). Preliminary biophysical experiments indicated that the acquired structural fold of seprio monomer is β -sheet enriched (also a characteristic feature of SRT). However unlike suckerin-19, which showed oligomeric behavior in the multiple buffer conditions, seprio monomer indicated a globule-like conformation with a hydrodynamic radius of 6 nm in sodium acetate. Based on these data, multidimensional NMR spectra were acquired in this buffer condition for structure calculation. The assignment of protein backbone was successfully done using automated assignment in Cyana. The core region of the protein is composed of four β -strands formed by residues from both M1 and M2 domains (in consensus with the MD simulation results) and on an average the length of each strand is 9-10 residues, which is in agreement with the X-ray diffraction data acquired for native SRT. The terminals conform into dynamic regions and are speculated to conform into β -strands during self-assembly of SRT or oligomerization due to inter-molecular β -sheet interactions. Along with the H/D-Ex experiments it is shown that seprio monomer adopts a dynamic structural fold, with weak hydrogen bond interactions.

Overall this work concludes that similar to SRT, its constituent proteins adopt a relatively dynamic structure with β -sheet interspersed within the amorphous domains. However, being similar in primary sequence design, on interaction with other suckerin proteins (as in the SRT) the dynamic conformations are transformed into more stable structures, which are relatively enriched in β -sheets. While these β -sheets impart mechanical strength, the dynamic and flexible domains in suckerins are responsible for their extensibility. This work describes the molecular level features of the suckerin proteins that can be engineered or modulated to provide an effective template for mimicking wide range of bio-inspired materials with targeted mechanical properties.

Outlooks and future work

A major portion of this research involved attempts at recovering suckerin-19 in its native conformation for which variety of refolding conditions were examined. In addition to this, screening for selection of soluble variants of the suckerin-19 was also carried out. For this, smaller constructs of suckerin-19 were cloned (either by deletion of the amorphous domain to smaller lengths or fragmenting the protein to a smaller size) and a few smaller molecular weight proteins were recombinantly expressed. Under high-levels of expression, several hydrophobic stretches (present in abundance in the suckerin protein) are exposed that result in aggregation and instability. These unstable buildups are formed due to an unbalanced equilibrium between protein solubilization and aggregation. All attempts undertaken to express selected suckerin proteins and the derived constructs lead to protein expression in inclusion bodies. While inclusion bodies do offer several advantages [301], it is not easy to refold the protein from its aggregated state. Hence it is clear that solubility is indeed a rate-limiting factor in obtaining suckerin protein, especially for the purpose of structural and biophysical characterization

Towards obtaining soluble proteins, different *E. coli* strains and cultivation strategies for heterologous expression have been experimented. We also tried to modify the target protein by deletion and fragmentation. Other strategies that can be explored are for example, the engineering of the suckerin proteins using fusion protein technology. While fusion tags or affinity tags are popularly used for purification of proteins, the fusion moiety can alternatively act as a solubilizing partner to improve refolding [301-303]. Among the potent solubility enhancing proteins, *E.coli* maltose binding protein (MBP, 40 kDa) is especially suited for proteins, which tend to form inclusion bodies [302], such as the suckerins. Moreover it has been proved to enhance folding/solubility than highly soluble GST and thioredoxin protein [304].

Thus, suckerin-19 can be engineered with fusion tags like the *E. coli* MBP, which might enhance its solubility from inclusion bodies to the cytoplasm. The fusion tag can be placed at the N- or the C-terminus of the suckerin to generate an in-frame fusion with the polypeptide tag-coding sequence. The expression vector can be designed to constitute a short hydrophilic linker peptide that will ensure spatial separation between the fused tag and the suckerin protein. An endoprotease cleavage site can be

incorporated in this linker, which can later (post purification) be exploited to cleave the suckerin protein from the MBP tag. While solubility enhancement by MBP is passive and not well understood, available evidence suggests that folding is mediated with the assistance of MBP-recruited chaperons present in the vicinity of the passenger protein [305, 306] or that the protein folds spontaneously [307]. The target suckerin can be purified easily from other bacterial proteins by one-step affinity chromatography. As MBP exhibits strong affinity to amylose resin, the tagged suckerin can be purified using amylose-agarose columns [308]. Subsequently the suckerin protein can be excised by specific protease cleavage in the linker for further structural characterization. In addition to the *E. coli* MBP there are other fusion tags such as the *E. coli* N-utilizing substance A (NusA, 54.8 kDa) [309] that is also considered to be an efficient solubility enhancers for aggregation-prone partner proteins [310]. However, since there are no structural or physiochemical similarities between the available solubility tags, the choice of the fusion tag that might work best for any given protein needs to be experimentally verified.

In addition, SRT is known to be a block copolymer of proteins, thus it is possible that synergistic protein complex formation of suckerins may initiate structural stabilization and hence solubility through intermolecular β -sheet interactions; suckerin-19 has indeed been shown to form stable β -sheets-enriched oligomers. The proteins can either be co-expressed to obtain soluble proteins or refolding conditions (for proteins expressed in inclusion bodies) can be established for the assembly of two suckerins to form a complex. The choice of proteins can be based on either difference in their isoelectric points, in the distribution of their modular domains, or in their molecular weights.

In Chapter 6, we have attempted to derive the folded structure of soluble seprio monomer using NMR spectroscopy. The sequential walking of residues was successfully achieved by combined analyses of triple resonance experiments with ^{15}N -HSQC. The final assignment of NOESY cross-peaks is currently in process to determine the final structural conformation of the seprio monomer. Subsequently after the determination of the fold by NMR, the structure and dynamics of this protein can be further investigated through MD simulations and the observations can be used to derive aspects of structural and mechanical properties of suckerin proteins. A major

limitation of assessing the mechanical properties by experimental methods is the lack of details at the nano- or atomic-scale, which is also limited by sample preparations. The experimental findings can be combined with computational models and simulations to examine the influence of building blocks of suckerin and their acquired structure on the overall strength, extensibility, and toughness. The NMR derived structure can also be used as a template to predict model tertiary structures of other proteins in suckerin family. To understand the mechanism underpinning self-assembly or oligomerization of the suckerin proteins, atomistic simulations cannot be carried out since the associated time scales of such physiological processes are extremely long and cannot reliably be accessed by current state-of-the-art atomistic simulations. To overcome this limitation, a promising strategy is to use coarse-grained representation of proteins and subsequent MD simulations [311, 312]. In such simulations, the number of degrees of freedom is decreased significantly by grouping atoms into pseudo-atoms (particles), often referred as beads. The associated parameters that define the interactions at the coarse grained level are developed from the interactions at a finer level of representation. This process in principle can be continued in a hierarchical manner and has potential to reach any level of representation in space and time. Thus using this method, the process of assembly or complex formation of full-length suckerin proteins can be modeled and studied at larger dimensions up, at length scales of micrometers or higher and timescales of milliseconds or higher [313].

One of the major goals of biomimetic research is to deliver sustainable materials that feature distinguishing mechanical properties. Suckerins and the derived biomaterials have already been shown to exhibit remarkable mechanical properties. Based on the knowledge of their structure, specifically the spatial organization of certain residues in the fold, various site-specific mutations can be suggested that can either enhance the mechanical properties or additional functionalities can be engineered by chemical modification of certain moieties. For instance using mussel adhesion system as an inspiration, the Tyr residues in the Gly-rich domains of suckerins can be modified into Dopa to engineer a suckerin-based material with additional adhesive properties. Thus additional functionality of adhesion can be incorporated in suckerins while conserving its own existing remarkable mechanical properties.

Bibliography

1. Mojzsis SJ, Arrhenius G, McKeegan KD, Harrison TM, Nutman AP, Friend CRL: Evidence for life on Earth before 3,800 million years ago. *Nature* 1996, 384(6604):55-59.
2. Sarikaya M: Biomimetics: materials fabrication through biology. *Proc Natl Acad Sci U S A* 1999, 96(25):14183-14185.
3. Fratzl P: Biomimetic materials research: what can we really learn from nature's structural materials? *J R Soc Interface* 2007, 4(15):637-642.
4. Bhushan B: Biomimetics: lessons from nature--an overview. *Philos Trans A Math Phys Eng Sci* 2009, 367(1893):1445-1486.
5. Wegst UGKA, M. F.: The mechanical efficiency of natural materials. *Philos Mag* 2004, 84:2167-2181.
6. Gosline JM, Guerette PA, Ortlepp CS, Savage KN: The mechanical design of spider silks: from fibroin sequence to mechanical function. *J Exp Biol* 1999, 202(Pt 23):3295-3303.
7. Wegst UGK, Bai H, Saiz E, Tomsia AP, Ritchie RO: Bioinspired structural materials. *Nat Mater* 2015, 14(1):23-36.
8. Jeronimidis G, Atkins A: Mechanics of biological materials and structures: Nature's lessons for the engineer. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 1995, 209(4):221-235.
9. Vincent JF, Bogatyreva OA, Bogatyrev NR, Bowyer A, Pahl AK: Biomimetics: its practice and theory. *J R Soc Interface* 2006, 3(9):471-482.
10. B. Alberts AJ, J. Lewis, M. Raff, K. Roberts and P. Walter: Molecular biology of Cell. 2002.
11. Silva NHCS, Vilela C, Marrucho IM, Freire CSR, Pascoal Neto C, Silvestre AJD: Protein-based materials: from sources to innovative sustainable materials for biomedical applications. *Journal of Materials Chemistry B* 2014, 2(24):3715-3740.
12. Wang N, Stamenovic D: Mechanics of vimentin intermediate filaments. *J Muscle Res Cell Motil* 2002, 23(5-6):535-540.
13. Gelse K, Poschl E, Aigner T: Collagens--structure, function, and biosynthesis. *Adv Drug Deliv Rev* 2003, 55(12):1531-1546.
14. Fratzl P, Weinkamer R: Nature's hierarchical materials. *Progress in Materials Science* 2007, 52(8):1263-1334.
15. Weisel JW: Biophysics. Enigmas of blood clot elasticity. *Science* 2008, 320(5875):456-457.
16. Moir R, Spann T: The structure and function of nuclear lamins: implications for disease. *Cellular and Molecular Life Sciences CMLS* 2001, 58(12-13):1748-1757.

17. Wilson KL, Zastrow MS, Lee KK: Lamins and disease: insights into nuclear infrastructure. *Cell* 2001, 104(5):647-650.
18. Lührs T, Ritter C, Adrian M, Riek-Loher D, Bohrmann B, Döbeli H, Schubert D, Riek R: 3D structure of Alzheimer's amyloid- β (1–42) fibrils. *Proceedings of the National Academy of Sciences of the United States of America* 2005, 102(48):17342-17347.
19. Gronau G, Krishnaji ST, Kinahan ME, Giesa T, Wong JY, Kaplan DL, Buehler MJ: A review of combined experimental and computational procedures for assessing biopolymer structure–process–property relationships. *Biomaterials* 2012, 33(33):8240-8255.
20. Ackbarow T, Chen X, Keten S, Buehler MJ: Hierarchies, multiple energy barriers, and robustness govern the fracture mechanics of α -helical and β -sheet protein domains. *Proceedings of the National Academy of Sciences* 2007, 104(42):16410-16415.
21. Bryson JW, Betz SF, Lu HS, Suich DJ, Zhou HX, O'Neil KT, DeGrado WF: Protein design: a hierarchic approach. *Science* 1995, 270(5238):935-941.
22. Yip S: Nanocrystals: The strongest size. *Nature* 1998, 391(6667):532-533.
23. Blanckenhagen Bv, Gumbsch P, Arzt E: Dislocation sources in discrete dislocation simulations of thin-film plasticity and the Hall-Petch relation. *Modelling and Simulation in Materials Science and Engineering* 2001, 9(3):157.
24. Buehler MJ, Keten S, Ackbarow T: Theoretical and computational hierarchical nanomechanics of protein materials: Deformation and fracture. *Progress in Materials Science* 2008, 53(8):1101-1241.
25. Csányi V: Hierarchy theory: A vision, vocabulary, and epistemology by Valerie Ahl and T.F.H. Allen. *Complexity* 1999, 4(6):29-30.
26. Meyers MA, Chen P-Y, Lin AY-M, Seki Y: Biological materials: Structure and mechanical properties. *Progress in Materials Science* 2008, 53(1):1-206.
27. Kaye R, Head E, Thompson JL, McIntire TM, Milton SC, Cotman CW, Glabe CG: Common structure of soluble amyloid oligomers implies common mechanism of pathogenesis. *Science* 2003, 300(5618):486-489.
28. Cranford S, Buehler MJ: Materiomics: biological protein materials, from nano to macro. *Nanotechnol Sci Appl* 2010, 3:127-148.
29. Knowles TP, Oppenheim TW, Buell AK, Chirgadze DY, Welland ME: Nanostructured films from hierarchical self-assembly of amyloidogenic proteins. *Nat Nanotechnol* 2010, 5(3):204-207.
30. Gao H, Ji B, Jager IL, Arzt E, Fratzl P: Materials become insensitive to flaws at nanoscale: lessons from nature. *Proc Natl Acad Sci U S A* 2003, 100(10):5597-5600.
31. Buehler MJ, Yung YC: Deformation and failure of protein materials in physiologically extreme conditions and disease. *Nat Mater* 2009, 8(3):175-188.
32. Buehler MJ: Multiscale mechanics of biological and biologically inspired materials and structures. *Acta Mechanica Sinica* 2010, 23(6):471-483.

33. Knowles TPJ, Oppenheim TW, Buell AK, Chirgadze DY, Welland ME: Nanostructured films from hierarchical self-assembly of amyloidogenic proteins. *Nat Nano* 2010, 5(3):204-207.
34. Watt SW, McEwen IJ, Viney C: Stability of Molecular Order in Silkworm Silk. *Macromolecules* 1999, 32(25):8671-8673.
35. Romer L, Scheibel T: The elaborate structure of spider silk: structure and function of a natural high performance fiber. *Prion* 2008, 2(4):154-161.
36. Vollrath F: Strength and structure of spiders' silks. *J Biotechnol* 2000, 74(2):67-83.
37. Van Hest JCM, Tirrell DA: Protein-based materials, toward a new level of structural control. *Chemical Communications* 2001(19):1897-1904.
38. Jin HJ, Kaplan DL: Mechanism of silk processing in insects and spiders. *Nature* 2003, 424(6952):1057-1061.
39. Altman GH, Diaz F, Jakuba C, Calabro T, Horan RL, Chen J, Lu H, Richmond J, Kaplan DL: Silk-based biomaterials. *Biomaterials* 2003, 24(3):401-416.
40. Simmons AH, Michal CA, Jelinski LW: Molecular orientation and two-component nature of the crystalline fraction of spider dragline silk. *Science* 1996, 271(5245):84-87.
41. Shao Z, Vollrath F, Sirichaisit J, Young RJ: Analysis of spider silk in native and supercontracted states using Raman spectroscopy. *Polymer* 1999, 40(10):2493-2500.
42. Sampath S, Isdebski T, Jenkins JE, Ayon JV, Henning RW, Orgel JP, Antipova O, Yarger JL: X-ray diffraction study of nanocrystalline and amorphous structure within major and minor ampullate dragline spider silks. *Soft Matter* 2012, 8(25):6713-6722.
43. Thiel BL, Guess KB, Viney C: Non-periodic lattice crystals in the hierarchical microstructure of spider (major ampullate) silk. *Biopolymers* 1997, 41(7):703-719.
44. Van Beek JD, Hess S, Vollrath F, Meier BH: The molecular structure of spider dragline silk: Folding and orientation of the protein backbone. *Proceedings of the National Academy of Sciences* 2002, 99(16):10266-10271.
45. Guerette PA, Ginzinger DG, Weber BH, Gosline JM: Silk properties determined by gland-specific expression of a spider fibroin gene family. *Science* 1996, 272(5258):112-115.
46. Teulé F, Furin WA, Cooper AR, Duncan JR, Lewis RV: Modifications of spider silk sequences in an attempt to control the mechanical properties of the synthetic fibers. *Journal of Materials Science* 2007, 42(21):8974-8985.
47. Lefèvre T, Rousseau M-E, Pézolet M: Protein secondary structure and orientation in silk as revealed by Raman spectromicroscopy. *Biophysical journal* 2007, 92(8):2885-2895.
48. Keten S, Xu Z, Ihle B, Buehler MJ: Nanoconfinement controls stiffness, strength and mechanical toughness of beta-sheet crystals in silk. *Nat Mater* 2010, 9(4):359-367.

49. Hayashi CY, Shipley NH, Lewis RV: Hypotheses that correlate the sequence, structure, and mechanical properties of spider silk proteins. *Int J Biol Macromol* 1999, 24(2-3):271-275.
50. Vepari C, Kaplan DL: Silk as a biomaterial. *Progress in Polymer Science* 2007, 32(8):991-1007.
51. Giesa T, Arslan M, Pugno NM, Buehler MJ: Nanoconfinement of spider silk fibrils begets superior strength, extensibility, and toughness. *Nano Lett* 2011, 11(11):5038-5046.
52. Chen Q, Pugno NM: Bio-mimetic mechanisms of natural hierarchical materials: a review. *J Mech Behav Biomed Mater* 2013, 19:3-33.
53. Hagn F, Eisoldt L, Hardy JG, Vendrely C, Coles M, Scheibel T, Kessler H: A conserved spider silk domain acts as a molecular switch that controls fibre assembly. *Nature* 2010, 465(7295):239-242.
54. Askarieh G, Nordling K, Saenz A, Casals C, Rising A, Johansson J, Knight SD: Self-assembly of spider silk proteins is controlled by a pH-sensitive relay. *Nature* 2010, 465(7295):236-238.
55. Sponner A, Vater W, Rommerskirch W, Vollrath F, Unger E, Grosse F, Weisshart K: The conserved C-termini contribute to the properties of spider silk fibroins. *Biochem Biophys Res Commun* 2005, 338(2):897-902.
56. Tao H, Kaplan DL, Omenetto FG: Silk Materials – A Road to Sustainable High Technology. *Advanced Materials* 2012, 24(21):2824-2837.
57. Qu Y, Payne SC, Apkarian RP, Conticello VP: Self-Assembly of a Polypeptide Multi-Block Copolymer Modeled on Dragline Silk Proteins. *Journal of the American Chemical Society* 2000, 122(20):5014-5015.
58. Gomes S, Leonor IB, Mano JF, Reis RL, Kaplan DL: Natural and genetically engineered proteins for tissue engineering. *Progress in Polymer Science* 2012, 37(1):1-17.
59. Hu X, Cebe P, Weiss AS, Omenetto F, Kaplan DL: Protein-based composite materials. *Materials Today* 2012, 15(5):208-215.
60. Fratzl P, Misof K, Zizak I, Rapp G, Amenitsch H, Bernstorff S: Fibrillar structure and mechanical properties of collagen. *J Struct Biol* 1998, 122(1-2):119-122.
61. Fratzl P: Collagen: structure and mechanics, an introduction: Springer; 2008.
62. Van der Rest M, Garrone R: Collagen family of proteins. *FASEB J* 1991, 5(13):2814-2823.
63. Chen P-Y, McKittrick J, Meyers MA: Biological materials: Functional adaptations and bioinspired designs. *Progress in Materials Science* 2012, 57(8):1492-1704.
64. Gautieri A, Vesentini S, Redaelli A, Buehler MJ: Hierarchical Structure and Nanomechanics of Collagen Microfibrils from the Atomistic Scale Up. *Nano Letters* 2011, 11(2):757-766.
65. Mouw JK, Ou G, Weaver VM: Extracellular matrix assembly: a multiscale deconstruction. *Nat Rev Mol Cell Biol* 2014, 15(12):771-785.

66. Buehler MJ: Nature designs tough collagen: Explaining the nanostructure of collagen fibrils. *Proceedings of the National Academy of Sciences* 2006, 103(33):12285-12290.
67. Chang S-W, Buehler MJ: Molecular biomechanics of collagen molecules. *Materials Today* 2014, 17(2):70-76.
68. Uzel SGM, Buehler MJ: Nanomechanical sequencing of collagen: tropocollagen features heterogeneous elastic properties at the nanoscale. *Integrative Biology* 2009, 1(7):452-459.
69. Vrhovski B, Weiss AS: Biochemistry of tropoelastin. *Eur J Biochem* 1998, 258(1):1-18.
70. Rodgers UR, Weiss AS: Cellular interactions with elastin. *Pathol Biol (Paris)* 2005, 53(7):390-398.
71. Gosline J, Lillie M, Carrington E, Guerette P, Ortlepp C, Savage K: Elastic proteins: biological roles and mechanical properties. *Philos Trans R Soc Lond B Biol Sci* 2002, 357(1418):121-132.
72. Daamen WF, Veerkamp JH, van Hest JC, van Kuppevelt TH: Elastin as a biomaterial for tissue engineering. *Biomaterials* 2007, 28(30):4378-4398.
73. Wise SG, Weiss AS: Tropoelastin. *Int J Biochem Cell Biol* 2009, 41(3):494-497.
74. Muiznieks LD, Weiss AS, Keeley FW: Structural disorder and dynamics of elastin. *Biochem Cell Biol* 2010, 88(2):239-250.
75. Hoeve CA, Flory PJ: The elastic properties of elastin. *Biopolymers* 1974, 13(4):677-686.
76. Aaron BB, Gosline JM: Elastin as a random-network elastomer: A mechanical and optical analysis of single elastin fibers. *Biopolymers* 1981, 20(6):1247-1260.
77. Hoeve C, Flory P: The elastic properties of elastin. *Biopolymers* 1974, 13(4):677-686.
78. Dorrington K, McCrum N: Elastin as a rubber. *Biopolymers* 1977, 16(6):1201-1222.
79. Mark JE, Erman B: Rubberlike elasticity: a molecular primer: Cambridge University Press; 2007.
80. Pometun MS, Chekmenev EY, Wittebort RJ: Quantitative observation of backbone disorder in native elastin. *Journal of Biological Chemistry* 2004, 279(9):7982-7987.
81. Perry A, Stypa MP, Tenn BK, Kumashiro KK: Solid-State ¹³ C NMR Reveals Effects of Temperature and Hydration on Elastin. *Biophysical journal* 2002, 82(2):1086-1095.
82. Urry DW, Harris RD, Long MM, Prasad KU: Polytetrapeptide of elastin. Temperature-correlated elastomeric force and structure development. *Int J Pept Protein Res* 1986, 28(6):649-660.
83. Muiznieks LD, Jensen SA, Weiss AS: Structural changes and facilitated association of tropoelastin. *Arch Biochem Biophys* 2003, 410(2):317-323.

84. Bochicchio B, Floquet N, Pepe A, Alix AJ, Tamburro AM: Dissection of human tropoelastin: solution structure, dynamics and self-assembly of the exon 5 peptide. *Chemistry* 2004, 10(13):3166-3176.
85. Roberts S, Dzuricky M, Chilkoti A: Elastin-like polypeptides as models of intrinsically disordered proteins. *FEBS Lett* 2015, 589(19 Pt A):2477-2486.
86. Urry DW: Physical chemistry of biological free energy transduction as demonstrated by elastic protein-based polymers. *The Journal of Physical Chemistry B* 1997, 101(51):11007-11028.
87. Urry DW: Five axioms for the functional design of peptide-based polymers as molecular machines and materials: Principle for macromolecular assemblies. *Peptide Science* 1998, 47(2):167-178.
88. Vendruscolo M, Dobson CM: Towards complete descriptions of the free-energy landscapes of proteins. *Philos Trans A Math Phys Eng Sci* 2005, 363(1827):433-450; discussion 450-432.
89. Rauscher S, Pomes R: Molecular simulations of protein disorder. *Biochem Cell Biol* 2010, 88(2):269-290.
90. Tompa P: Structural disorder in amyloid fibrils: its implication in dynamic interactions of proteins. *FEBS J* 2009, 276(19):5406-5415.
91. Rauscher S, Pomès R: Structural disorder and protein elasticity. In: *Fuzziness*. Springer; 2012: 159-183.
92. Li B, Alonso DO, Bennion BJ, Daggett V: Hydrophobic hydration is an important source of elasticity in elastin-based biopolymers. *Journal of the American Chemical Society* 2001, 123(48):11991-11998.
93. Li B, Alonso DO, Daggett V: The molecular basis for the inverse temperature transition of elastin. *Journal of molecular biology* 2001, 305(3):581-592.
94. Zagrovic B, Lipfert J, Sorin EJ, Millett IS, van Gunsteren WF, Doniach S, Pande VS: Unusual compactness of a polyproline type II structure. *Proc Natl Acad Sci U S A* 2005, 102(33):11698-11703.
95. Almine JF, Bax DV, Mithieux SM, Nivison-Smith L, Rnjak J, Waterhouse A, Wise SG, Weiss AS: Elastin-based materials. *Chem Soc Rev* 2010, 39(9):3371-3379.
96. Kim W, Chaikof EL: Recombinant elastin-mimetic biomaterials: emerging applications in medicine. *Advanced drug delivery reviews* 2010, 62(15):1468-1478.
97. Nettles DL, Chilkoti A, Setton LA: Applications of elastin-like polypeptides in tissue engineering. *Advanced drug delivery reviews* 2010, 62(15):1479-1485.
98. McDaniel JR, Callahan DJ, Chilkoti A: Drug delivery to solid tumors by elastin-like polypeptides. *Advanced drug delivery reviews* 2010, 62(15):1456-1467.
99. Coyne KJ, Qin XX, Waite JH: Extensible collagen in mussel byssus: a natural block copolymer. *Science* 1997, 277(5333):1830-1832.
100. Coyne KJ, Waite JH: In search of molecular dovetails in mussel byssus: from the threads to the stem. *J Exp Biol* 2000, 203(Pt 9):1425-1431.

101. Brown C: Some structural proteins of *Mytilus edulis*. *Quarterly Journal of Microscopical Science* 1952, 3(24):487-502.
102. Lee BP, Messersmith PB, Israelachvili JN, Waite JH: Mussel-Inspired Adhesives and Coatings. *Annu Rev Mater Res* 2011, 41:99-132.
103. Yu J, Wei W, Danner E, Ashley RK, Israelachvili JN, Waite JH: Mussel protein adhesion depends on interprotein thiol-mediated redox modulation. *Nat Chem Biol* 2011, 7(9):588-590.
104. Wei W, Yu J, Broomell C, Israelachvili JN, Waite JH: Hydrophobic enhancement of Dopa-mediated adhesion in a mussel foot protein. *J Am Chem Soc* 2013, 135(1):377-383.
105. Yu J, Kan Y, Rapp M, Danner E, Wei W, Das S, Miller DR, Chen Y, Waite JH, Israelachvili JN: Adaptive hydrophobic and hydrophilic interactions of mussel foot proteins with organic thin films. *Proc Natl Acad Sci U S A* 2013, 110(39):15680-15685.
106. Hwang DS, Zeng H, Masic A, Harrington MJ, Israelachvili JN, Waite JH: Protein- and metal-dependent interactions of a prominent protein in mussel adhesive plaques. *J Biol Chem* 2010, 285(33):25850-25858.
107. Petrone L, Kumar A, Sutanto CN, Patil NJ, Kannan S, Palaniappan A, Amini S, Zappone B, Verma C, Miserez A: Mussel adhesion is dictated by time-regulated secretion and molecular conformation of mussel adhesive proteins. *Nat Commun* 2015, 6:8737.
108. Guerette PA, Hoon S, Seow Y, Raida M, Masic A, Wong FT, Ho VH, Kong KW, Demirel MC, Pena-Francesch A *et al*: Accelerating the design of biomimetic materials by integrating RNA-seq with proteomics and materials science. *Nat Biotechnol* 2013, 31(10):908-915.
109. Buehler MJ, Yung YC: How protein materials balance strength, robustness, and adaptability. *HFSP Journal* 2010, 4(1):26-40.
110. Zeidberg LD, Robison BH: Invasive range expansion by the Humboldt squid, *Dosidicus gigas*, in the eastern North Pacific. *Proceedings of the National Academy of Sciences* 2007, 104(31):12948-12950.
111. Wormuth JH: The biogeography and numerical taxonomy of the oegopsid squid family Ommastrephidae in the Pacific Ocean, vol. 23: Univ of California Press; 1976.
112. Staaf DJ, Zeidberg LD, Gilly WF: Effects of temperature on embryonic development of the Humboldt squid *Dosidicus gigas*. *Marine Ecology Progress Series* 2011, 441:165-175.
113. Tan Y, Hoon S, Guerette PA, Wei W, Ghadban A, Hao C, Miserez A, Waite JH: Infiltration of chitin by protein coacervates defines the squid beak mechanical gradient. *Nature chemical biology* 2015, 11(7):488-495.
114. Miserez A, Li Y, Waite JH, Zok F: Jumbo squid beaks: inspiration for design of robust organic composites. *Acta Biomater* 2007, 3(1):139-149.
115. Miserez A, Schneberk T, Sun C, Zok FW, Waite JH: The transition from stiff to compliant materials in squid beaks. *Science* 2008, 319(5871):1816-1819.

116. Miserez A, Rubin D, Waite JH: Cross-linking chemistry of squid beak. *J Biol Chem* 2010, 285(49):38115-38124.
117. Miserez A, Weaver JC, Pedersen PB, Schneeberk T, Hanlon RT, Kisailus D, Birkedal H: Microstructural and biochemical characterization of the nanoporous sucker rings from *Dosidicus gigas*. *Advanced Materials* 2009, 21(4):401-406.
118. Zok FW, Miserez A: Property maps for abrasion resistance of materials. *Acta Mater* 2007, 55(18):6365-6371.
119. Ding D, Guerette PA, Hoon S, Kong KW, Cornvik T, Nilsson M, Kumar A, Lescar J, Miserez A: Biomimetic Production of Silk-Like Recombinant Squid Sucker Ring Teeth Proteins. *Biomacromolecules* 2014, 15(9):3278-3289.
120. Guerette PA, Hoon S, Ding D, Amini S, Masic A, Ravi V, Venkatesh B, Weaver JC, Miserez A: Nanoconfined beta-sheets mechanically reinforce the supra-biomolecular network of robust squid Sucker Ring Teeth. *ACS Nano* 2014, 8(7):7170-7179.
121. Wang R, Gupta HS: Deformation and fracture mechanisms of bone and nacre. *Annual Review of Materials Research* 2011, 41:41-73.
122. Studart AR: Towards high-performance bioinspired composites. *Adv Mater* 2012, 24(37):5024-5044.
123. Andersen SO: Insect cuticular sclerotization: a review. *Insect Biochemistry and Molecular Biology* 2010, 40(3):166-178.
124. Gatesy J, Hayashi C, Motriuk D, Woods J, Lewis R: Extreme diversity, conservation, and convergence of spider silk fibroin sequences. *Science* 2001, 291(5513):2603-2605.
125. Yano M, Nagai K, Morimoto K, Miyamoto H: Shematrin: A family of glycine-rich structural proteins in the shell of the pearl oyster *Pinctada fucata*. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* 2006, 144(2):254-262.
126. Rousseau ME, Lefevre T, Beaulieu L, Asakura T, Pezolet M: Study of protein conformation and orientation in silkworm and spider silk fibers using Raman microspectroscopy. *Biomacromolecules* 2004, 5(6):2247-2257.
127. Ulrich S, Glisovic A, Salditt T, Zippelius A: Diffraction from the beta-sheet crystallites in spider silk. *Eur Phys J E Soft Matter* 2008, 27(3):229-242.
128. Martel A, Burghammer M, Davies RJ, Riekel C: Thermal behavior of Bombyx mori silk: evolution of crystalline parameters, molecular structure, and mechanical properties. *Biomacromolecules* 2007, 8(11):3548-3556.
129. Latza V, Guerette PA, Ding D, Amini S, Kumar A, Schmidt I, Keating S, Oxman N, Weaver JC, Fratzl P *et al*: Multi-scale thermal stability of a hard thermoplastic protein-based material. *Nat Commun* 2015, 6:8313.
130. Ding D, Guerette PA, Fu J, Zhang L, Irvine SA, Miserez A: From Soft Self-Healing Gels to Stiff Films in Suckerin-Based Materials Through Modulation of Crosslink Density and beta-Sheet Content. *Adv Mater* 2015, 27(26):3953-3961.

131. Karas M, Hillenkamp F: Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical chemistry* 1988, 60(20):2299-2301.
132. Karas M, Bahr U, Gießmann U: Matrix-assisted laser desorption ionization mass spectrometry. *Mass spectrometry reviews* 1991, 10(5):335-357.
133. Jurinke C, Oeth P, van den Boom D: MALDI-TOF mass spectrometry. *Molecular biotechnology* 2004, 26(2):147-163.
134. Bradford MM: A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* 1976, 72:248-254.
135. (Ed.) RP: Dynamic light scattering—applications of photon correlation spectroscopy, Plenum Press, New York. 1985.
136. Schmitz KS: An introduction to dynamic light scattering by macromolecules. Boston: Academic Press.; 1990.
137. Lorber B, Fischer F, Bailly M, Roy H, Kern D: Protein analysis by dynamic light scattering: methods and techniques for students. *Biochemistry and Molecular Biology Education* 2012, 40(6):372-382.
138. Greenfield NJ: Using circular dichroism spectra to estimate protein secondary structure. *Nature protocols* 2007, 1(6):2876-2890.
139. Kelly SM, Jess TJ, Price NC: How to study proteins by circular dichroism. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 2005, 1751(2):119-139.
140. Savitzky A, Golay MJ: Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry* 1964, 36(8):1627-1639.
141. Baker MJ, Trevisan J, Bassan P, Bhargava R, Butler HJ, Dorling KM, Fielden PR, Fogarty SW, Fullwood NJ, Heys KA *et al*: Using Fourier transform IR spectroscopy to analyze biological materials. *Nat Protoc* 2014, 9(8):1771-1791.
142. Yang H, Yang S, Kong J, Dong A, Yu S: Obtaining information about protein secondary structures in aqueous solution using Fourier transform IR spectroscopy. *Nat Protoc* 2015, 10(3):382-396.
143. Krimm S, Bandekar J: Vibrational spectroscopy and conformation of peptides, polypeptides, and proteins. *Adv Protein Chem* 1986, 38:181-364.
144. Infrared Analysis of Peptides and Proteins. Principles and Applications, vol. 750: American Chemical Society; 1999.
145. Bandekar J: Amide modes and protein conformation. *Biochim Biophys Acta* 1992, 1120(2):123-143.
146. Pelton JT, McLean LR: Spectroscopic methods for analysis of protein secondary structure. *Anal Biochem* 2000, 277(2):167-176.
147. Williams DB, Carter C: Barry Transmission electron microscopy: A textbook for materials science. In.: New York: Plenum Press; 1996.
148. Moult J, Fidelis K, Kryzhtafovych A, Schwede T, Tramontano A: Critical assessment of methods of protein structure prediction (CASP)--round x. *Proteins* 2014, 82 Suppl 2:1-6.

149. McGuffin LJ, Bryson K, Jones DT: The PSIPRED protein structure prediction server. *Bioinformatics* 2000, 16(4):404-405.
150. Drozdetskiy A, Cole C, Procter J, Barton GJ: JPred4: a protein secondary structure prediction server. *Nucleic Acids Research* 2015:gkv332.
151. Rost B, Sander C: Prediction of Protein Secondary Structure at Better than 70% Accuracy. *Journal of molecular biology* 1993, 232(2):584-599.
152. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H: Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 2000, 16(5):412-424.
153. Dorn M, MB ES, Buriol LS, Lamb LC: Three-dimensional protein structure prediction: Methods and computational strategies. *Comput Biol Chem* 2014, 53PB:251-276.
154. Rost B, Fariselli P, Casadio R: Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci* 1996, 5(8):1704-1718.
155. Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J: A method for the improvement of threading-based protein models. *Proteins* 1999, 37(4):592-610.
156. Ingwall RT, Scheraga HA, Lotan N, Berger A, Katchalski E: Conformational studies of poly-L-alanine in water. *Biopolymers* 1968, 6(3):331-368.
157. Lesk AM: CASP2: report on ab initio predictions. *Proteins* 1997, Suppl 1:151-166.
158. Zemla A, Venclovas C, Reinhardt A, Fidelis K, Hubbard TJ: Numerical criteria for the evaluation of ab initio predictions of protein structure. *Proteins* 1997, Suppl 1:140-150.
159. Lesk A, Chothia C: The response of protein structures to amino-acid sequence changes. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 1986, 317(1540):345-356.
160. Johnson MS, Srinivasan N, Sowdhamini R, Blundell TL: Knowledge-based protein modeling. *Crit Rev Biochem Mol Biol* 1994, 29(1):1-68.
161. Xiang Z: Advances in homology protein structure modeling. *Curr Protein Pept Sci* 2006, 7(3):217-227.
162. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *Journal of molecular biology* 1990, 215(3):403-410.
163. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: The Protein Data Bank. *Nucleic Acids Res* 2000, 28(1):235-242.
164. Lee J, Wu S, Zhang Y: Ab Initio Protein Structure Prediction. In: *From Protein Structure to Function with Bioinformatics*. Edited by Rigden DJ. Dordrecht: Springer Netherlands; 2009: 3-25.
165. Anfinsen CB: Principles that govern the folding of protein chains. *Science* 1973, 181(4096):223-230.
166. Helles G: A comparative study of the reported performance of ab initio protein structure prediction algorithms. *J R Soc Interface* 2008, 5(21):387-396.

167. Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A: Evaluation of template-based models in CASP8 with standard measures. *Proteins: Structure, Function, and Bioinformatics* 2009, 77(S9):18-28.
168. Kryshtafovych A, Krysko O, Daniluk P, Dmytriv Z, Fidelis K: Protein structure prediction center in CASP8. *Proteins: Structure, Function, and Bioinformatics* 2009, 77(S9):5-9.
169. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O *et al*: Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* 2009, 77 Suppl 9:89-99.
170. Zhang Y: I-TASSER: fully automated protein structure prediction in CASP8. *Proteins* 2009, 77 Suppl 9:100-113.
171. Peng J, Xu J: RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins* 2011, 79 Suppl 10:161-171.
172. Kim DE, Chivian D, Baker D: Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res* 2004, 32(Web Server issue):W526-531.
173. Simons KT, Bonneau R, Ruczinski I, Baker D: Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* 1999, Suppl 3:171-176.
174. Rohl CA, Strauss CE, Misura KM, Baker D: Protein structure prediction using Rosetta. *Methods Enzymol* 2004, 383:66-93.
175. Kallberg M, Margaryan G, Wang S, Ma J, Xu J: RaptorX server: a resource for template-based protein structure modeling. *Methods Mol Biol* 2014, 1137:17-27.
176. Peng J, Xu J: Boosting Protein Threading Accuracy. *Res Comput Mol Biol* 2009, 5541:31-45.
177. Peng J, Xu J: A multiple-template approach to protein threading. *Proteins* 2011, 79(6):1930-1939.
178. Zhang Y: I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 2008, 9:40.
179. Roy A, Kucukural A, Zhang Y: I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protocols* 2010, 5(4):725-738.
180. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y: The I-TASSER Suite: protein structure and function prediction. *Nat Methods* 2015, 12(1):7-8.
181. Karplus M, Gelin BR, McCammon JA: Internal dynamics of proteins. Short time and long time motions of aromatic sidechains in PTI. *Biophysical journal* 1980, 32(1):603.
182. Leach AR: Molecular modelling: principles and applications: Pearson education; 2001.
183. Laskowski RA: Integrated Servers for Structure-Informed Function Prediction. In: *From Protein Structure to Function with Bioinformatics*. Edited by Rigden DJ. Dordrecht: Springer Netherlands; 2009: 251-272.
184. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S, Weiner P: A new force field for molecular mechanical simulation of nucleic

- acids and proteins. *Journal of the American Chemical Society* 1984, 106(3):765-784.
185. Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham TE, DeBolt S, Ferguson D, Seibel G, Kollman P: AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications* 1995, 91(1):1-41.
  186. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M: CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry* 1983, 4(2):187-217.
  187. Van Gunsteren WF, Billeter S, Eising A, Hünenberger PH, Krüger P, Mark AE, Scott W, Tironi IG: Biomolecular simulation: the {GROMOS96} manual and user guide. 1996.
  188. Jorgensen WL, Tirado-Rives J: The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society* 1988, 110(6):1657-1666.
  189. Gunsteren WF, Mark AE: On the interpretation of biochemical data by molecular dynamics computer simulation. *European Journal of Biochemistry* 1992, 204(3):947-961.
  190. Allen MP: Introduction to molecular dynamics simulation. *Computational soft matter: from synthetic polymers to proteins* 2004, 23:1-28.
  191. Case DA, Darden TA, Cheatham TE, Simmerling CL, Wang J, Duke RE, Luo R, Walker RC, Zhang W, Merz KM *et al*: AMBER 12. In.: University of California, San Francisco; 2012.
  192. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C: Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* 2006, 65(3):712-725.
  193. Swope WC, Andersen HC, Berens PH, Wilson KR: A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of chemical physics* 1982, 76(1):637-649.
  194. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML: Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* 1983, 79(2):926-935.
  195. Darden T, York D, Pedersen L: Particle mesh Ewald: An N*log(N) method for Ewald sums in large systems. *The Journal of chemical physics* 1993, 98(12):10089-10092.
  196. Ryckaert J-P, Ciccotti G, Berendsen HJC: Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* 1977, 23(3):327-341.
  197. McQuarrie D: Statistical Mechanics, Harper &. In.: Row Publishers; 1976.

198. Humphrey W, Dalke A, Schulten K: VMD: visual molecular dynamics. *J Mol Graph* 1996, 14(1):33-38, 27-38.
199. Matthews B: X-ray crystallographic studies of proteins. *Annual Review of Physical Chemistry* 1976, 27(1):493-493.
200. Ilari A, Savino C: Protein structure determination by x-ray crystallography. *Methods Mol Biol* 2008, 452:63-87.
201. Shi Y: A glimpse of structural biology through X-ray crystallography. *Cell* 2014, 159(5):995-1014.
202. Durbin SD, Feher G: Protein crystallization. *Annu Rev Phys Chem* 1996, 47:171-204.
203. Wooh JW, Kidd RD, Martin JL, Kobe B: Comparison of three commercial sparse-matrix crystallization screens. *Acta Crystallogr D Biol Crystallogr* 2003, 59(Pt 4):769-772.
204. Dyson HJ, Wright PE: Insights into protein folding from NMR. *Annu Rev Phys Chem* 1996, 47:369-395.
205. Kay LE: NMR studies of protein structure and dynamics. *J Magn Reson* 2005, 173(2):193-207.
206. Jahn TR, Radford SE: Folding versus aggregation: Polypeptide conformations on competing pathways. *Archives of Biochemistry and Biophysics* 2008, 469(1):100-117.
207. Cavanagh J, Fairbrother WJ, Palmer Iii AG, Rance M, Skelton NJ: Chapter 6 - Experimental ¹H NMR Methods. In: *Protein NMR Spectroscopy (Second Edition)*. Burlington: Academic Press; 2007: 405-532.
208. Kwan AH, Mobli M, Gooley PR, King GF, Mackay JP: Macromolecular NMR spectroscopy for the non-spectroscopist. *FEBS J* 2011, 278(5):687-703.
209. Cavanagh J, Fairbrother WJ, Palmer III AG, Skelton NJ, Rance M: Protein NMR Spectroscopy: Principles and Practice: Academic Press; 2010.
210. Cavanagh J, Fairbrother WJ, Palmer Iii AG, Rance M, Skelton NJ: Chapter 7 - Heteronuclear NMR Experiments. In: *Protein NMR Spectroscopy (Second Edition)*. Burlington: Academic Press; 2007: 533-678.
211. Evans J: Biomolecular NMR Spectroscopy. 1995. In.: Oxford: Oxford University Press.
212. Cavanagh J, Fairbrother WJ, Palmer Iii AG, Rance M, Skelton NJ: Chapter 10 - Sequential Assignment, Structure Determination, and other applications. In: *Protein NMR Spectroscopy (Second Edition)*. Burlington: Academic Press; 2007: 781-817.
213. Grzesiek S, Bax A: An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins. *Journal of Magnetic Resonance (1969)* 1992, 99(1):201-207.
214. Grzesiek S, Bax A: Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR. *Journal of the American Chemical Society* 1992, 114(16):6291-6293.

215. KAY LE, IKURA M, TSCHUDIN R, BAX A: Three-Dimensional Triple-Resonance NMR Spectroscopy of Isotopically Enriched Proteins. *JOURNAL OF MAGNETIC RESONANCE* 1990, 89(4965):14.
216. Marion D, Driscoll PC, Kay LE, Wingfield PT, Bax A, Gronenborn AM, Clore GM: Overcoming the overlap problem in the assignment of ^1H NMR spectra of larger proteins by use of three-dimensional heteronuclear ^1H - ^{15}N Hartmann-Hahn-multiple quantum coherence and nuclear Overhauser-multiple quantum coherence spectroscopy: application to interleukin 1 beta. *Biochemistry* 1989, 28(15):6150-6156.
217. Marion D, Kay LE, Sparks SW, Torchia DA, Bax A: Three-dimensional heteronuclear NMR of nitrogen-15 labeled proteins. *Journal of the American Chemical Society* 1989, 111(4):1515-1517.
218. Bax A, Clore GM, Gronenborn AM: ^1H - ^1H correlation via isotropic mixing of ^{13}C magnetization, a new three-dimensional approach for assigning ^1H and ^{13}C spectra of ^{13}C -enriched proteins. *Journal of Magnetic Resonance (1969)* 1990, 88(2):425-431.
219. Wuthrich K: NMR of proteins and nucleic acids: Wiley; 1986.
220. Guntert P, Buchner L: Combined automated NOE assignment and structure calculation with CYANA. *J Biomol NMR* 2015, 62(4):453-471.
221. Keller R: The computer Aided Resonance Assignment Tutorial; 2004.
222. Lee W, Tonelli M, Markley JL: NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* 2015, 31(8):1325-1327.
223. Ramachandran GN, Ramakrishnan C, Sasisekharan V: Stereochemistry of polypeptide chain configurations. *J Mol Biol* 1963, 7:95-99.
224. Williams DH, Fleming I: Spectroscopic methods in organic chemistry. 1995.
225. Wishart DS, Bigam CG, Holm A, Hodges RS, Sykes BD: ^1H , ^{13}C and ^{15}N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. *J Biomol NMR* 1995, 5(1):67-81.
226. Wishart DS, Sykes BD, Richards FM: Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *Journal of molecular biology* 1991, 222(2):311-333.
227. Englander SW, Mayne L, Bai Y, Sosnick TR: Hydrogen exchange: the modern legacy of Linderstrom-Lang. *Protein Sci* 1997, 6(5):1101-1109.
228. Li R, Woodward C: The hydrogen exchange core and protein folding. *Protein Sci* 1999, 8(8):1571-1590.
229. Kim KS, Fuchs JA, Woodward CK: Hydrogen exchange identifies native-state motional domains important in protein folding. *Biochemistry* 1993, 32(37):9600-9608.
230. Ding D: Preparation and characterization of biomimetic materials inspired by jumbo squid sucker ring teeth. Nanyang Technological University 2016.
231. Caprioli RM, Farmer TB, Gile J: Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. *Analytical Chemistry* 1997, 69(23):4751-4760.

232. Spector T: Refinement of the Coomassie blue method of protein quantitation: A simple and linear spectrophotometric assay for ≤ 0.5 to 50 μg of protein. *Analytical Biochemistry* 1978, 86(1):142-146.
233. Vincentelli R, Canaan S, Campanacci V, Valencia C, Maurin D, Frassinetti F, Scappucini-Calvo L, Bourne Y, Cambillau C, Bignon C: High-throughput automated refolding screening of inclusion bodies. *Protein science* 2004, 13(10):2782-2792.
234. Singh SM, Panda AK: Solubilization and refolding of bacterial inclusion body proteins. *Journal of bioscience and bioengineering* 2005, 99(4):303-310.
235. Ptitsyn OB: Molten Globule and Protein Folding. In: *Advances in Protein Chemistry*. Edited by C.B. Anfinsen FMRJTE, David SE, vol. Volume 47: Academic Press; 1995: 83-229.
236. Fink AL: Molten Globule. In: *eLS*. John Wiley & Sons, Ltd; 2001.
237. Sunde M, Serpell LC, Bartlam M, Fraser PE, Pepys MB, Blake CC: Common core structure of amyloid fibrils by synchrotron X-ray diffraction. *Journal of molecular biology* 1997, 273(3):729-739.
238. Baker M: Making membrane proteins for structures: a trillion tiny tweaks. *Nat Meth* 2010, 7(6):429-434.
239. Wan W, Wille H, Stohr J, Kendall A, Bian W, McDonald M, Tiggelaar S, Watts JC, Prusiner SB, Stubbs G: Structural studies of truncated forms of the prion protein PrP. *Biophys J* 2015, 108(6):1548-1554.
240. Yee AA, Savchenko A, Ignachenko A, Lukin J, Xu X, Skarina T, Evdokimova E, Liu CS, Semesi A, Guido V *et al*: NMR and X-ray crystallography, complementary tools in structural proteomics of small proteins. *J Am Chem Soc* 2005, 127(47):16512-16517.
241. Slabinski L, Jaroszewski L, Rodrigues AP, Rychlewski L, Wilson IA, Lesley SA, Godzik A: The challenge of protein structure determination--lessons from structural genomics. *Protein Sci* 2007, 16(11):2472-2482.
242. Baker D, Sali A: Protein structure prediction and structural genomics. *Science* 2001, 294(5540):93-96.
243. Venselaar H, Krieger E, Vriend G: Homology modeling. *Structural Bioinformatics* 2009:715-732.
244. Venselaar H, Joosten RP, Vroiling B, Baakman CA, Hekkelman ML, Krieger E, Vriend G: Homology modelling and spectroscopy, a never-ending love story. *European Biophysics Journal* 2010, 39(4):551-563.
245. Sanchez R, Sali A: Advances in comparative protein-structure modelling. *Curr Opin Struct Biol* 1997, 7(2):206-214.
246. Al-Lazikani B, Jung J, Xiang Z, Honig B: Protein structure prediction. *Curr Opin Chem Biol* 2001, 5(1):51-56.
247. Petrey D, Honig B: Protein structure prediction: inroads to biology. *Mol Cell* 2005, 20(6):811-819.
248. Ginalski K: Comparative modeling for protein structure prediction. *Curr Opin Struct Biol* 2006, 16(2):172-177.

249. Vitkup D, Melamud E, Moult J, Sander C: Completeness in structural genomics. *Nat Struct Biol* 2001, 8(6):559-566.
250. Venclovas C, Zemla A, Fidelis K, Moult J: Assessment of progress over the CASP experiments. *Proteins* 2003, 53 Suppl 6:585-595.
251. Tress M, Ezkurdia I, Grana O, Lopez G, Valencia A: Assessment of predictions submitted for the CASP6 comparative modeling category. *Proteins* 2005, 61 Suppl 7:27-45.
252. Greenfield NJ: Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc* 2006, 1(6):2876-2890.
253. Jones DT: Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999, 292(2):195-202.
254. Buchan DWA, Minneci F, Nugent TCO, Bryson K, Jones DT: Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Research* 2013, 41(W1):W349-W357.
255. Drozdetskiy A, Cole C, Procter J, Barton GJ: JPred4: a protein secondary structure prediction server. *Nucleic Acids Research* 2015.
256. Garnier J, Robson B: The GOR method for predicting secondary structures in proteins. In: *Prediction of protein structure and the principles of protein conformation*. Springer; 1989: 417-465.
257. Russell RB, Barton GJ: The limits of protein secondary structure prediction accuracy from multiple sequence alignment. *J Mol Biol* 1993, 234(4):951-957.
258. Chivian D, Robertson T, Bonneau R, Baker D: Ab initio methods. *Structural Bioinformatics, Volume 44* 2003:547-557.
259. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA: A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society* 1995, 117(19):5179-5197.
260. Duan Y, Kollman PA: Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* 1998, 282(5389):740-744.
261. Neria E, Fischer S, Karplus M: Simulation of activation free energies in molecular systems. *The Journal of chemical physics* 1996, 105(5):1902-1921.
262. MacKerell Jr AD, Bashford D, Bellott M, Dunbrack Jr RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S: All-atom empirical potential for molecular modeling and dynamics studies of proteins†. *The Journal of Physical Chemistry B* 1998, 102(18):3586-3616.
263. Jorgensen WL, Maxwell DS, Tirado-Rives J: Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society* 1996, 118(45):11225-11236.
264. Hagler AT, Huler E, Lifson S: Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. *J Am Chem Soc* 1974, 96(17):5319-5327.

265. Glaves R, Baer M, Schreiner E, Stoll R, Marx D: Conformational dynamics of minimal elastin-like polypeptides: the role of proline revealed by molecular dynamics and nuclear magnetic resonance. *Chemphyschem* 2008, 9(18):2759-2765.
266. Kannan S, Zacharias M: Simulated annealing coupled replica exchange molecular dynamics--an efficient conformational sampling method. *J Struct Biol* 2009, 166(3):288-294.
267. Brooks CL, 3rd: Methodological advances in molecular dynamics simulations of biological systems. *Curr Opin Struct Biol* 1995, 5(2):211-215.
268. Ma B, Nussinov R: Molecular dynamics simulations of the unfolding of beta(2)-microglobulin and its variants. *Protein Eng* 2003, 16(8):561-575.
269. Sotomayor M, Schulten K: Single-molecule experiments in vitro and in silico. *Science* 2007, 316(5828):1144-1148.
270. Feig M, MacKerell AD, Brooks CL: Force Field Influence on the Observation of π -Helical Protein Structures in Molecular Dynamics Simulations. *The Journal of Physical Chemistry B* 2003, 107(12):2831-2836.
271. Rhee YM, Pande VS: Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys J* 2003, 84(2 Pt 1):775-786.
272. Miyashita N, Straub JE, Thirumalai D, Sugita Y: Transmembrane structures of amyloid precursor protein dimer predicted by replica-exchange molecular dynamics simulations. *J Am Chem Soc* 2009, 131(10):3438-3439.
273. Sugita Y, Okamoto Y: Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters* 1999, 314(1):141-151.
274. Tobias DJ, Mertz JE, Brooks CL, 3rd: Nanosecond time scale folding dynamics of a pentapeptide in water. *Biochemistry* 1991, 30(24):6054-6058.
275. Constantine KL, Friedrichs MS, Stouch TR: Extensive molecular dynamics simulations of a β -hairpin-forming peptide. *Biopolymers* 1996, 39(4):591-614.
276. Daura X, Mark AE, van Gunsteren WF: Peptide folding simulations: no solvent required? *Computer Physics Communications* 1999, 123(1):97-102.
277. Gnanakaran S, Nymeyer H, Portman J, Sanbonmatsu KY, Garcia AE: Peptide folding simulations. *Current opinion in structural biology* 2003, 13(2):168-174.
278. Ebrahimi D, Tokareva O, Rim NG, Wong JY, Kaplan DL, Buehler MJ: Silk-Its Mysteries, How It Is Made, and How It Is Used. *ACS Biomater Sci Eng* 2015, 1(10):864-876.
279. Kluge JA, Rabotyagova O, Leisk GG, Kaplan DL: Spider silks and their applications. *Trends Biotechnol* 2008, 26(5):244-251.
280. Heras B, Totsika M, Peters KM, Paxman JJ, Gee CL, Jarrott RJ, Perugini MA, Whitten AE, Schembri MA: The antigen 43 structure reveals a molecular Velcro-like mechanism of autotransporter-mediated bacterial clumping. *Proc Natl Acad Sci U S A* 2014, 111(1):457-462.
281. Wang X, Kluge JA, Leisk GG, Kaplan DL: Sonication-induced gelation of silk fibroin for cell encapsulation. *Biomaterials* 2008, 29(8):1054-1064.

282. Clément R: From secondary structure to electron beam lithography resist: going green with squid proteins. Université Paris-Saclay; 2014.
283. Georgiou G, Bowden G, Ho C, Prokop A, Dajpai A: Recombinant DNA Technology and Applications. In.: McGraw-Hill, New York; 1990.
284. Thomas JG, Baneyx F: Protein misfolding and inclusion body formation in recombinant *Escherichia coli* cells overexpressing heat-shock proteins. *Journal of Biological Chemistry* 1996, 271(19):11141-11147.
285. Khrapunov S: Circular dichroism spectroscopy has intrinsic limitations for protein secondary structure analysis. *Anal Biochem* 2009, 389(2):174-176.
286. Sreerama N, Woody RW: Structural composition of betaI- and betaII-proteins. *Protein Sci* 2003, 12(2):384-388.
287. Sreerama N, Woody RW: Computation and analysis of protein circular dichroism spectra. *Methods Enzymol* 2004, 383:318-351.
288. Selkoe DJ: Folding proteins in fatal ways. *Nature* 2003, 426(6968):900-904.
289. Orban G, Volgyi K, Juhasz G, Penke B, Kekesi KA, Kardos J, Czurko A: Different electrophysiological actions of 24- and 72-hour aggregated amyloid-beta oligomers on hippocampal field population spike in both anesthetized and awake rats. *Brain Res* 2010, 1354:227-235.
290. Micsonai A, Wien F, Kernya L, Lee YH, Goto Y, Refregiers M, Kardos J: Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proc Natl Acad Sci U S A* 2015, 112(24):E3095-3103.
291. Data mining of the PDB [http://www.douglas.co.uk/PDB_data.htm]
292. Fields BN, Knipe DM, Howley PM: Fields virology. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2013.
293. Koradi R, Billeter M, Wüthrich K: MOLMOL: A program for display and analysis of macromolecular structures. *Journal of Molecular Graphics* 1996, 14(1):51-55.
294. Krishna MM, Hoang L, Lin Y, Englander SW: Hydrogen exchange methods to study protein folding. *Methods* 2004, 34(1):51-64.
295. Bockmann A, Juy M, Bettler E, Emsley L, Galinier A, Penin F, Lesage A: Water-protein hydrogen exchange in the micro-crystalline protein crh as observed by solid state NMR spectroscopy. *J Biomol NMR* 2005, 32(3):195-207.
296. Van Melckebeke H, Schanda P, Gath J, Wasmer C, Verel R, Lange A, Meier BH, Bockmann A: Probing water accessibility in HET-s(218-289) amyloid fibrils by solid-state NMR. *J Mol Biol* 2011, 405(3):765-772.
297. Schanda P, Forge V, Brutscher B: Protein folding and unfolding studied at atomic resolution by fast two-dimensional NMR spectroscopy. *Proc Natl Acad Sci U S A* 2007, 104(27):11257-11262.
298. Wishart DS, Sykes BD: Chemical shifts as a tool for structure determination. *Methods Enzymol* 1994, 239:363-392.
299. Osapay K, Case DA: Analysis of proton chemical shifts in regular secondary structure of proteins. *J Biomol NMR* 1994, 4(2):215-230.

300. Jensen MR, Zweckstetter M, Huang JR, Blackledge M: Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy. *Chem Rev* 2014, 114(13):6632-6660.
301. Makrides SC: Strategies for achieving high-level expression of genes in *Escherichia coli*. *Microbiol Rev* 1996, 60(3):512-538.
302. Sorensen HP, Mortensen KK: Advanced genetic strategies for recombinant protein expression in *Escherichia coli*. *J Biotechnol* 2005, 115(2):113-128.
303. Walls D, Loughran ST: Tagging recombinant proteins to enhance solubility and aid purification. *Methods Mol Biol* 2011, 681:151-175.
304. Kapust RB, Waugh DS: *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci* 1999, 8(8):1668-1674.
305. Randall LL, Hardy SJ, Topping TB, Smith VF, Bruce JE, Smith RD: The interaction between the chaperone SecB and its ligands: evidence for multiple subsites for binding. *Protein Sci* 1998, 7(11):2384-2390.
306. Bach H, Mazor Y, Shaky S, Shoham-Lev A, Berdichevsky Y, Gutnick DL, Benhar I: *Escherichia coli* maltose-binding protein as a molecular chaperone for recombinant intracellular cytoplasmic single-chain antibodies. *J Mol Biol* 2001, 312(1):79-93.
307. Raran-Kurussi S, Waugh DS: The ability to enhance the solubility of its fusion partners is an intrinsic property of maltose-binding protein but their folding is either spontaneous or chaperone-mediated. *PLoS One* 2012, 7(11):e49589.
308. Di Guan C, Li P, Riggs PD, Inouye H: Vectors that facilitate the expression and purification of foreign peptides in *Escherichia coli* by fusion to maltose-binding protein. *Gene* 1988, 67(1):21-30.
309. Davis GD, Elisee C, Newham DM, Harrison RG: New fusion protein systems designed to give soluble expression in *Escherichia coli*. *Biotechnol Bioeng* 1999, 65(4):382-388.
310. Nallamsetty S, Waugh DS: Solubility-enhancing proteins MBP and NusA play a passive role in the folding of their fusion partners. *Protein Expr Purif* 2006, 45(1):175-182.
311. Ayton GS, Noid WG, Voth GA: Multiscale modeling of biomolecular systems: in serial and in parallel. *Curr Opin Struct Biol* 2007, 17(2):192-198.
312. Noid WG, Chu JW, Ayton GS, Krishna V, Izvekov S, Voth GA, Das A, Andersen HC: The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J Chem Phys* 2008, 128(24):244114.
313. Wang Y, Noid WG, Liu P, Voth GA: Effective force coarse-graining. *Phys Chem Chem Phys* 2009, 11(12):2002-2015.

Publications

1. Squid's Suckerin Proteins in Bits & Bytes; (Feb 16, 2016) **Akshita Kumar**, *et al. Biophysical Journal*. 110 (3):341a.
2. Mussel Adhesion is Dictated by Time-Regulated Secretion and Molecular Conformation of Mussel Adhesive Proteins; (Oct 28, 2015) Luigi Petrone, **Akshita Kumar**, *et al. Nature Communications*. 6:8737.
3. Multi-scale thermal stability of a hard thermoplastic protein-based material; (Sept 21 2015) Victoria Latza, Paul A Guerette, Dawei Ding, Shahrouz Amini, **Akshita Kumar**, *et al. Nature Communications*. 6:8313.
4. Biomimetic Production of Silk-Like Recombinant Squid Sucker Ring Teeth Proteins (Sep 8, 2014); Dawei Ding, Paul A. Guerette, Shawn Hoon, Kiat Whye Kong, Tobias Cornvik, Martina Nilsson, **Akshita Kumar**, *et al.*, *Biomacromolecules*. 15(9):3278-89.
5. Discovery and characterization of pseudocyclic cystine-knot α -amylase inhibitors with high resistance to heat and proteolytic degradation (Aug 8, 2014); Phuong Q. T. Nguyen, Shujing Wang, **Akshita Kumar**, *et al. FEBS Journal*. 281:4351–4366.
6. *Aspergillus fumigatus* harbouring the sole Y121F mutation shows decreased susceptibility to voriconazole but maintained susceptibility to itraconazole and posaconazole; (July 19, 2014); Julien Lescar, Isabelle Meyer, **Akshita Kumar**, *et al. Journal of Antimicrobial Chemotherapy*. 69(12):3244-7.

Poster Presentation

- Biophysical Society 60th Annual Meeting, Los Angeles, California, USA (2016)
- 2nd NTU-ICES (A*STAR) Joint Science Workshop, Singapore (2016)
- MGMS “Exploring Mechanisms in Biology: Theory and Experiment”, Singapore (2015)
- 1st Lübeck–Singapore Symposium, Singapore (2013)

Oral Presentation

- MGMS “Exploring Mechanisms in Biology: Theory and Experiment”, Singapore (2015)

Appendix A

Table A1: Suckerin proteins identified in *D. gigas* SRT.

Name	M.W. (kDa)	pI
Suckerin-1	4.8	5.5
Suckerin-2	9.6	7.1
Suckerin-3	12.4	7.2
Suckerin-4	12.9	6.7
Suckerin-5	14.6	9.4
Suckerin-6	15.7	7.3
Suckerin-7	16.4	9.2
Suckerin-8	16.9	8.3
Suckerin-9	18.6	8.7
Suckerin-10	22.5	7.4
Suckerin-11	24.7	8.9
Suckerin-12	22.9	8.0
Suckerin-13	23.0	8.9
Suckerin-14	23.6	8.1
Suckerin-15	25.0	8.7
Suckerin-16	30.4	8.3
Suckerin-17	35.5	7.3
Suckerin-18	37.5	8.0
Suckerin-19	39.3	8.3
Suckerin-20	49.7	7.9
Suckerin-21	44.4	7.3

>*D. gigas* Suckerin-1

MASTLFVLLAVIALASYASG ALTYGAGGWAL
 FVGGWAVFTGGVGGYGVFGGYGVFGGFGG FGGYGGGFGGGY*

>*D. gigas* Suckerin-2

MAVKIILLAVMALSNYAYGHFTGGVGGYGYGAGAGYGAGYGGGLGGGGVGSYGGVSGGYGGYGGYGGYGG
 YGGYGGHGGYGGYGGYGGYAGHGGYGGGLGGYGGHGGYGGYGSGGHHGFPYGGYGGYL*

>*D. gigas* Suckerin-3

MAHTQLIFVLLAAFGFACSEDEKQAF TAKATAF SNLLRSSWYT- - - - -
 FSSSYGIGLGYGGYGLGLGYGGYGAYGGYGGYGGYLFYGYGGILFGYFGGY
 PAVTSSSVTHHAF TY - - - - -
 PAVT-SSSVTHHAFYGVGVGYLGGLYGGYLGLH*

>*D. gigas* Suckerin-4

MASTKLIFVLLAAFGFACSEETKETTT KPFT- - - - -
 PAASNLLGSSWYTF TLGLGYGGYGLGLGYGGYGMGLGYGGYGLGLGYGGYGLGLGYGGYGGGLGGFYPGYHGG
 -VSTSSVTHHAFVYSV
 PHVTSSSVTHHAFYGLGAGYLGGLYGGYLGLH*

>*D. gigas* Suckerin-5

MTSKILLAVLSVSNYASAIL
 PATISVSQATHRGYR TYAGILKGFGLTYG STIYSKLTIRYGGLLGGYGYGLF SAGYGYGYGLF LAGYGGLYGGYGG
 YGGWGYFVGAVGAYGGLLGGYGYGLF LGGYGGFFGGYGA FYAGYGYGVLLGF VGF PATISVSQTTHH*

Figure A1: Protein sequences of respective suckerins found in *D. gigas* SRT. The M1 domain is highlighted in red and M2 domain is highlighted in cyan. The Pro residues that flank the two domains are highlighted in green.

[illegible]

MASTIFTLLAVIALSSYAAAIL
PATTSVSRITTHHAFVY----GGYGLLHGGLYGGYGLG
--AATVSHITTHHAPYGYGGLYGGYGLAHGGLYGGYGLGAYGLGY
PAAAAVSHITTHHAPLGFGGGLYGGYGLG
---AVSHITTHHAPLGFGGVLGGYGLG
---AVSHITTHHAPLGFGGLYGGYGLG
---AVSHITTHHAPLGFGGVVGGYGLG
---AVSHITTHHAPLGFGGLYGGYGLG
---AVSHITTHHAPLGFGGGLYGGY
---AVSHITTHAPLGYGGLAGYAWLH*

MAAAVLFIYSHIAALSCQTEAIL -
 PGTSVKITSSVHHIVV
 PSVRHTVSTVSHAGAHFAYGNLYGGLYGGPYGV
 PGATVHTVSHGVHPS
 PLITTSVHHHTHGGGLIGGLYGAGYFLGFGGYGLHPAGLGYGLGGLYGGLYGY
 PAAASVTHGFHPAGFGLGLGGFYGGSYGY
 PGATVSHHTHGLHPGGFGLGLGGFYGGAYGF
 PAASSVSHVTHGVHPAGLFGGVYGTGYGI
 PAGTTVSHHTHGVHHL
 PAASTVTHHTHGVHHPMGVSYGNVFLH*

MAAIFTL~~LLAVLAISNYASAIL~~-----
 PAATSVSR~~THRT~~GYGYGGLLGGYGLHY-----
 PATTAVSHT~~THHAPAA~~LG-VYGGYGLGAYGFGY-----
 PGAATVSH~~THHAPY~~GYGGLYGGYGLAH-GGLYGGY-GLGAYGYGY-----
 P-AAAVSHV~~THHAPY~~GYGL--GGY-----GGLYGGLYGGHGLLHGGLYGGYGLGAYGYGY-----
 P-AAAVSHV~~THHAPY~~GYGLAGYGGLHGGYGFLHGGLYGGHGLLHGGLYGGYGLGAYGYGY-----
 P-AAATVSH~~PTHHAPY~~GYGGLYGGYGH~~PGY~~GGLY-GGYGFAH--GGLYGGYGFFHGGLYGGYGF-----
 -----PHGGLYGGYGLGGYGLGGYGLGGYGGWFH*

MAATYFIIAAFETVTGLVSAQE -
 PFGPGPEYIGIDIGFGNLEAGAFTHGFGPFGAFFFFFPAPYPGITPHFISWSFHVSI
 PTHVTHTSGVSV
 PHHVSTHVGGS
 - TTKTQTVTHH
 PGSHVTVVHHPAT
 PMATHVQHVNYGSGFPVGAPELLPMGYGPGYPVEFYFGMGHPMSW -
 PTFGGVGGTTHVTHT
 PVGGVTHHVGTGHTTTTKTQTVTHH
 PSSQVTHVTHHPA
 PMVTSHVNVGAGFPGGFGYGHGFLAP
 PGITHQTFSHGFGFGFGGMYDQFGVGYGFDTGMPFFYIHDS
 MGMQN VHTVHHHH
 PGGVTTHITYPHPLTAYPMGNPPYPYGSTTTVTKHTKTVTHKAGSGKKN*

157

>*D. gigas* Suckerin-18
MAASILTLAVIALSSCTLAVV -----
PAATVSRTTHHAPYGYGGVVGGLYGGYGLGGYGLAGYGGGLHGGLYGGYGLGAYGFGY
P-AATVSHHTTHHAPYGFGLYGGHGLLHGGLYGGYGLHHGGLYGGY-----GLGAYGFGY
PVATAVSHVTHHAPYGYGLAGYGGF--H----GGYGLLHGGLYGGYGLHHGGLYGGYGLGAYGFGY
P-AATVSHHTTHHAPYGFGLYGGHGLLHGGLYGGYGLHHGGLYGGY-----GLGAYGFGY
P-AATVSHHTTHHAPYGFGLYGGYGLHHGGLYGGYGLLHGGLYGGYGLHHGGLYGGYGLGAYGFGY
PGAATVSHHTTHHAPYGFGLYGGYGLHHGGLYGGYGLLHGGLYGGYGL-----
---GAVSHHTTHHAPYGY-GL-GGY-----GGLYGGW-LH*

> *D. gigas* Suckerin-19
MAAIFTLAVLAISNYASAIL -----
PAATSVSRTHRTGYGYGGLLGGYGLHY-----
PATTAVSHHTTHHAPAALGVY--GGY-----GLGAYGFGY
PGAATVSHHTTHHAPYGY-----GGLYGGYGLAHGGLYGGYGLGAYGYGY
PAATAVSHHTTHHAPYGYGLGGYGGYGGYGLAHGGLYGGYGLGGYGLHY
PAATAVSHHTTHHAPLGYGLAGYGGLY-----GGLYGGYGLGGYGL---
--AAVSHHTTHHAPLGYGLGGYGGLY-----GGLYGGYGLGGYGLHY
PAATAVSHHTTHHAPLGYGLGGYGGLY-----GGLYGGYGLGGYGL---
--AAVSHHTTHHAPLGF-----GGLYGGYGVGGYGYGY
PAAATVSHHTTHHAPYGY-----GGLYGGYGLAHGGLYGGYGFG-----
---AVSHHTTHHAPLGF-----GGLYGGYGLAHGGLYGGYGLG-----
---AVSHHTTHHAPYGF-----GGLYGGYGLLH*

> *D. gigas* Suckerin-20
MATTIFAILAVIALSNYASAI-----
PAATSVSRTHHAGYGYGGFLGGYGLGAYFGS-----
P-AATVSHHTTHHAPYGY-----GGLYGGYGGYGGHGLLH-----GGLYGGYGHGGLYGGYGLGAYGFGY
PAAAVSHHTTHHAPLGFGLAGY-----GGLYGGYGLGGYGLG-----LHGGLYGGY-----GLGAYGFGY
PAATAVSHHTTHHAPYGYGLGGYGGYGGYGLAGYGGYGGHGLLHGGLYGGY-----GLGAYGFGY
PAAATVSHHTTHHAPYGY-----GGLYGGYGLAH-----GGLYAHGGLYGGYGL-----GGYGLHY
PAATAVSHHTTHHAPLGYGLAGYGGYGGYGGYGL-----GGY-----GGYGLHY
--AAVSHHTTHHAPLGYGLGGYGGYGLYGGYGL-----GGY-----GGYGLHY
PAATAVSHHTTHHAPLGYGLGGYGGYGGYGGYGL-----GGY-----GGYGLHY
---AAVSHHTTHHAPLGF-----GGLYGGYGV-----GGYGGY-----GGYGGY
PAAATVSHHTTHHAPYGY-----GGLYGGYGLAH-----GGLYGGYGFG-----
---AVSHHTTHHAPLGF-----GGLYGGYGLAH-----GGLYGGYGLG-----
---AVSHHTTHHAPYGF-----GGLYGGYGLLH*

> *D. gigas* Suckerin-21
MAASILTLAVIALSSCTLAVV -----
PAATVSRTTHHAPYGYGGLL-----GGYGGLH-----GGYGLGGYGLAGY--GGLHGGLYGGYGLGAYGFGY
P-AATVSHHTTHHAPYGF-----GGLYGGHGLLHGGLYGGYGLH-----HGGLYGGYGLGAYGFGY
PAATAVSHVTHHAPYGYGLAGYGGYGGYGLAHGGLYGGY-----GLAGY-----GGLYGGYGLGAYGFGY
PAATAVSHHTTHHAPYGYGL-----GGYGGYGGLYGGYGLAGYGGYGGHGLLHGGLYGGYGLGAYGFGY
PAAATVSHHTTHHAPYGY-----GGLYGGYGLAHGGLYGDMDLAHTDTLPSCYSCQ
P-----HT-HHALLRIWSGRICGLYEICLA--HGGLYGYGLGGYGLHY-----
PAATAVSHHTTHHAPLGYGLAGY-----GGLYGGYGGYGLGGYGL-----
--AAVSHHTTHHAPLGYGL-----GGYGGYGGYGGYGLGGYGLHY
PAATAVSHHTTHHAPLGYGL-----GGYGGYGGYGGYGLGGYGL-----
---AAVSHHTTHHAPLGF-----GGLYGGYGVGGYGYGY-----
PAAATVSHHTTHHAPYGY-----GGLYGGYGLAHGGLYGGYGFG-----
---AVSHHTTHHAPLGF-----GGLYGGYGLAHGGLYGGYGLG-----
---AVSHHTTHHAPYGF-----GGLYGGYGLLH*

Figure A1 (cont.): Protein sequences of respective suckerins found in *D. gigas* SRT. The M1 domain is highlighted in red and M2 domain is highlighted in cyan. The Pro residues that flank the two domains are highlighted in green.

Table A2: Suckerin proteins identified in *S. lessoniana* SRT.

Name	M.W. (kDa)	pI
Suckerin-1	11.9	9.8
Suckerin-2	12.3	8.2
Suckerin-3	12.8	8.2
Suckerin-4	18.0	8.2
Suckerin-5	21.9	7.5
Suckerin-6	31.5	9.2
Suckerin-7	33.2	7.2
Suckerin-8	51.2	9.2

> *S. lessoniana* Suckerin-1

MAAKLLTLLAVIALSNYAYALL -----
-----PGLMGGYGY-----
PAATTYRRTI ---LNGYGGLYG-----GLGY--HY
PAATAVSHITTHHAPYGYGGLYGWGY-----
PAAASVS-IVHR-PVGY----GGWGLGGYGLGGYGLGYGLHY
PAATV-----GYS-----GL--GL-GYGSYVVL*

> *S. lessoniana* Suckerin-2

MASYRLIFVYLLAAFLAYGEQTKK -----
TITISNTGAQNYYNFWYYTYNYGWGNFWGLGYGYGMLFNYGTH
---TVTHHAPQYFYYGF-----
PQVSTSHVTHHAPIV-----
PHVTSNTVTHHAPYTFGNLGLYGGLFGYH*

> *S. lessoniana* Suckerin-3

MTTSSLLALMTVIGLVTCCAGYPSWG -----
STYSQVSTHHAPVGWGGGLGGLYGYSLGGLGYGGVYGGGVHVAQS
PVRVHGYSVGT-----
PHVQSYGVNYPT-----
PTVTHHHNTWH-GGYGGFGLAGFVGH-----
SVQTVSQGVHTPGWGLGYGLSGYGY*

> *S. lessoniana* Suckerin-4

MTATVLFLLFGIAALTCOTEAAITG -----
-----TSSVKTVVHP-----
PTVVHPVHTVTHTDHHLVWNTGVYGI-----
PSAGSTSVKTVTHGIHHTGGISVTS-----
PGGATVTHTTTHGISH-PFGLGHGL-GGLYGGVYGLPM-----
PGATTVSHSTHGV---PYSFGY---GGLGYGGIGYGGGLGYGGLGYGLGGAYGLPYNTAISHSYGFQSPSVYGFQFGGYYGFH*

> *S. lessoniana* Suckerin-5

MVSTVLIIMSMIAALSCQSEAL -----
SVGTSVKSVHHSVHH-----
PISAVGQTVKTVTHAVPQIYFPGGLPFGDAY-----GGLYGALHGGVYGV
PAATSVQTVSHGLH--PTV-----
PVGSTSVSHTTTHGIHHPVTY GGLGVGGLGFGGLGYGGVGGGLGGLGYGGVYGLHY--
PGAAFYH-----PAG-----LGYGLGGLYGGLYGLHL--
PAATSVSHTTTHGVHH-PALGYGLGLYGAHY-----
PAASSVTHTTTHAVPH-PGFLNLYGVFH*

Figure A2: Protein sequences of respective suckerins found in *S. lessoniana* SRT. The M1 domain is highlighted in red and M2 domain is highlighted in cyan. The Pro residues that flank the two domains are highlighted in green.

Table A3: Suckerin proteins identified in *S.esculanta* SRT

Name	M.W. (kDa)	pI
Suckerin-1	15.1	6.8
Suckerin-2	17.5	8.6
Suckerin-3	18.4	8.2
Suckerin-4	26.1	8.1
Suckerin-5	28.7	9.1
Suckerin-6	32.5	9.3
Suckerin-7	37.8	9.6
Suckerin-8	41.3	7.3
Suckerin-9	47.8	9.5

>*S. esculenta* Suckerin-1

MAATATLLVLM SVIGLTSC - - - - -
 - - - - - YPGYFGGYGGAL - - - - -
 PAASTYSQVSHHHGWGGLGGLYGYGVHL
 PVSTTSHVSHGAIYGGWGYPGLYGGVH -
 - - - - - VSQSPVRYHGYSVGT - - - - -
 PHVQSYGVHY - - - - -
 PTTSVTTHSHGDWGLGEYGYGA - - - - -
 PLAHSVQTVSHGFGHNLHLELGYGLHSFGHF*

>*S. esculenta* Suckerin-2

MAIKTFETTATLIAVLGLIACSEAIF - - - - -
 PA - - - - - YYGSYY - - - - -
 PGYRRSVSSVTHHPVGYGGFGYGSYGGFYGGY - - - - -
 PAASTVTHTAHGVP - - YSGVGYGWGLAGLHGGYGWHYPAGYGGWGLNGVYGFHH
 PVGASVS - TVSHHP - - - - - GFYGGYGLGF - - GYGGWGY - - - - -
 PAATY - SH - VAHQPFGLGYGYGAWGY - - - - -
 PGASYTSH - VAHHPFGFY*

>*S. esculenta* Suckerin-3

MMANVLEFILSLIAALSCQLAAALAAIK - - - - -
 - - - - - TSSTAVHHSP - - - - -
 PTVAHTVHSLAHAVPHFYTFGVWPGYDFYGGLYGYF
 PTTSVKTVSHGLH - - - - -
 PAVGTTSVSHTSQGIHQELPY - - - - - GGFHGGFY - -
 PGAAFHY - - - - - PLGLGYGFGGLYGGFQ - GFHN -
 PGATTVSHTTQGVHHPLGLGYGFL - - - - -
 PYGAAHV - - - - -
 PTTSVSHTTQGLSHPAFGWNYGNVFGFLH*

Figure A3: Protein sequences of respective suckerins found in *S. lessoniana* SRT. The M1 domain is highlighted in red and M2 domain is highlighted in cyan. The Pro residues that flank the two domains are highlighted in green.

MAYKIITFLAVIALSNYHA - - - - -
 - - - - - IIPGLLGHHGY - - - - -
 SGATTYKQTHHA-PYGYGSVYGGWGL
 PAAGVSTYSHGLHHPVGWGLGGYGFHY
 PAATSYSHTHHA-PYGYGSVYGGWGL
 PAAGVSTYSHGLHHPVGWGLGGYGFHY
 PAATSYSHTHHA-PYGYGSVYGGWGL
 PAAGVSTYSHGLHHPVGWGVGGYGFHY
 PVATYSHTTHHA-PYGYGSVYGGWGL
 PAAGVSTYSHGLHHPVGLGFGGYG LHY
 PAATSYSH*

MATTITLFAFIAVLQMSCSVSG
 ----- LFNGAWGSLGTSHYGYRY
PSFAKSVSSISH ----- **GAHY** -----
 PAYTRWRYS ----- P-GMWGHGLSGWSRYSSYY
PASSTVSHLAE ----- **GAH** ----- Y -
 PIEAVAYNTH ----- Y -
 PVGQTISSASHGAYQPFAWGYSHGSGYGLNNL ---
 - ATSSASHSTHGIHQPIAWRGLYGLHY -----
 PAITTVYSQSNHGVVNPISWGYSHGGLYGGYGfNY
 PAATSVSHAIHGNHQPVAWGAHYGLHY -----
 PAVTNLSQASHGIVSPISWGSYHGGLYGGYGfHY
 PTSTLQHTSGWHSFGAHSPPFNOCYGFY*

MAAGHFIVSALLAVIALSNCANGVL
 ---PGVAGGYGTSYG
 TFSHGSHFPGSGWWGSLSGFGHG
 HTVSQVSHGFPVGYGWGYGNLGGYGGYGLGGISGHYGGYGLGGISGHYGGYGLGGLYGHY
 ---GYLGLGLHGHFGGYGLGGYGGYGLPGMYGAYGGYGHGGVSHYGIIGGR--
 TINQVTRRLHPLGYGIPY
 PVGHAVHRYTQPIAHYGTISRRAFYQEVRVVS
 PVTHRYIQMVTVSKPYVVPREYEMHVRQYQVPVPRYYVTMAQY
 PVVIFRVAIAESHQ
 PVTHISHVSRDYHQPIIGMHY
 PMGOTIESVSHHEPFVYGATHYGGGY*

MTAKYVIVAFLAGIVSGQLPFHAF -
 PAYGIGGIGSLGYAGGAGAFSHGFG QPFGFSYYGISYSLFL -
 PVSAAGAAGTTHTVSHTSAAVPH -
 PAASHVNKAVTVAHIF -
 PSTHVTTVHH -
 PVSAVTTHEQHVNHGAFHEVVPYSFFYGETLSQFGYQYRTSFVKHQFI -
 PATSHSTVVKSTSQTSSIPGSYLDLVKKL -
 - - - - -SSSAANVAQESTIYSYQNYK -
 PVASVVTKTQHLPVGYPYNL -
 FNIATRHHQTVSHVGPFGVAHLPSLF -
 PGLTSASTIKYQQTAYPLGLGSYGLPGTFNSPFGAFGTFFYVNAGF -
 PGGAVSSSAVHTVHHPGAFALYSAFQNYPVGTSSPVSFYSV -
 PSSHLKITKSKOKSSSNKK*

162

>*S. esculenta* Suckerin-8

MVASVFATALYLVLISSQAIF
PGTSTFAS--YY-----
PYYRSTVSHGVNFPGFYGLGYGVH-----
PVGFYGAYDSVS**PVSHGFHGF**GGWGYGLSGLFGGYGLHY
PAATSVSHTTTHGY**PYPVL**HEGWSLG SVYGGFGWHS-----
PTGAAVSSVSHY**PSGWG**FSGLYSGHHY-----
PTATSFSQTRY**HPVDW**GFGGLYDGFHY-----
PAATAVSHITH**HPVGW**GFGGLYGLHY-----
PSTTAVSHMTH**HPVGW**GFGDLYGRLPY-----
PSATSVSH**THLPLGW**GFGGLYGGFHY-----
PSATAVRHM**THLPVGW**GFGGHYSAFLY-----
PAATSFSK**TIHHPVGW**EFRGIYAGLHY-----
PTITTVSHMTH**HPVGW**GFGGAYGMHY-----
PVGAAVSIFN**HPVGY**GGFGLGFGGLYGS**PF**FGFY-----
PAAVSNTHIT**TYGFHY**PEVYGGFGHGIFY*

>*S. esculenta* Suckerin-9

MGLMQFAVTLIIVLGISCSAKG
-----LYNGSRGGLGNLHYGYRY-----
PYYKGFANTVSHGSQY**PSYSGW**GSDISKG-----
-L**SLANYGRQYSRAHY**P--FVNYGLGGNSHHY
PYSTISQTINR**PTRW**IFGGIYRGSGGQVYQS
PVRYHGYSVGN-----
PTVLTYGFY**PRKGIF**FHNFQHGGHSQSGLGEFISSYR---
PGASVNAASHGISHL-----
PIHSSWNFGF-----
PAVQSVSAVQ**PTID**-YRSGSIGDFYRGYSTNY-----
PLGAAASTLSYDASQ**FIGLGF**GNSYSRYSFHY-----
PKSVSHTTQ**RPP**---YLGGWSGYGQGYGIHY-----
PIGASLSTVSQGIQ**QPI**IGYRTWGYDHGAHY-----
PVVQSLNTVSQGIQ**HP**IYGGWGLTRGHGIDY-----
PSGYSSRTVSHGFQ**TPVAF**GGLVGGFGL**PGQV**GGSELG-
-----GLDGFGLGGQVGRSSFGLVIRGY
-----GLGGQVGGSNFWG**PLRGY**GLGGQ
-----VSGYGLGGLSGGYGIEI
PVNSLSHH**PVTP**---GSGF*

Figure A3 (cont.): Protein sequences of respective suckerins found in *S. lessoniana* SRT. The M1 domain is highlighted in red and M2 domain is highlighted in cyan. The Pro residues that flank the two domains are highlighted in green.

Appendix B

Figure B1. 1D NMR spectrum of sepia monomer in 20 mM sodium acetate, pH 5.0 (green), water (blue) and 5% acetic acid (red).

Figure B2. ^{15}N -HSQC spectrum for sepio monomer in 20 mM sodium acetate pH 5.0 at 298 K