Biophysical characterization of chitin-binding beak proteins in dosidicus gigas

Heymann, Dario Benjamin Jonas

2020

Heymann, D. B. J. (2020). Biophysical characterization of chitin-binding beak proteins in dosidicus gigas. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/138896>

<https://doi.org/10.32657/10356/138896>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 09:05:21 SGT

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**Biophysical Characterization of Chitin-binding Beak Proteins
in *Dosidicus Gigas***

DARIO BENJAMIN JONAS HEYMAN

SCHOOL OF MATERIALS SCIENCE AND ENGINEERING

2019

**Biophysical Characterization of Chitin-binding Beak Proteins
In Dosidicus Gigas**

DARIO BENJAMIN JONAS HEYMANN

SCHOOL OF MATERIALS SCIENCE AND ENGINEERING

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

2019

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

1 December 2019
.....
Date

.....
Dario Heymann

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

1 December 2019
.....
Date

.....
Ali Miserez

Authorship Attribution Statement

This thesis contains material from one paper submitted to JACS in which I am listed as first author.

Chapter 4 is submitted for publication to the *Journal of the American Chemical Society* as:

Heymann, Dario; Mohanram, Harini; Kumar, Akshita; Verma, Chandra; Lescar, Julien; Miserez, Ali. Structure of a Consensus Chitin-Binding Domain Revealed by Solution NMR.

The contribution of the co-authors is as follows:

- M.A. and J.L. provided the initial project directions and edited the manuscript draft.
- I prepared the initial manuscript draft. The manuscript draft was revised and edited by M.H.
- I designed the study and performed all laboratory work at the Biological & biomimetic Material Laboratory at Nanyang Technological University.
- M.H. supported on executing and analysing the NMR experiment
- A.K. and C.V. performed the molecular dynamics studies at the Bioinformatics Institute (BII) at A*STAR.

1 December 2019
.....
Date

.....
Dario Heymann

Abstract

Carbohydrate-binding proteins (CBPs) are a versatile group of proteins occurring in almost every organism on earth. These proteins are imperative to various functions, ranging from enzymatic carbohydrate degradation and serve as templating scaffolds with load-bearing properties in the exoskeleton of crustaceans or insects. A large amount of data on how lectins bind highly specifically to carbohydrates exist, however, the actual binding mechanisms involved is unknown in most biological materials where they are found.

Primary focus of the subsequent study is the determination of the structure of CBD- γ , a chitin-binding domain found in the Chitin-binding Beak Protein-3 (DgCBP-3) using NMR spectroscopy. DgCBP-3 is one of four unique chitin-binding proteins (DgCBPs) found in the beak structure of *Dosidicus gigas* and is believed to interact with chitin to form a net-like structure in which a second class of proteins, namely Histidine-binding Beak Proteins (DgHBPs) infiltrate to assemble the beak. Previous research has identified one chitin-binding motif, the extended Rebers and Riddiford" consensus (R&R), to play a critical step in the biogenesis and hardening process in exoskeletons of arthropods. However, till date no experimental tertiary structure of a protein containing an R&R Consensus has been reported and the mechanism how the R&R Consensus binds chitin remains unclear.

The following study includes the structural determination of CBD- γ , which appears to be the first experimentally solved three-dimensional protein structure of a CBP containing the R&R Consensus. Further, solution state NMR spectroscopy has been used to elucidate the molecular interactions between CBD- γ and the soluble chitin derivative pentaacetyl-chitopentaose (PCP). The study shows that PCP triggers a folding of CBD- γ upon its interaction.

The results of this study can be used as a template to understand in more details the role of the R&R motif and may inform other research conducted on other structural chitin-binding domains across species, including other cephalopods, crustaceans, and insects.

Lay Summary

Proteins are the versatile workhorses found across all organisms on earth which are essential in all known functional, regulatory and structural processes. Proteins are assembled from amino acids, organic compounds, whose function is described by its tertiary, three-dimensional structure that cannot easily be interpreted from secondary patterns. The three-dimensional structure of a protein provides useful information, such as how binding to other macromolecules occurs or to describe its unique function.

This study focusses on the determination of the three-dimensional structure of a functional portion of a protein found in the beak assembly, namely CBD- γ , of the Humboldt Squid, a large predator found in the Eastern Pacific Ocean. The squid's beak forms a wholly organic biogradient that exhibits a stiffness that exceeds that of steel and is therefore of high interest in the biomimetic research community. CBD- γ shares the pattern of a carbohydrate-binding domains, a group of functional domains that are widely found to be involved in the degradation of carbohydrates but that also play a function in the integumentary system of insects.

This study determines the three-dimensional structure of CBD- γ which not only represents a first-of-its-kind structure of structural carbohydrate-binding domain but also introduces a novel binding mechanism. Though CBD- γ is thought to be unique, it interestingly shares to a certain extent similarity to other carbohydrate-binding domains found in insect's integumentary system and in rubber tree.

The study further assesses the impact chitin derivatives have on the tertiary structure of CBD- γ and maps the location where carbohydrate and CBD- γ are interacting to provide further insights on the underlying molecular mechanism that plays a major role in the assembly of the Humboldt Squid's beak.

Ultimately, insights gathered from the study on the beak's assembly can be used to design wholly organic, graded biomaterials which share the common advantage over artificial materials to be biosimilar and degradable. Such materials can be used in biomedical applications or in renewable architecture.

Acknowledgements

Thinking about and consciously expressing gratitude and appreciation in written words speak louder than just spoken. Keeping this in mind, I would like to express my deepest gratitude and appreciation to the following individuals, who have supported me during the course of my studies and during the writing of my thesis.

Up amongst all, I would like to thank my two principle investigators:

Ali, for giving me the opportunity to pursue my studies in his laboratory, and his guidance and tranquillity for the past years and let me apply my existing knowledge to a new field of research, I would have never thought to explore. I am further thankful about the freedom that you have provided me to shape my own future.

Julien, for giving me a space in his laboratory where I could conduct most of my research, even though I have never been a proper member of your laboratory. I am also thankful about the few but conclusive discussions we had which always supported me to not lose sight of my goal.

Despite them, I would like to thank two additional people in particular, who have supported me along the way.

Firstly Harini, who helped me especially during my final year with all NMR related experiments and data interpretation. I know we had many ups and downs along the way from reacquiring whole new sets data to reassessing sets that have already been analysed.

Secondly Abbas, who helped me especially during the first half of my research while all X-ray experiments have failed, from setting up screens to measuring none-existing crystals overseas.

I would also like to thank all members of Ali's and Julien's labs who have supported me along the way for their friendship, providing an enjoyable working environment and endless discussions.

I would also like to extend my gratitude to Shu Hui for her help with SEM and continuous positivity and Akshita for her help conducting molecular dynamics predictions for me.

I would further like to acknowledge School of Materials Science & Engineering, NTU and the Ministry of Education for funding my research project and giving me the opportunity to pursue my degree in Singapore.

I would further like to thank my parents, family and friends for always supporting me whenever required. Without encouraging me I wouldn't be where I am standing today. Thank you also for always giving me the freedom to always follow the path which I believe is best for me.

This is the real secret of life — to be completely engaged with what
you are doing in the here and now.

And instead of calling it work, realize it is play.

(Alan Watts)

Table of Contents

Statement of Originality	v
Supervisor Declaration Statement.....	vii
Authorship Attribution Statement	ix
Abstract	xi
Lay Summary	xiii
Acknowledgements.....	xv
Table of Contents	xix
Table Captions.....	xxv
Figure Captions	xxvi
Abbreviations	xxxv
Chapter 1 Introduction.....	1
1.1 Background and Hypotheses	2
1.1.1 Carbohydrate Binding in Nature.....	2
1.1.2 Carbohydrate Binding in <i>Dosidicus Gigas</i>	3
1.1.3 Hypotheses.....	3
1.2 Objectives and Scope.....	4
1.3 Dissertation Overview	6
1.4 Findings and Outcomes/Originality.....	7
1.5 References.....	7

Chapter 2 Literature Review	11
2.1 Biomimicry in Material Science	12
2.2 Proteins in Material Science	13
2.2.1 The structure and property of material proteins	14
2.3 Dosidicus Gigas and its beak: a heroic biomaterial	16
2.3.1 Dosidicus Gigas	16
2.3.2 D. gigas' un-mineralised beak and its composition	18
2.4 Chitin	19
2.4.1 General Overview	19
2.4.2 Common Applications	20
2.4.3 Structural characterisation	21
2.5 Carbohydrate-binding domains	22
2.5.1 Rebers and Riddiford Consensus	23
2.5.2 Chitin-Binding Beak Proteins	24
2.6 NMR as a Tool to Provide Structural and Biochemical Insights of the	
Structure of Carbohydrate Binding proteins	25
2.6.1 Chemical Shift Perturbation Analysis	25
2.6.2 Amide Proton Temperature Dependence in Proteins	27
2.6.3 STD-NMR as Tool to Study the Protein-Carbohydrate Interactions.....	27
2.7 MD Simulations and Docking Analysis of CBD- γ and PCP	29
2.8 Scope of the PhD Project in Context of Previous Research	31
2.9 References.....	33

Chapter 3 Experimental Methodology.....	45
3.1 Rational.....	46
3.2 Construct Design, Cloning & Analytical Small-Scale Expression of DgCBPs	46
3.3 Construct Optimisation for DgCBP-3	47
3.4 Expression of DgCBP-3	48
3.5 Protein purification of GST-PreSci-His6-TEV- CBD- γ	48
3.6 Expression of DgCBP-3 Constructs for Nitrogen-15 NMR Spectroscopy	50
3.7 Analysis of DgCBP-3 Constructs	51
3.8 Circular Dichroism Spectroscopy.....	52
3.9 Analytical FPLC	53
3.10 Fluorescence-based Melting Temperature Determination.....	54
3.11 Structural Analysis of CBD- γ	54
3.11.1 Protein X-ray Crystallography	55
3.11.2 Nuclear Magnetic Resonance Spectroscopy	56
3.11.2.1 1-Dimensional Proton NMR.....	57
3.11.2.2 2-Dimensional Proton NMR.....	57
3.11.2.3 Heteronuclear Correlation Spectroscopy	58
3.11.2.4 2-Dimensional ^1H - ^{15}N HSQC.....	59
3.11.2.5 3-Dimensional ^1H - ^{15}N NOESY-HSQC & ^1H - ^{15}N TOCSY-HSQC.....	59
3.11.2.6 NMR Assignment	59
3.11.2.7 Chemical Shift Perturbation.....	60
3.12 Saturation Transfer Difference NMR	61
3.13 Predictive Ligand Docking Using Schrödinger's Glide	63
3.14 Amide Proton Temperature Coefficient	64
3.15 References.....	65

Chapter 4	Results	71
4.1	Introduction.....	71
4.2	Protein Construct Design of DgCBPs and Small-Scale Purification	72
4.3	Purification of CBD- γ	73
4.4	Primary Sequence Analysis of CBD- γ	77
4.5	Identification of Secondary Structure Elements Integrity of CBD- γ Using Circular Dichroism Spectroscopy	79
4.6	Qualitative Binding Analysis of CBD- γ to Chitin Magnetic Beads	80
4.7	Ligand-Receptor Binding Affinity Determination Using STD NMR.....	81
4.8	One-Dimensional Homonuclear NMR	85
4.9	Protein Stability Assessment Using Two-Dimensional ^1H NMR	87
4.10	Two-Dimensional Homonuclear NMR	89
4.11	Solution Three-Dimensional Structure of CBD- γ	94
4.12	Determination of the Structure of Free CBD- γ & in Complex with PCP	95
4.13	Chemical Shift Perturbation of CBD- γ	101
4.14	Conclusion	104
4.15	References.....	105

Chapter 5	Discussion and Future Work	109
5.1	General Discussion	110
5.2	Expression and Purification of Chitin Binding Beak Proteins	112
5.3	Structural Characterisation of CBD- γ and the Impact of PCP	113
5.4	CBD- γ - a First Protein Structure of a Structural CBM	116
5.5	CBD- γ in Context of the Humboldt's Beak	118
5.6	Future Work	119
5.8	References	120
Chapter 6	Appendix	123
6.1	Construct Design	123
6.2	2-Dimensional Homonuclear NMR	125
6.3	Analytical FPLC	127

Table Captions

Table 2-1 | Amino acid content of DgCBPs and DgHBPs

Table 3-1 | Protein Purification Buffers

Table 4-1 | Forward and Reverse Primers of the CBD- γ Insert

Table 4-2 | K_d of acetyl-chitooligomers determined using STD-NMR spectroscopy

Table 4-3 | NOE restrains of CBD- γ in Absence and in Complex with Pentaacetyl-chitopentaose.

Figure Captions

Figure 2-1 | Proteins in Material Science. (A) Schematic representation of the Collagen assembly (Adapted from: [19]) (B) Spider-silk fibre assembly is being formed by shear forces that cause the C-terminus to expose its hydrophobic surfaces [20] (C) well-known cytoskeletal filament proteins i) lamin, ii) F-actin, iii) microtubule that provide mechanical support in cells (Adapted from: [21-23])

Figure 2-2 | Anatomy and Geographical Distribution of *D. gigas*. (A) Zoological representation of the Humboldt Squid with flaring tentacles and its beak by Alcide D'Orbigny (1835) [29], (B) Geographical distribution map of *D. gigas* in the Eastern Pacific Ocean (Adapted from [30-33]).

Figure 2-3 | Humboldt Squid's Key Features in Detail. (A) (A) Schematic illustration of the Humboldt Squid. (B) Upper and lower beak embedded in the buccal mass C graphical representation of the beak's component gradient between the hydrated proximal and dehydrated most distant end of the beak.

Figure 2-4 | Graphical Representation of Glucose Polysaccharides. (A) Cellulose, a natural abundant polysaccharide consisting of β linked D-glucose units primarily found in plant's cell walls. (B) Chitin, a natural polymer consisting of N-acetylglucosamine and the primary component of the exoskeleton of arthropods and also abundant in other organisms, C chitosan, the linearised deacetylated polysaccharide from chitin which is composed of β -linked D-glucosamine and N-acetyl-D-glucosamine.

Figure 2-5 | Secondary Structure of Chitin-Binding Beak Proteins of *D. Gigas*. *Dosidicus gigas*' carbohydrate-binding proteins consist of a 18-25 amino acid long signal peptide and two or three 44-59 amino acid long chitin-binding domains.

Figure 2-6 | Introduction to Chemical Shift Perturbation. (A) Slow exchange, the intensity of the free peak (black) decreases while the bound peak (orange) is increasing

during titration. **(B)** During a fast exchange the peaks move gradually in H and N direction from a free (black) to a bound (orange) stage. (Adapted from: [82]).

Figure 2-7 | Representation of STD NMR Spectroscopy. (Adapted from [91]) During the saturation process of a protein the ligands that are in exchange with the protein become saturated which then can be detected in solution. When subtracting the saturated spectrum from a spectrum without protein irradiation only signals from molecules that bind to the protein appear, while in the difference spectrum only non-binding ligands appear. The signals of the bind protein in the resonance spectrum can then be used for epitope mapping to identify which section of the ligand is in contact with the protein.

Figure 3-1 | Construct Optimisation of His6-TEV-DgCBP-3 Δ 141-201. The initial construct of CBP-3 comprising the third carbohydrate-binding domain was cloned into two vectors to optimise the construct design. While pNIC-GST adds an N-terminal GST-PreSci-His6-TEV segment, pNIC-CH2 only carries a C-terminal His6-tag.

Figure 3-2 | Reference circular dichroism spectroscopy spectra. Reference spectra for a (red) α -helical, (blue) β -sheet, (green) random-coiled protein

Figure 4-1 | Structures of tachycitin and hevein. Depiction of two previously published structures, tachycitin **(A)** and hevein **(B)** structures, containing the the proposed CBM-14 (taken and adapted from Suetake et. al [4]).

Figure 4-2 | Soluble Fraction of CBP-3 Constructs. Small batch purification analysis of 22 protein constructs of DgCBP-3. The construct pNIC-28-BSA-His₆-TEV-DgCBP-3 Δ R141-V201 with a medium high expression of protein in the soluble fraction (yellow) was further optimised by using the plasmids pNIC-GST and pNIC-CH2, which resulted in an increased yield of protein in the soluble fraction after lysis and ultracentrifugation (green).

Figure 4-3 | Construct Design of CBD- γ . (A) Segmentation of the full-length DgCBP-3 protein with its signal peptide (green) and three chitin-binding domains CBD- α - γ (B) Final construct design of CBD- γ with double cleavable N-terminal GST and His₆-tag. (C) Protein sequence of the final TEV-cleaved protein construct with cleavage fragments (QSM).

Figure 4-4 | Gel Filtration Profile and SDS-PAGE of CBD- γ . (Chromatogram) The gel filtration was monitored using an NGC Quest™ 10 Chromatography System at two wavelengths, UV260 and UV280. Due to the low amount of aromatic amino acid residues in the protein, only low absorption was recorded using the UV280 filter. **(SDS-PAGE)** The SDS-PAGE confirms a high content of the full construct of GST-PreSci-His₆-TEV-CBD- γ in the sample after GST affinity chromatography, running at about 80kDa (left band). After cleavage and subsequent gel filtration the final construct of CBD- γ collected from 20mL of elution volume, appears at a molecular size of about 10kDa (based on Thermo Scientific™ Spectra™ Multicolor Broad Range Protein Ladder 10 – 260 kDa).

Figure 4-5 | Homology matching of CBD- γ with the R&R Consensus and Peritrophin-A Domain Sequence comparison of CBD- γ to the original R&R Consensus [6], two proteins with a sequence similarity >80%, the Peritrophin-A domain of *A. duodenale* and two cuticle proteins found in *H. americanus* and *L. polyphemus*. Alignments with the R&R Consensus and the Peritrophin-A domain are highlighted in green and red, respectively. Similarity analysis identified high sequence similarity of CBD- γ to the R&R Consensus in the N-terminus and middle region of the protein and to C-terminus of the Peritrophin-A domain. The two protein a similarity of >80% to CBD- γ , a hypothetical protein found in the octopus *O. bimaculoides* (KOF83921.1) and an obstructor-E-like

protein found in *O. vulgaris* (XP_029637511.1), share high similarity with both the R&R Consensus and the Peritrophin-A domain. In contrast, two cuticle proteins, namely AMP1B from *H. americanus* (P81385) and the cuticle protein 14 from *L. polyphemus* (P83355), feature only similarity with the middle region of CBD- γ and the original R&R Consensus.

Figure 4-6 | Circular Dichroism of CBD- γ in Absence and Presence of PCP. The secondary structure characteristics of CBP- γ were examined using circular dichroism (CD) spectroscopy in the absence (black line) and presence (red line) of PCP to assess structural changes the protein undergoes at an incremental ligand concentration. In the absence of PCP the ellipticity exposes a minimum at 200 nm and a maximum at 192 nm. In presence of PCP CBD- γ exhibits a minimum at 205 nm, a second minimum at 214 nm, and a shift of the maximum from 192 nm to 197 nm.

Figure 4-7 | Binding Analysis of CBD- γ Using Analytical FPLC. Binding of CBP- γ to chitin magnetic beads. A defined amount of protein was loaded onto 100 μ L of chitin magnetic beads and incubated for 2h at 4°C and the supernatant analysed via analytical FPLC using a Superdex 75 10/300 column to measure the amount of unbounded protein in the sample. All positive samples (A-C) indicated that approximately 45% of protein were bound to the beads. As a reference, the same amount of protein was kept under same conditions without adding the chitin magnetic beads. 5 μ g of GST was used as a negative control, showing no binding (D).

Figure 4-8 | Saturation Transfer Difference NMR Spectroscopy of CBD- γ in Presence of PCP. (A) 1D STD-reference and difference NMR spectra of CBD- γ :PCP ($t_{\text{sat}} = 2.0$ s) with different proton resonances (H1, H2, H3, H4, H5, NH(H2) and CH₃ (H2)) of PCP labelled. (B) STD amplification factor of five different proton resonances (H1, H2, H3, NH(H2) and CH₃ (H2)) as a function of increasing saturation times ($t_{\text{sat}} = 0.5$ s, 1 s, 1.5 s, 2 s, 3 s, 4 s, 5 s) to calculate STD_{max} and K_{st} from the equation, $\text{STD-AF} = \text{STD}_{\text{max}}(1 - e^{-k_{\text{st}} \cdot t_{\text{sat}}})$. (C) Stacked 1D STD-difference NMR spectra of acetamide (CH₃) region at a saturation time $t_{\text{sat}} = 2.0$ s, for increasing PCP concentrations. (Inset: fitting of STD amplification factor (STD-AF) against increasing PCP concentrations to obtain K_{D}). (D)

Group epitope mapping of one chitin subunit of PCP based on the STD_{total} value calculated from the equation $STD_{total} = STD_{max} \cdot K_{st}$. The proton that exhibits the highest STD_{total} value was defined as 100% and all other protons are normalized with respect to the highest STD_{total} .

Figure 4-9 | K_D Value Determination of the Binding of Chito-oligomers to CBD- γ . Plotting of the STD-AF vs the ligand concentrations for CH₃COOH (blue), NH₃ (red), H₁ (grey) to determine the proton K_D values. (A) TCT, (B) PCP, (C) HCH, and (D) HpCH.

Figure 4-10 | ¹H-NMR Spectroscopy in Presence and Absence of PCP. (A) in absence of PCP, CBD- γ expresses a dispersed spectrum in the protein's fingerprint amide region that relates to the internal structure of the protein. (B) Upon addition of PCP, the spectrum shows changes, including an increased resonance in the NH region (1), a more intense dispersion of peaks in the fingerprint amide region (2), a new peak in the methyl region (3) and a more intense resonance of the CH₃ peak (4).

Figure 4-11 | Unfolding of CBD- γ in Presence and Absence of PCP. (A-C) In presence of PCP, CBD- γ shows strong resonances which remained distinct even after addition of 8M urea. Adding 20 mM DTT, most of the resonance peaks dispersed and only a few remained. (D-F) In the absence of the PCP carbohydrate, CBD- γ displays less resonance peaks, with many peaks disappearing upon addition of urea. With the addition of DTT, all resonance signals strongly decreased in intensity, suggesting that CBD- γ is fully denatured.

Figure 4-12 | 2D ¹H-¹H TOCSY of DgCBP in Absence and Presence of PCP. (A) In the absence of PCP only a limited number resonance peak arise for CBD- γ , indicating the presence of a structure with only a few structural elements. (B) In presence of PCP resonance peaks of both the CBD- γ and PCP arose, including resonance cross-peaks between the carbohydrate and the protein.

Figure 4-13 | Solution ¹H ¹H NMR Characterisation of CBD- γ in Absence of PCP. Sequence specific resonance assignment of CBD- γ without PCP (A) in the fingerprint

amide region of the two dimensional ^1H - ^1H TOCSY spectrum and **(B)** the two dimensional ^1H - ^1H NOESY spectrum. **(C)** NOE distribution of CBD- γ without of PCP, showing primarily short-range NOEs. **(D)** Histogram of the final structure with sequential (white bars), medium-long NOEs ($i, i+2 - i, i+4$) (grey) and long-range NOEs ($i, i+5$ and above).

Figure 4-14 | Solution ^1H ^1H NMR Characterisation of CBD- γ in Complex with PCP.

Sequence specific resonance assignment of CBD- γ /PCP complex **(A)** in the fingerprint amide region of the two dimensional ^1H - ^1H TOCSY spectrum and **(B)** the two dimensional ^1H - ^1H NOESY spectrum. **(C)** NOE assignment of CBD- γ /PCP complex showing short- and medium-range NOEs. **(D)** Histogram of the final structure with sequential (white bars), medium-long NOEs ($i, i+2 - i, i+4$) (grey) and long-range NOEs ($i, i+5$ and above).

Figure 4-15 | CSD of CBD- γ in Absence and at an Equimolar Protein-Carbohydrate Ratio.

(A) Chemical shift deviations (CSDs) of $\text{H}\alpha$ of CBD- γ residues from random coil without PCP, showing positive chemical shift deviations between Q₁₇₆ and V₂₀₁. **(B)** CSDs of $\text{H}\alpha$ of CBD- γ residues from random coil at an equimolar ratio of PCP with negative chemical shift deviations between G₁₅₉ and Q₁₇₄ and positive chemical shift deviations between Q₁₇₆ and V₁₉₆.

Figure 4-16 | Heteronuclear Protein Assignment using ^1H - ^{15}N HSQC and ^1H - ^{15}N - ^1H NOESY

(A) Assignment of the fingerprint amide region of CBD- γ using ^1H - ^{15}N HSQC spectrum. **(B)** Example sequential walk-through of the folded region of CBD- γ using the ^1H - ^{15}N - ^1H NOESY spectrum.

Figure 4-17 | Solution NMR Structure of CBD- γ .

(A) An ensemble of the backbone of ten lowest energy structures of CBD- γ without PCP. **(B)** A representative conformation of CBD- γ with colour gradient representing flexible, semiflexible and rigid domains of the protein in accordance with the ten lowest energy structures. **(C)** An ensemble of the backbone of ten lowest energy structures of CBD- γ in the presence of PCP. **(D)** Representative conformation of CBD- γ in the presence of PCP where the backbone atoms

of the structurally defined residues are highlighted. The side chain interactions of Y₁₆₄, D₁₆₈ in the helical region interfacing with S₁₇₉ and Q₁₇₅ are represented as sticks.

Figure 4-18 | Three - Dimensional Protein Structure of CBD- γ in Absence (left) and Presence (right) of PCP. Surface net charge of CBD- γ alone (left) and in complex with pentaacetyl-chitopentaose (right). While the folded region of the free protein is primarily negatively charged, the folded regions of the carbohydrate – protein complex is split between a negative net charge of the α -helical region and a positive net charge of the β -sheets.

Figure 4-19 | MD Simulation of the CBD- γ -PCP Complex and Mutant using Glide (A) Prediction of the CBD- γ -PCP complex with the amino acid residues in proximity to the carbohydrate highlighted. F₁₈₃ is predicted to interact with the acetyl group of the acetlyamine of H2, while Y₁₆₄ remains in close proximity with another C2 domain and the saccharide ring system. Other predicted residues in interaction include polar residues located on the α -helix and β ₂-sheet. **(B)** Energy components involved in the binding between CBD- γ and PCP with both, electrostatic (ele) and Van der Waals (VdW) being predominant. **(C)** Binding mechanism of PCP to the Y₁₆₄A, N₁₇₀A, Q₁₇₆A mutant of CBD- γ over 80ns. The MD simulation demonstrates that PCP does bind to the mutant of CBD- γ . **(D)** Distance between the centre of mass of the amino acid residues of CBD- γ and the centre of mass of PCP. While the distance of the trajectory between PCP and the wild-type of CBD- γ remains constant and is even slightly decreasing over time, the distance of the trajectory between PCP and the mutant of CBD- γ expands significantly. The distance between the centre of masses including the wild-type are initially closer than that of the mutant because of the longer side chains of Y, N and F compared to A.

Figure 4-20 | Chemical shift perturbation (CSP) analysis and amide temperature coefficients of protein-carbohydrate complex. (A) The chemical shift perturbation plot of protein in complex with 0.7 mM of carbohydrate plotted against residue number. The residues that show stronger perturbations (above the mean + 2 x standard deviation (SD) line) are marked. **(B)** Amide temperature coefficients of free protein and protein-

carbohydrate complex plotted against residue number. The residues that show stronger CSP show an increase in hydrogen bonding efficiency as marked. **(C)** Strips of 2D ¹H-¹⁵N HSQC of residues that show weaker perturbation in the presence of 0.7 mM PCP. As the protein exhibit tight binding with carbohydrate, at concentration as high as 1.2 mM, the peaks corresponding to slow exchange between free and bound form can be clearly seen. **(D)** Mapping of strong, medium and weakly perturbed residues (decreasing gradient of blue spheres) onto the molecular structure of the complex.

Appendix Figure 6-1 | Soluble Fraction of CBP-1 Constructs. Small batch purification analysis of 21 protein constructs of DgCBP-1 with none of the constructs being in the soluble fraction after lysis and ultracentrifugation

Appendix Figure 6-2 | Soluble Fraction of CBP-2 Constructs. Small batch purification analysis of 19 protein constructs of DgCBP-1 with none of the constructs being in the soluble fraction after lysis and ultracentrifugation

Appendix Figure 6-3 | Soluble Fraction of CBP-4 Constructs. Small batch purification analysis of 12 protein constructs of DgCBP-1 with none of the constructs being in the soluble fraction after lysis and ultracentrifugation

Appendix Figure 6-4 | Ramachandran Plot of CBD-γ The Ramachandran plot of free CBD-γ reports 16 amino acids in most favoured, 38 amino acids in additionally allowed regions and 1 amino acid in a generously allowed region. The residues are generally dispersed over the spectrum in comparison to Ramachandran Plot of CBD-γ in presence of PCP

Appendix Figure 6-5 | Ramachandran Plot of CBD-γ in Presence of PCP The Ramachandran plot of CBD-γ in presence of PCP reports 27 amino acids in most favoured, 27 amino acids in additionally allowed regions and 1 amino acid in a generously allowed region. The residues are well dispersed over the α -helical region between ψ of -45° and 45° and the β -sheet region between a ψ of 90° and 180° .

Appendix Figure 6-6 | Overlay of 2D ^1H - ^1H TOCSY spectra of 700 μM of CBD- γ spectra Overlay of 2D ^1H - ^1H TOCSY spectra of 700 μM of CBD- γ in the presence (green) and absence (red) of 0.7 mM PCP. The amino acid residues of CBD- γ that displays cross peaks with H2, H3 protons of PCP are labelled.

Appendix Figure 6-7 | Binding Analysis of CBD- γ (mut) Using Analytical FPLC Analytical PFLC of (A) CBD- γ and (B) CBD- γ (mut) (Y164 $\rightarrow$ A, N170 $\rightarrow$ A, Q176 $\rightarrow$ A). (C) GST-fusion protein was used as a negative control. Samples were incubated either in the presence of 100 μL of chitin-magnetic beads (blue curves) or without the addition of beads (orange curves) for 2h at 4°C. In the presence of chitin magnetic beads, CBD- γ shows a loss of absorption as compared to both the CBD- γ mutant and the negative control, suggesting that the amino acid residues Y164, N170 and Q176 identified in the NMR experiments are key residues enabling the binding between CBD- γ and chitin.

Abbreviations

HEPES	(4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid)
CBM	Carbohydrate-binding module
CSD	Chemical shift deviation
CSP	Chemical shift perturbation
CBD	Chitin-binding domain
CBD- γ	Chitin-binding domain- γ
CD	Circular Dichroism
DgCBP	Dosidicus gigas Chitin-binding beak proteins
DgHBP	Dosidicus gigas histidine-rich beak protein
kD	equilibrium dissociation constant
FPLC	Fast Protein Liquid Chromatography
GEM	Group epitope mapping
Gly	Glycine
HpCT	Heptaacetyl-chitoheptaose
HSQC	Heteronuclear Single Quantum Coherence
HCT	Hexaacetyl-chitohexaose
IPTG	Isopropyl β -D-1-thiogalactopyranoside
ITC	Isothermal titration calorimetry
kDa	kilodalton
LB Medium	Luria Bertani Medium
μ M	micromolar
mM	millimolar

MWCO	Molecular weight cut-off
GlcNAc	N-Acetyl-Glucosamine
NMR	Nuclear Magnetic Resonance
NOESY	Nuclear Overhauser spectroscopy
OD	Optical density
nOE	Overhauser enhancements
PCP	Pentaacetyl-chitopentaose
R&R	Rebers and Riddiford
RNA-Seq	RNA-sequencing
STD-NMR	Saturation transfer difference NMR
preCols	semi-crystalline collagenous proteins
NaCl	Sodium Chloride
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
TCT	Tetraacetyl-chitotetraose
TSA	Thermal Shift Assay
TEV	Tobacco Etch Virus
TOCSY	Total correlation spectroscopy
UV-VIS	Ultraviolet–visible spectroscopy
XRD	X-ray diffraction
XPS	X-ray photoelectron spectroscopy

Chapter 1 Introduction

*This chapter will briefly introduce distinct types of carbohydrate binding that is found in nature and the translation to the subsequent research conducted on the carbohydrate binding proteins of *Dosidicus gigas*. The chapter will also outline the hypotheses and the key objectives of this project.*

1.1 Background and Hypotheses

1.1.1 Carbohydrate Binding in Nature

Carbohydrates are one of the most abundant biomolecules found in nature across species. Apart from being stored as sugars and fibres, primarily important for the metabolism and used as an energy source, carbohydrates can also be found as cellulose in the cell walls of plants or as chitin in crustaceans' and insects' exoskeletons. And while most metabolic carbohydrates are easily digestible, the majority of carbohydrates are not as easily accessible and require enzymes to either degrade or store those insoluble polysaccharides. Such enzymes are widely distributed across many organisms including, bacteria, fungi, insects, plants, and mammals and support with the digestion, pathogenicity, immunity or moulting [1]. The degradation of the inaccessibility of the glycosidic bond in polysaccharides represents often the limiting factor in the process to make polysaccharides usable degradable [2].

Multiple Carbohydrate-Binding Modules (CBMs) have been described that bind to sugars. Their role has been described primarily in three ways: 1) CBMs increase the targeted direction of catalytic modules towards carbohydrates, 2) CBMs actively lower the proximity in order to concentrate hydrolytic enzymes on the carbohydrate or 3) by shearing the carbohydrates' ordered structure to increase their accessibility [3-5]. Whilst these effects are primarily in relation with the degradation of carbohydrates some CBMs have been reported to play a critical role in the formation of cuticles in arthropods and the exoskeleton in crustaceans or arthropods. Unlike other CBMs that primarily demonstrate enzymatic properties, these CBMs express structural properties and are often correlated with chitin [6, 7]. These structural CBMs share a common motif, known as the Rebers and Riddiford Consensus (R&R Consensus) that has been reported in a wide range of arthropods and crustaceans, but for which no experimental tertiary structure has been described to date. The R&R Consensus is a 35-36 amino acid long conserved region that has been primarily reported for cuticle proteins in both arthropods and crustaceans and which is widely believed to be bind to chitin [7-9]. Previous research has recognised the R&R Consensus to be homologous to the bovine plasma retinol binding protein and

proposed that the basic folding motif is likely to be a half-barrel of anti-parallel β -sheets [10].

1.1.2 Carbohydrate Binding in *Dosidicus Gigas*

Dosidicus gigas, often also referred as the Humboldt squid, belongs to the family of *Ommastrephidae* and can reach up to 2 m in length, up to 45 kg in weight and is a stealthy, elusive predator [11-12]. Research on the beaks and sucker rings (SRT) of cephalopods arose from various angles, including the fish industries and ecological or population research studies, and more recently within the material science field due to its morphometric characteristics, e.g. the material of SRTs can compete with best structural synthetic polymers with an elastic modulus of ~ 8 GPa while representing a fully natural thermoplastic biomaterial [13-15].

The squid beak is a fully-functional organic material built from four components: proteins, chitin, water and pigment [16]. The proportion of the components controls the stiffness of the beak by forming a mechanical bio-gradient that increases between the untanned, hydrated section at the proximal wing and the tanned rostrum [17]. RNA-Seq combined with proteomic studies identified that two protein families, “histidine rich beak proteins” (DgHBPs) and “chitin-binding beak protein”s (DgCBPs) are present in the beak [16]. The later has only recently been found and is proposed to be associated with the formation of a protein-carbohydrate network in which DgHBPs infiltrate in. Preliminary sequence similarity analysis has further promoted an analogy to the CBM-14 motif, a family of chitin binding proteins, found primarily in peritrophic matrix proteins of insects and animal chitinases.

1.1.3 Hypotheses

DgCBPs are a new family of structural proteins for which no extensive research has been conducted on. The proteins are of special interest to understand the mechanism that leads to the formation of the biogradient formed in the Humboldt squid beak, which could be translated to similar natural structures appearing in other cephalopods or in the exoskeletons of arthropods. Based on the limited knowledge about DgCBPs, the following

assumptions were made for the following project:

- a. During the development of the squid embryo, DgCBPs start to interact with chitin in the egg capsules to form a protein-carbohydrate network. The beak continues to grow over the entire lifespan of the squid.
- b. The binding of DgCBPs to chitin shares similarities to the binding of Peritrophin-A, a protein containing an CBM-14 motif, to chitin.
- c. The DgCBP-chitin network forms the base in which DgHBPs infiltrate and coacervate to harden the beak, as DgHBPs are the main proteins present in the rostrum of the beak.

1.2 Objectives and Scope

This thesis aimed to discover the role of DgCBPs in the formation of the squid's beak, as well as to identify the protein structure and its fundamental binding mechanism. Further, the project anticipated to confirm the previous assumption that DgCBPs share a common structure with proteins containing a CBM-14 motif, such as hevein and tachycitin.

Initially all variants of the DgCBP family, namely DgCBP-1, DgCBP-2, DgCBP-3 and DgCBP-4 were considered for structural assessment and to study their binding activity to chitin. The densely cross-linked beak material does not allow the proteins to be extracted easily without harsh methods [16]. In fact, no method has been reported that allows the extraction of full-length proteins from the beak. Consequently, a recombinant expression system was required to study the DgCBP proteins. In view of this, the following objectives were set for the project:

- a. To set up a recombinant expression and purification system for DgCBP-1, DgCBP-2, DgCBP-3 and DgCBP-4

The DNA sequences for the proteins were previously identified by probing a transcriptome of beccublast cells (where the beak proteins are expressed) with proteomic data of protein fragments extracted from mature beaks. The genes encoding for the sequences were designed and cloned into various vectors and transferred either into *E. coli* or into insect cells for recombinant protein expression. Because none of the full-length proteins

expressed in the soluble fraction, protein constructs of the chitin-binding domains of the proteins were designed and again transferred either into *E. coli* or into insect cells. The chitin-binding domain CBD- γ discussed later was purified by FPLC using affinity chromatography and size exclusion chromatography. The purified proteins were analysed by SDS-PAGE.

- b. To qualitatively characterise the activity and to solve the native protein structure of the DgCBPs

Purified CBD- γ was qualitatively characterised by circular dichroism spectroscopy to assess its structure, by analytical FPLC after binding to chitin magnetic beads to assess its binding activity, and by one-dimensional homonuclear NMR to study the structural integrity of the protein.

Studies on the structure of CBD- γ have initially been conducted using X-ray Crystallisation without outcome and subsequently using homo- and heteronuclear NMR spectroscopy. NMR spectroscopy was used to investigate the native structure of CBD- γ both as free CBD- γ and in complex with the soluble chitin derivative PENTAACETYL-CHITOPENTAOSE.

- c. To study the fundamental binding mechanism that is involved in the formation of the beak's structure

The binding between CBD- γ and PCP was assessed using different NMR methods. The methods included Saturation transfer difference (STD) NMR, and the assessment of the chemical shift perturbation (CSP), the chemical shift deviation (CSD) and the amine temperature coefficient using different NMR spectra.

Ultimately, based on the NMR ensemble, molecular dynamics (MD) simulations we were conducted to confirm, visualise and interpret specific molecular interactions between CBD- γ and PCP which were anticipated from the findings from the above-mentioned NMR studies.

1.3 Dissertation Overview

The dissertation on the examination of the structure and mechanism of DgCBPs has a primary focus on the study of CBD- γ from DgCBP-3. The dissertation is separated in five chapters:

Chapter 1 provides a general introduction on carbohydrate binding and a brief excursion on the proposed binding that appears in *D. gigas*, as well as the hypotheses the dissertation is based on, its key objectives, and the scope of the project.

Chapter 2 provides an in-depth review on the literature of protein-based biostructures, on previous studies that have been conducted on *D. gigas* and the beak of *D. gigas*, along with an analysis of previous research that has been conducted on carbohydrate binding domains with particular interest for those that interact with chitin. Chapter 2 also summarises NMR methods that can be used to identify carbohydrate binding in solution.

Chapter 3 contains the methodologies and principles of the techniques and procedures that have been conducted within this project.

Chapter 4 describes the key results of the project in the following way.

Firstly, the results of preliminary sequence analysis are outlined.

Secondly, outcomes of the construct design and protein expression are presented.

Thirdly, findings of the protein purification and qualitative analysis of CBD- γ are described.

Fourthly, the outcomes of the structural analysis of free and CBD- γ in complex with PCP are reported.

Fifthly, the predictions from molecular dynamics studies based on the NMR ensembles.

Finally, the results of the binding studies between CBD- γ and PCP are shared.

Chapter 5 summarises the key outcomes of the previous chapter and concludes on the

results shared. This Chapter also highlights on the impact of the findings on future research and on studies on other cephalopods, crustaceans and insects, that may contain a similar binding mechanism to the one apparent in the beak of *D. gigas*.

1.4 Findings and Outcomes/Originality

This dissertation contains several outcomes that may impact other future projects by:

- a. Describing a purification strategy that led to the expression of a native carbohydrate-binding domain of DgCBP and proposes other optimisation strategies that could lead to the successful purification of other domains.
- b. Describing the first structure of a protein containing an R&R consensus and the impact of PCP on the structural propensity of CBD- γ .
- c. Explaining the binding mechanism that is involved in the complexation CBD- γ with PCP and other related chitooligomers.

1.5 References

1. Arakane, Y. and S. Muthukrishnan, Insect chitinase and chitinase-like proteins. *Cell Mol Life Sci*, 2010. 67(2): p. 201-16.
2. Boraston, A.B., et al., Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J*, 2004. 382(Pt 3): p. 769-81.
3. Duan, C.J., et al., Identification of a novel family of carbohydrate-binding modules with broad ligand specificity. *Sci Rep*, 2016. 6: p. 19392.
4. Herve, C., et al., Carbohydrate-binding modules promote the enzymatic deconstruction of intact plant cell walls by targeting and proximity effects. *Proc Natl Acad Sci U S A*, 2010. 107(34): p. 15293-8.

5. Wang, L., Y. Zhang, and P. Gao, A novel function for the cellulose binding module of cellobiohydrolase I. *Sci China C Life Sci*, 2008. 51(7): p. 620-9.
6. Rebers, J.E. and L.M. Riddiford, Structure and expression of a *Manduca sexta* larval cuticle gene homologous to *Drosophila* cuticle genes. *J Mol Biol*, 1988. 203(2): p. 411-23.
7. Rebers, J.E. and J.H. Willis, A conserved domain in arthropod cuticular proteins binds chitin. *Insect Biochem Mol Biol*, 2001. 31(11): p. 1083-93.
8. Togawa, T., H. Nakato, and S. Izumi, Analysis of the chitin recognition mechanism of cuticle proteins from the soft cuticle of the silkworm, *Bombyx mori*. *Insect Biochem Mol Biol*, 2004. 34(10): p. 1059-67.
9. Suzuki, M., et al., Structural and functional analyses of a strong chitin-binding protein-1 (SCBP-1) from the exoskeleton of the crayfish *Procambarus clarkii*. *Biosci Biotechnol Biochem*, 2013. 77(2): p. 361-8.
10. Hamodrakas, S. J.; Willis, J. H.; Iconomidou, V. A., A Structural Model of the Chitin-Binding Domain of Cuticle Proteins. *Insect Biochemistry and Molecular Biology* **2002**, 32, 1577-1583.
11. Jerep, P. and C.F.E. Roper, An annotated and illustrated catalogue of cephalopod species known to date, in *Cephalopods of the world*. 2010: FAO Species Catalogue for Fishery Purposes.
12. Markaida, U., C. Quiñónez-Velázquez, and O. Sosa-Nishizaki, Age, growth and maturation of jumbo squid *Dosidicus gigas* (Cephalopoda: Ommastrephidae) from the Gulf of California, Mexico. *Fish Res*, 2004. 66: p. 31-47.
13. Martinez, P., A. Sanjuan, and A. Guerra, Identification of *Illexcoindetii*, *I. illecebrosus* and *I. argentinus* (Cephalopoda: Ommastrephidae) throughout the Atlantic Ocean; by body and beak characters. *Mar Biol*, 2002. 141(1): p. 131-143.
14. Groger, J., U. Piatkowski, and H. Heinemann, Beak length analysis of the Southern Ocean squid *Psychroteuthis glacialis* (Cephalopoda: Psychroteuthidae) and its use for size

and biomass estimation. *Polar Biol*, 2000. 23: p. 70-74.

15. Miserez, A., et al., Jumbo squid beaks: inspiration for design of robust organic composites. *Acta Biomater*, 2007. 3(1): p. 139-49.

16. Tan, Y., et al., Infiltration of chitin by protein coacervates defines the squid beak mechanical gradient. *Nat Chem Biol*, 2015. 11(7): p. 488-95.

17. Miserez, A., D. Rubin, and J.H. Waite, Cross-linking chemistry of squid beak. *J Biol Chem*, 2010. 285(49): p. 38115-24.

18. Ventura, S. and A. Villaverde, Protein quality in bacterial inclusion bodies. *Trends Biotechnol*, 2006. 24(4): p. 179-85.

Chapter 2 Literature Review

Evolution in nature has demonstrated the ability to develop and tune highly sophisticated biological systems as a solution to highly specific environments and functions [1]. As part of the evolution, those that were not able to adapt to the environment can now either be found as fossils or will always remain unknown. Materials which surround us today, underwent multiple steps of evolution over millions of years, at an unparalleled efficiency, durability and simplicity into highly complex hierarchical structures, but are yet built from a fairly limited selection of components [2]. Some of these combine mechanical, biological and other distinctive functions that inspired mankind traceable back to at least Leonardo da Vinci and Galileo Galilei, who first observed the relationships between functions and anatomy in living organisms [3, 4].

2.1 Biomimicry in Material Science

Biomimetic or bioinspired research has gained increasing attention in the past decades, opening up many opportunities to study functional and structural features from the biological world [1]. For example, *collagen*, a structural protein most abundant in skin and connective tissues, is widely used as wound healing aids in burn patients, in cosmetics, but also to support digestion and satiety. Another example, *silk*, has demonstrated a wide range of usability across biomedical and engineering applications, including applications in tissue engineering, drug delivery, or photonics [101-103]. A more exotic example, *sericin* is a waste product of the silk industry, that can be used for various applications, including cosmetics, wound healing and as an anti-microbial product, due to its biocompatibility, biodegradability and bioactivity [5].

Research in biomimicry is inspired by solving current technological problems, for which human technology has so far failed to find solutions. One field of interest includes the engineering of graded bio-composite materials. Organisms generally consist of an assemblage of hard and soft tissues (such as bones, tendons, ligaments, nails) that are connected by attachment sites – so called entheses - which result in tremendous interfacial or mechanical stresses that commonly lead to damage or fracture over time [6, 7]. However, graded materials mitigate contact damage by introducing a mechanical gradient. In human, the transition between skin and fingernails or the dentin-enamel junction in teeth are classical example of adaptive materials that minimise interfacial stresses [8, 9]. A similar strategy to reduce contact damage has also been adopted in marine organisms. For example, mussels form tough, fibrous, self-healing filaments, so called byssus, that consists of an array of semi-crystalline collagenous proteins (preCols) and that function as an anchoring holdfast [10]. Another example, the jaw of the annelid marine worm (polychaete) is made a natural collagen-based block copolymer that forms a mechanical stiffness gradient controlled by the ratio between three individual secretory proteins [11-13].

The most prominent strategy to form these tough, graded materials involves biomineralization, which leads to the formation of load-bearing structures that are far tougher than their individual components. Biominerals are organic-inorganic hybrid structures that precisely arrange an inorganic phase (such as calcium carbonate, calcium phosphate, silicate, iron oxides and other metal sulphides) in a “Lego” block-like fashion over several length scales, with an organic phase between the building blocks that direct mineral formation and also “glue” the inorganic phase together [14, 15].

A second strategy was discovered in the jaw of the polychaete bloodworm (*Glycera*) that consists of an organic cross-linked network, in which only a small fraction of a copper-based biomineral (atacamite) is incorporated in the organic phase [11]. Compared to the classic biomineralization, the bloodworm’s jaw incorporates much less minerals, however.

A third strategy to form graded materials rigorously avoids the use of any inorganic material (mineral phase), which is very unusual. An example is the beak of cephalopods that smoothly couples mismatching interfaces by incorporating a biogradient consisting of water, proteins, chitin and pigments [16, 17].

2.2 Proteins in Material Science

To develop methods of polymerisation that allow simultaneous control over molecular length, sequence and stereochemistry has been a major challenge in polymer chemistry. Natural polymers, particularly extracellular matrix proteins such as collagens, elastins, silk or biological adhesives, have received a strong interest by polymer chemists in recent years, notably because of their genetically-controlled polymerisation or post-translational modifications that provide specific functionalities. [18].

Proteins are generally classified into eight distinctive types based on their function: enzymatic, hormonal, defensive, storage, carrier, contractile, receptor and structural. The latter is of major interest for material science.

Figure 2-1 | Proteins in Material Science

(A) Schematic representation of the Collagen assembly (Adapted from: [19]) (B) Spider-silk fibre assembly is being formed by shear forces that cause the C-terminus to expose its hydrophobic surfaces [20] (C) well-known cytoskeletal filament proteins i) lamin, ii) F-actin, iii) microtubule that provide mechanical support in cells (Adapted from: [21-23])

2.2.1 The Structure and Property of Material Proteins

Common structural proteins, also known as fibrous protein, can be distinguished as either extracellular or intracellular proteins, based on their spatial location. Prominent extracellular proteins include *collagen*, *elastin*, *keratin* or *silk* (Figure 2.1, A & B). Fibrous proteins are secreted from the cell into the extracellular matrix and have load-bearing functions for tissues such as muscles and skin, but are also abundant in hair, nails and teeth. In contrast, intracellular proteins support the architecture of a cell and include proteins such as *microtubule*, *actin* or *lamin* [24] (Figure 2.1, C).

Structural proteins share a hierarchical architecture and feature common secondary protein structures. A classic example is collagen, which features a distinctive repeating triplet of Gly-X-Y, in which Y is often a proline (Pro). The triplet forms a poly-proline helix governed by the reduced backbone flexibility of Pro present in high abundance along the collagen sequences. Additionally, Pro is often hydroxylated to provide enhanced thermal stability. Three coiled subunits form collagen fibrils by winding around each other in a characteristic right-handed helix. These fibrils are building blocks of collagen that consists of multiple collagen fibrils [25] (Figure 2.1, A). In some organisms, the structure of the protein is highly dependent on the environment. For example, spider dragline silk consists of two proteins, *Spidroin I* and *Spidroin II*, that form α -helices in solution when it is stored in its secretory gland, but transform into β -sheet structures during assembly into strong silk fibres [26] (Figure 2.1, B).

Though many natural materials have been studied extensively over the past years, numerous challenges remain unsolved in order to develop a more extensive understanding of the structure-property correlation to the protein's mechanical response. The main challenge describes that fact that most of the processes that regulate the activity or properties, of proteins, such as load-bearing properties, are not understood. Another obstacle during the synthesis of nature-like materials is the formation of hydrogen bonds which represents the most abundant bonds in nature [27].

Unlike human-made materials, biological materials coevolved with the environment and adapted to its changes as a response to the evolution. These adaptations include but not exhaust the arrangement and physical properties, such as elasticity and robustness, of the material. As the formation of such materials occurs under more eco-friendly conditions compared to synthetic materials, understanding its assembly and “structure-forming” process can help develop novel sustainable materials with various application purposes.

2.3 Dosidicus Gigas and its Beak: A Heroic Biomaterial

Dosidicus gigas is an elusive predator found in the Humboldt Current in the Eastern Pacific Ocean. The organism has triggered the interest in the biomimetic research community for its two un-mineralized tissues, its sucker ring teeth and its beak, of which the squid beak is the focus of this thesis.

2.3.1 *Dosidicus Gigas*

Dosidicus gigas - also known as jumbo flying squid or Humboldt squid - belongs to the family of *Ommastrephidae* and can reach up to 2 m in length and up to 45 kg in weight [28]. Its anatomy was first described by the Zoologist Alcide D'Orbigny, in 1835, after his expedition to the South American continent between 1826 and

Figure 2-2 | Anatomy and Geographical Distribution of *D. Gigas*

(A) Zoological representation of the Humboldt Squid with flaring tentacles and its beak by Alcide D'Orbigny (1835) [29], (B) Geographical distribution map of *D. gigas* in the Eastern Pacific Ocean (Adapted from [30-33]).

1833, including drawings of its sucker rings and intriguing beak (**Figure 2.2; A**) [29]. Perennially inhabited in the neritic-zone of Humboldt Current in the Eastern Pacific Ocean, the Humboldt squid has invasively expanded its geographical range in the last 40 years to the waters off central California (**Figure 2.2; B**) [31, 33]. *D. gigas* is a stealthy, elusive, and aggressive predator with long arms, each with 100-200 sucker rings equipped with hooked teeth and whomping tentacles that thump at a fast speed. Propelled by water pulsed to a speed of up to 24 kmph, jumbo squids are considered to be one of the fastest swimmer in the ocean [34]. Usually acting in shoals of often more than a thousand, *D. gigas* also exhibits a cannibalistic behaviour in the event of an injured conspecific [35, 36]. Research on the beaks of cephalopods arose from various angles, including the fish industries and ecological or population research studies, and within the material science field due to its morphometric characteristics [16, 37, 38].

Figure 2-3 | Humboldt Squid's Key Features in Detail

(A) Schematic illustration of the Humboldt Squid. (B) Upper and lower beak embedded in the buccal mass C graphical representation of the beak's component gradient between the hydrated proximal and dehydrated most distant end of the beak.

2.3.2 *D. Gigas'* Un-mineralised Beak and its Composition

The Humboldt squid's beak is one of the most astonishing examples of how Nature's prowess seamlessly designs fully-functional organic materials (**Figure 2-3; B**). It is one of the toughest and stiffest materials build entirely from organic components [39]. Not only its robustness, but also its embedment of the sharp beak within the soft buccal envelope is its most exciting feature, which is used to penetrate the thick dorsal integument of its prey. The buccal mass is gelatinous with a Young Modulus E between 10 – 50 MPa (**Figure 2-3: C**). In contrast, the rostrum at the very tip of the squid beak shows a Young Modulus E of about 5 GPa [16, 17]. However, the two materials make no pivotal contact in the squid. Instead, the beak incorporates a 200-fold mechanical bio-gradient that decreases the modulus incrementally to $E = 50$ MPa towards the untanned, hydrated section at the proximal wing that is rooted in the buccal mass [40].

It has been discovered that hardness and stiffness are controlled by a relative ratio variation of the beak's components, water, chitin, protein and a pigment. The hydrated, soft, pliant section forms the proximal wing of the beak and contains 70 % water, 25 % chitin and 5 % protein. With decreasing water content, the concentration of protein and pigment increases towards the rostrum. It has been shown in other organisms that the pigmentation, a quinone tanning, is associated with the proteins bound to chitin layers [41]. At the very tip, the levels of water and chitin decreased to 5% and 15%, respectively, whereas the content of protein and pigment increased to 60 %, and 20 %, respectively [40]. The higher amount of pigments is also responsible for the tan of the rostrum.

For the development of bio-inspired, gradual and functional materials based on the Humboldt Squid's beak, the basic understanding of the structure and biosynthesis of the material has to be obtained. First studies included the characterisation of the beak proteins using tandem mass spectrometry initially and RNA-Seq to profile the transcriptome [28]. The characterisation identified two protein families, "histidine rich beak proteins" (DgHBPs) and chitin-binding beak proteins (DgCBPs). The first, DgHBPs, were found to contain high contents of histidine, glycine and alanine and are located mainly the pigmented, hard region of the beak. The later, DgCBPs, were identified as having a much lower content of histidine, glycine and alanine (Table 2-1).

Table 2-1| Amino acid content of DgCBPs and DgHBPs. Contents (in mol %) of selected amino acids for DgCBPs and DgHBPs, of the *D. gigas* squid beak. (Taken from [39])

	Asx	Gly	Ala	His
DgCBP-1	8.9	5.5	3.1	3.1
DgCBP-2	13.5	5.7	2.1	5.7
DgCBP-3	13.8	6.9	7.4	1.5
DgCBP-4	11.2	3.4	4.9	2.4
DgHBP-1	4.0	32.2	19.3	10.4
DgHBP-2	10.3	20.7	16.3	15.8
DgHBP-3	1.9	34.4	19.1	14.8

2.4 Chitin

2.4.1 General Overview

After cellulose, chitin is the second most abundant natural mucopolysaccharide on earth and yet is also one of the most underutilised resources. While cellulose consists of β linked D-glucose units and is primarily found in plant's cell walls, chitin features the structure of a repeating sequence of N-Acetylglucosamine, a glucose derivative, that is polymerised in β -1,4 conjugation, and is found in various organisms, including the exoskeletons of crustaceans, the cell walls of fungi and in insects [42]. Due to its high natural

Figure 2-4 | Graphical Representation of Glucose Polysaccharides

(A) Cellulose, a natural abundant polysaccharide consisting of β linked D-glucose units primarily found in plant's cell walls. (B) Chitin, a natural polymer consisting of N-acetylglucosamine and the primary component of the exoskeleton of arthropods and also abundant in other organisms, (C) Chitosan, the linearised deacetylated polysaccharide from chitin which is composed of β -linked D-glucosamine and N-acetyl-D-glucosamine.

abundance chitin represents a huge potential for biomimicry applications. Through the years several techniques have been proposed and published to extract chitin from different natural sources with crustaceans shells being the most commonly used source [39, 43-45].

2.4.2 Common Applications

Despite its high content of nitrogen, chitin has exceptionally low immunogenicity. Over the past years chitin and its derivatives have been introduced as source of a variety of solutions, including of biomedical such as drug delivery systems, wound dressings, in food and chemical industries and other applications such as in dentistry, cosmetics and agriculture among others [46-50]. Another promising polysaccharide represents chitosan,

the cationic N-deacetylated derivative of chitin which shares structurally similarities with hyaluronic acid and glycosaminoglycan in the extracellular matrix [51]. Due to its antibacterial and antifungal properties, chitosan can be firmly anchored to titanium alloys to be used in prosthesis. Till date no clinical trial on Chitosan in conjunction with prosthesis has been conducted [52, 53].

Other potential applications for chitosan is in the field of cartilage tissue regeneration. Studies on chitosan / hyaluronic acid tuned composite glycosaminoglycans scaffolds in the extracellular matrix resulted in a better pore network configuration and exhibited enhanced extracellular matrix cartilage formation [51].

Accessing the U.S National Library of Medicine identified 77 clinical trials on Chitosan till date, of which 38 have been completed (accessed via ClinicalTrials.com in October 2019). Therapeutic fields range from applications in oncology to chronic disease management and wound healing.

2.4.3 Structural Characterisation

The characterisation of chitin and its derivatives goes back to 1937 when Clark and Smith conducted the X-ray diffraction (XRD) studies on chitin, chitosan and its derivatives which they have obtained from lobster [54]. Subsequent studies include the use of X-ray photoelectron spectroscopy (XPS), solid state NMR, solution NMR, UV-VIS or Raman spectrometry, to study the structure and acetylation content of chitin among other properties [55-57].

2.5 Carbohydrate-Binding Domains

Enzymes to degrade or store insoluble polysaccharides are widely distributed in many organisms including, bacteria, fungi, insects, plants, mammals, and play an essential role in digestion, pathogenicity, immunity and in the moulting process of arthropods [58]. Despite the importance of polysaccharides in various processes, their degradation is rather inefficient as the glycosidic bonds are often inaccessible to the active site of glycoside hydrolases [59]. Due to their sugar-binding functions, Carbohydrate-binding modules (CBMs) are found to promote three functional roles, namely 1) targeting, 2) proximity, and 3) disruptive effects.

CBMs that demonstrate targeting of carbohydrates and proximity capability have been reported numerous in literature [60-62]. As an example, CBMs are often appended to cell wall hydrolases and play a pivotal role in plants by directing the catalytic module towards polysaccharides in intact cell walls through the recognition of non-substrate polysaccharides. Hence they actively exhibit strong proximity effects where they concentrate hydrolytic enzymes on the cell-wall polysaccharides [63]. Moreover, carbohydrate recognition by proteins has further been reported to be a pivotal step in viral and microbial infection, defence in plants, and inflammatory responses among others [64]. In these cases, it has been reported that the binding to carbohydrates is often driven by aromatic amino acids within the recognition domain of CBMs [65].

Another type of CBMs are reported to disrupt the structure of polysaccharides. Initially documented for the N-terminal family 2a CBM of Cel6A from *Cellulomonas fimi*, disruptive CBMs appear to shear the ordered structure of highly-ordered substrates and increase their accessibility to the appended enzymes by weakening and disrupting the non-covalently linkage of the hydrogen bonding network. [66, 67].

Finally, CBMs have also been reported to play a critical role in the formation of cuticles in arthropods and the exoskeleton in crustaceans. Unlike discussed above, such CBMs mainly play a mechanical support role, which is influenced by the proportion of chitin and the type of protein found therein among other [68]. This type of CBMs have been found to share a chitin-binding sequence, known as the Rebers and Riddiford Consensus (R&R Consensus).

2.5.1 Rebers and Riddiford Consensus

The R&R Consensus is a 35 – 36 amino acid long sequence that was first identified in cuticular proteins and is present throughout many arthropods [69, 70]. While the original consensus was described as G-x(8)-G-x(6)-Y-x-A-x-E-x-GF-x(7)-P-x-P, it has gone through a number of alterations since its discovery. One modification, the extended R&R Consensus (pfam00379 (IPR000618)), has a length of about 64 amino acids, which does not include any cysteine residues, and has been shown to bind to chitin in both insects and crusades [68, 71, 72]. Further modifications, namely RR-1, RR-2 and RR-3, were introduced by Andersen in 1998 and 2000, which were isolated first from soft cuticles, hard cuticles and postecdysial cuticle in insects, respectively [73, 74]. A study combining Raman, infrared and Circular Dichroism (CD) spectroscopy measurements on a soft cuticle protein inferred that the R&R Consensus was likely to adopt a β -pleated sheet conformation [104]. Hamodrakos et al. recognized that the R&R consensus was homologous to bovine plasma retinol binding protein and proposed that the basic folding motif is likely to be a half-barrel of anti-parallel β -sheets. They further hypothesized that the interactions with chitin are mediated via the flat surface of the protein, which is composed of aromatic side chains that would interact with the polysaccharide chains of chitin [105]. While the function of proteins containing the R&R Consensus has been widely explored, to date, no experimental study has reported the structure of a CBD harbouring the R&R Consensus.

2.5.2 Chitin-Binding Beak Proteins

Previous analysis of DgCBPs confirmed the presence of four protein variants. They feature the presence of an 18-25 amino acid signal peptide and the existence of either two or three, proposedly, chitin-binding type 2 domains that contain six conserved cysteines (**Figure 2-5**) [39]. Such motive, also commonly referred as Peritrophin-A domain, is found in the peritrophic matrix of insects and was first reported for the protein Peritrophin-44 found in the larvae of the fly *Lucilia cuprina* [75]. The sequence of the Peritrophin-A domain contains a sequence of six conserved cysteines that are proposed to form three disulphide bonds and belongs to the carbohydrate-binding module-14 (CBM-14) family (pfam 01607), also known as chitin-binding domain type 2.

Earlier studies by Miserez *et al.* conducted in 2010 also suggested the presence of the above-mentioned R&R Consensus, however for DgHBPs in the squid beak [40]. Their suggestion, however, was made before the presence of DgCBPs was confirmed in 2015 [39]. Conclusively, the initially suggested binding via a β -sheet conformation to the catecholic region of chitin can also be applied to the DgCBPs.

Figure 2-5 | Secondary Structure of Chitin-Binding Beak Proteins of *D. Gigas*

Dosidicus gigas' carbohydrate-binding proteins consist of a 18-25 amino acid long signal peptide and two or three 44-59 amino acid long chitin-binding domains.

2.6 NMR as a Tool to Provide Structural and Biochemical Insights of the Structure of Carbohydrate Binding Proteins

Due to the commonly high molecular weights of carbohydrates, detailed studies on the binding between proteins and carbohydrates have historically been reported using X-ray crystallography but recently also more frequently by using Nuclear Magnetic Resonance (NMR) spectroscopy [64, 76-78].

The use of NMR spectroscopy in studies of carbohydrates is generally favoured over X-ray crystallography, primarily due to the reoccurring presence of the CBD motifs and to the use of smaller chitin derivatives (GlcNAc polysaccharides). NMR spectroscopy demonstrates the advantage whereby the structure and thermodynamic of the affinity of carbohydrates to the active sides of the carbohydrate-binding proteins can be studied dynamically at the molecular level [79, 80]. As an example, Asensio *et al* have published a detailed analysis of the binding strength between Hevein and (GlcNAc)_n oligomers which shows that a minimum of five GlcNAc units are critical to express all possible protein-carbohydrate interactions simultaneously [64].

2.6.1 Chemical Shift Perturbation (CSP) Analysis

CSP analysis is one noteworthy analytical method to determine the protein residues involved in carbohydrate interaction. CSP analysis is carried out by titrating an unlabelled ligand into a ¹⁵N-labelled protein solution while acquiring a 2D Heteronuclear single quantum coherence spectroscopy (HSQC) spectrum [81]. The accuracy of CSP analysis is based on the theory that the chemical shift change is sensitive to structural changes of the protein, and consequently almost any genuine binding interaction will produce CSPs. Unlike ¹H shifts which are largely influenced by through-space interactions (such as hydrogen bonds), ¹⁵N or ¹³C shifts can be influenced by through-bond (determined via TOCSY, HSQC experiments) and through-space interactions (determined via NOESY experiments). Therefore, the calculated geometrical distance detected for a peak in presence and absence of a ligand in the ¹⁵N HSQC can be used to determine the location of a bound ligand in a complex [82].

Like any enzymatic reaction the binding between the ligand (L) and the protein (P) in a single side, is described by the reversible equation:

where the forward and backward reactions are described by the rate constants k_{on} and k_{off} , respectively. As a result, if the k_{off} is significantly slower than the chemical shift difference between the bound and free proteins, then the peak signals gradually disappear (**Figure 2-6: A**). In contrast, if the k_{off} is greater than the chemical shift difference, then the peak signals will gradually migrate from their initial position in the free spectrum to the position in the bound spectrum (**Figure 2-6; B**).

As an example, in a study conducted by Farina *et al* on Cyclophilin A and the Apoptosis Inducing Factor a dose-response reduction of cross peak intensities was reported, which stood in correlation with the formation of a larger complex, in a slow-intermediate exchange regime between the free and the bound protein [83].

Figure 2-6 | Introduction to Chemical Shift Perturbation

(A) Slow exchange, the intensity of the free peak (black) decreases while the bound peak (orange) is increasing during titration. (B) During a fast exchange the peaks move gradually in H and N direction from a free (black) to a bound (orange) stage. (Adapted from: [82]).

2.6.2 Amide Proton Temperature Dependence in Proteins

Amide protons ($^1\text{H}_\text{N}$) demonstrate a strong dependency on temperatures which is fundamentally based on the presence of inter- and intramolecular hydrogen bonds in a protein [106]. Intermolecular hydrogen bonds are generally weaker than intramolecular bonds. As a result, their structural perturbations are more prominent and consequently the $^1\text{H}_\text{N}$ chemical shifts more temperature dependent. This correlation between the amide proton temperature and the hydrogen bonds can be used as a supporting method to identify the amino acid residues that are in interaction with a ligand. Previous analyses have defined the values of the amide proton ($^1\text{H}_\text{N}$) temperature coefficient that are till today widely used to predict hydrogen bonds in folded proteins [84-86]. A later study by *Cierpicki and Otlewski* further defined the value for $^1\text{H}_\text{N}$ with a temperature gradient more positive than -4.6 ppb/K to be form a hydrogen bond with a probability of 85% and with a probability of 93% if the amides are within the range of -4 to -1 ppb/K [87].

2.6.3 STD-NMR as Tool to Study Protein-Carbohydrate Interactions

The elucidation of the binding mechanism between oligosaccharides and proteins is of high interest in research due to its potential for clinical applications. While x-ray crystallography and modelling can provide detailed information on the protein-ligand complex, the methods lack the analysis of the thermodynamic driving force for the complexation [64]. Saturation transfer difference NMR spectroscopy (STD-NMR) is a powerful ligand-observed NMR method, based on the nuclear Overhauser effect between bound protons and all protons, to study the protein-ligand interaction in solution [88]. Initially designed to screen a library of carbohydrates against wheat-germ agglutinin, STD-NMR has developed to be a versatile tool to also identify the epitope binding of a ligand and determine the dissociation constant (K_D) using a single-ligand titration approach [89, 90]. The STD spectrum represents the difference between a spectrum in which the protein is saturated and one in which the protein is unsaturated (Figure 2-7).

Figure 2-7 | Representation of STD NMR Spectroscopy.

(Adapted from [91]) During the saturation process of a protein the ligands that are in exchange with the protein become saturated which then can be detected in solution. When subtracting the saturated spectrum from a spectrum without protein irradiation only signals from molecules that bind to the protein appear, while in the difference spectrum only non-binding ligands appear. The signals of the bind protein in the resonance spectrum can then be used for epitope mapping to identify which section of the ligand is in contact with the protein.

STD-NMR has the advantage to determine not only ligands of a protein but also to provide information about the specificity of the binders by returning binding constants when performed in a titration experiment and for epitope mapping among other applications [92, 93].

Epitope mapping represents a useful approach to understand the binding mechanism between a ligand and a protein. For example, a study of the interaction of Polymyxin B to lipopolysaccharide (LPS) from *E.coli* 0111:B4 showed unambiguously that aromatic ring protons of residue D-Phe6 exhibited the strongest STD signals and therefore interacted with the LPS micelles [94]. Hence, the epitope mapping could considerably support the understanding of the interaction between the micelle and the antibiotic at an atomic level.

Despite the fact that a direct evaluation of the dissociation constant (K_D) using variation of STD-NMR is not feasible, combining it with a single-ligand titration experiment does allow a direct measurement of the protein-ligand dissociation constant. A key variable of the STD-NMR experiment is the saturation time t_{sat} , which has an impact on the peak intensities and the signal-to-noise ratio, and consequentially the sensitivity of the experiment. [95]. Although adjustable, typically saturation times of 2 sec are used for STD-NMR experiments to achieve a better signal-to-noise ratio. However, a study done by Angulo *et al* showed that the saturation time of 2 sec has an impact on the normalised STD amplification factor, which led to an increase of the apparent K_D value. Thus an analysis of the STD-NMR saturation time should be performed prior to determining the K_D of the ligand [89].

Once an adequate saturation time t_{sat} has been determined and by ignoring the competitive binding aside, the K_D can then be determined using a model for enzymatic kinetics (Eq. 2):

$$K_D = \frac{\alpha_{STD} [L]}{STD_{amp} + [L]} \quad (\text{Eq. 2}).$$

Where α_{STD} is the maximum amplification and STD_{amp} is the amplification factor of the STD experiment

2.7 Molecular Dynamics Simulation and Docking of CBD- γ and PCP

Virtual molecular screening is an often assessed approach to support the identification of new ligands that bind to a protein of interest or to optimise lead candidate structures of ligands in structure-based drug discovery projects. While for lead identification large libraries of ligands are being screened in high-throughput, the subsequent step, lead optimisation, commonly only requires a smaller set of compounds [96]. Further, virtual screening can also be used to predict the geometries in which proteins and ligands are oriented in a complex by applying predictive methods on their structures.

In brief, such methods are using hierarchical filters to identify possible conformations in which the ligand can attach to the binding site of a protein. Programs such as *Glide* use collections of ligand conformations, based on various torsions of the compound, which are assessed during the docking process. In a last step, a limited number of ligand conformations are minimised and docked to the binding site with full flexibility.

For the docking analysis of CBD- γ and PCP, the 3D structure of PCP was retrieved from ChemSpider (CSID27445378) for docking and molecular dynamic (MD) simulations on CBD- γ . The ligand structure was minimized using the Macromodel module employing OPLS 2005 force field in Schrodinger 9.0 [97]. The parameters for PCP were generated using the package Antechamber and were further optimised for energy minimisation and MD simulations using ff14SB force-field in AMBER18 to obtain lowest energy structures [98, 99]. Clusters were generated from the MD simulations and the conformations containing the highest populated cluster was selected for docking analysis.

The optimised PCP was docked onto the MD energy minimised structure of CBD- γ residues C₁₆₁, Y₁₆₄, S₁₆₆, N₁₇₀ and D₁₇₄ using *Glide* [96]. The inner box was kept as default with a size of 10 $\times$ 10 $\times$ 10 Å and the search space confined to the centre of the aforementioned amino acid residues for docking of PCP. The *Glide* Grid Generation settings were kept at default.

Since the protein conformation was determined using NMR spectroscopy, a rigid receptor docking protocol was used in *Glide* that fixes the protein's conformation, but allowed PCP to remain flexible. 5 docked conformations were identified and each subjected to MD simulations. The simulations were carried out for 100 ns using the standard protocol in AMBER18. Subsequently, the trajectories in which the ligand was bound stably, were taken for further analysis. Clustering was done on the three MD trajectories in which the ligand was stably bound. The top clusters populated with more than 25-30% of sampled conformations were assessed and the interactions which are the closest (within an RMSD of 0.5-0.6 Å) to the centroid of the highly populated cluster analysed respectively.

2.8 Scope of the PhD Project in Context of Previous Research

The elucidation of the mechanism that leads to the recognition and improvement of the carbohydrate accessibility by carbohydrate-binding proteins remains of high interest in current research. Among the myriad of carbohydrates, chitin represents one of the most abundant carbohydrates on earth. Though abundant it remains also one of the most inaccessible energy sources.

The Humboldt Squid, *Dosidicus gigas*, is just one example in nature that makes use of chitin by incorporating chitin in its wholly organic, multi-layered bio-graded beak. From a materials engineering perspective, the wholly organic, biodegradable, graded squid beak is a model load-bearing structure that can provide lessons in regenerative medicine or sustainable biocomposites. Like any other biological material, the understanding of the fundamentals of the squid beak structure and binding mechanism is critical to elucidate its properties and also help interpret the mechanism found in other organisms.

Based on the limited knowledge available about the binding mechanism of chitin-binding proteins, we propose to elucidate the structural features of the chitin-binding domain (CBD) of one chitin-binding beak protein from *Dosidicus gigas* and the binding mechanism to various chitin derivatives.

Since CBDs are functionally similar by binding to carbohydrates across organisms, the aim is to characterise one domain from one chitin-binding beak protein, namely CBD- γ , which encompasses the third carbohydrate-binding domain of CBP-3. This domain has been chosen for the study for the following reasons: (i) its molecular weight (7.15 kDa) is within range to determine the structure of the protein using NMR even when unlabelled; (ii) after screening a wide range of constructs it was present in the soluble fraction after expression; and (iii) it includes the six cysteines that are being reported to form three disulphide bonds in peritrophin-A. The in-depth analysis of the structure and intrinsic binding mechanism will not only help to clarify how chitin recognition is achieved in the squid beak but also in other organisms that contain chitin and CBPs.

Using standardised recombinant expression tools, we aimed to express the recombinant protein CBD- γ in the soluble fraction for basic characterisation using NMR spectroscopy. A major challenge of the project was that initial attempts to express the proteins failed and an initial sequence analysis on the basis of the correlation between stability of a protein and its dipeptides classified the protein as instable [100]. Given the timeframe of the project, various protein constructs were initially designed, and promising candidates shortlisted to further refine the expression before moving the project into the purification and characterisation phase. For structural characterisation, both X-Ray crystallography and NMR spectroscopy were initially explored, with the later finally been used to solve the structure in solution and to assess the binding properties to derivatives of the proteins' natural ligand.

The ultimate research goal underlying this project is to better understand the inter-molecular interactions between proteins and carbohydrate in the squid beak, and to understand how these correlates to the mechanical gradient of the beak.

In summary, this project outlines: (i) microbiological methods, namely construct design of recombinant proteins, expression and purification; (ii) structural biological methods, namely NMR spectroscopy and limited use of X-Ray crystallography; (iii) biophysical methods, namely CD spectroscopy, surface plasmon resonance spectroscopy and analytical FPLC. Analysis of the tertiary structure and binding affinity of the protein will be analysed by integrating experimental methods and computational tools. We envision that our insights can be translated into organisms with similar biomacromolecular components, namely chitin and carbohydrates, and also help to establish the design of bio-graded, degradable, protein/carbohydrate biocomposites.

2.9 References

1. Fratzl, P., Biomimetic materials research: what can we really learn from nature's structural materials? *J R Soc Interface*, 2007. 4(15): p. 637-42.
2. Wegst, U.G., et al., Bioinspired structural materials. *Nat Mater*, 2015. 14(1): p. 23-36.
3. Da Vinci, L., *The notebooks of Leonardo da Vinci*. 1952: Oxford world's classics.
4. Ascenzi, A., Biomechanics and Galileo Galilei. *J Biomech*, 1993. 26(2): p. 95-100.
5. Srinivas, N., et al., Extraction & characterization of sericin and its immobilization on hydroxylapatite nanoparticles for tissue engineering applications. *International Journal of ChemTech Research*, 2015. 7(5): p. 2117-2124.
6. Benjamin, M., et al., Where tendons and ligaments meet bone: attachment sites ('entheses') in relation to exercise and/or mechanical load. *J Anat*, 2006. 208(4): p. 471-90.
7. Suresh, S., Graded materials for resistance to contact deformation and damage. *Science*, 2001. 292(5526): p. 2447-51.
8. Bendit, E., Mechanical properties of keratins III: volume swelling and the longitudinal modulus in relation to matrix protein content. *Proc. 6th Quinquennial Int. wool textile research*, 1980. 2: p. 247-257.
9. Marshall, G.W., Jr., et al., Mechanical properties of the dentinoenamel junction: AFM studies of nanohardness, elastic modulus, and fracture. *J Biomed Mater Res*, 2001. 54(1): p. 87-95.
10. Priemel, T., et al., Rapid self-assembly of complex biomolecular architectures during mussel byssus biofabrication. *Nat Commun*, 2017. 8: p. 14539.

11. Lichtenegger, H.C., et al., High abrasion resistance with sparse mineralization: copper biomineral in worm jaws. *Science*, 2002. 298(5592): p. 389-92.
12. Gosline, J., et al., Elastic proteins: biological roles and mechanical properties. *Philos Trans R Soc Lond B Biol Sci*, 2002. 357(1418): p. 121-32.
13. Waite, J.H., et al., Exploring molecular and mechanical gradients in structural bioscaffolds. *Biochemistry*, 2004. 43(24): p. 7653-62.
14. Colfen, H., Biomineralization: A crystal-clear view. *Nat Mater*, 2010. 9(12): p. 960-1.
15. Lowenstam, H.A. and S. Weiner, *On Biomineralization*. 1989, Oxford University Press.
16. Miserez, A., et al., Jumbo squid beaks: inspiration for design of robust organic composites. *Acta Biomater*, 2007. 3(1): p. 139-49.
17. Miserez, A., et al., The transition from stiff to compliant materials in squid beaks. *Science*, 2008. 319(5871): p. 1816-9.
18. Tirrell, D.A., M.J. Fournier, and T.L. Mason, Protein engineering for materials applications. *Current Opinion in Structural Biology*, 1991. 1: p. 638-641.
19. Mouw, J.K., G. Ou, and V.M. Weaver, Extracellular matrix assembly: a multiscale deconstruction. *Nat Rev Mol Cell Biol*, 2014. 15(12): p. 771-85.
20. Rising, A., Controlled assembly: a prerequisite for the use of recombinant spider silk in regenerative medicine? *Acta Biomater*, 2014. 10(4): p. 1627-31.
21. Stuurman, N., S. Heins, and U. Aebi, Nuclear lamins: their structure, assembly, and interactions. *J Struct Biol*, 1998. 122(1-2): p. 42-66.
22. Lodish, H., A. Berk, and S.L. Zipursky, *The Dynamics of Actin Assembly*, in *Molecular Cell Biology*, W.H. Freeman, Editor. 2000: New York.

23. Rochlin, M.W., M.E. Dailey, and P.C. Bridgman, Polymerizing microtubules activate site-directed F-actin assembly in nerve growth cones. *Mol Biol Cell*, 1999. 10(7): p. 2309-27.
24. Alberts, B., A. Johnson, and J. Lewis, *Molecular Biology of the Cell*. Vol. 4. 2002, New York: Garland Science.
25. Lodish, H., A. Berg, and S.L. Zipursky, *Collagen: The Fibrous Proteins of the Matrix*, in *Molecular Cell Biology*. 2000, W. H. Freeman: New York.
26. Simmons, A.H., C.A. Michal, and L.W. Jelinski, Molecular orientation and two-component nature of the crystalline fraction of spider dragline silk. *Science*, 1996. 271(5245): p. 84-7.
27. Ackbarow, T., et al., Hierarchies, multiple energy barriers, and robustness govern the fracture mechanics of alpha-helical and beta-sheet protein domains. *Proc Natl Acad Sci U S A*, 2007. 104(42): p. 16410-5.
28. Jerep, P. and C.F.E. Roper, An annotated and illustrated catalogue of cephalopod species known to date, in *Cephalopods of the world*. 2010: FAO Species Catalogue for Fishery Purposes.
29. Orbigny, A.D.d., et al., *Voyage dans l'Amérique méridionale : (le Brésil, la république orientale de l'Uruguay, la République argentine, la Patagonie, la république du Chili, la république de Bolivie, la république du Pérou)*, exécuté pendant les années 1826, 1827, 1828, 1829, 1830, 1831, 1832, et 1833. 1835, Paris
Strasbourg: Pitois-Levrault ; Ve. Levrault.
30. Wormuth, J.H., The biogeography and numerical taxonomy of the oegopsid squid family Ommastrephidae in the Pacific Ocean. *Bulletin of the Scripps Institution of Oceanography*, 1976. 23: p. 1-90.
31. Nigmatullin, C., K. Nesis, and A. Arkhipkin, A review of the biology of the jumbo squid *Dosidicus gigas* (Cephalopoda: Ommastrephidae). *Fish Res*, 2001. 54(9-19).

32. Gilly, W.F., Spreading and stranding of jumbo squid., in *Exosystems observations for the Monterey Bay National Marine Sactuary*. 2006. p. 25-26.
33. Zeidberg, L.D. and B.H. Robison, Invasive range expansion by the Humboldt squid, *Dosidicus gigas*, in the eastern North Pacific. *Proc Natl Acad Sci U S A*, 2007. 104(31): p. 12948-50.
34. Markaida, U., C. Quiñónez-Velázquez, and O. Sosa-Nishizaki, Age, growth and maturation of jumbo squid *Dosidicus gigas* (Cephalopoda: Ommastrephidae) from the Gulf of California, Mexico. *Fish Res*, 2004. 66: p. 31-47.
35. Nesis, K., The biology of the jumbo squid of Peru and Chile, *Dosidicus gigas*. *Oceanology*, 1970. 10: p. 108-118.
36. Nesis, K., *Dosidicus gigas*. *Cephalopod Life Cycles*, ed. P. Boyle. Vol. 1. 1983, London: Academic Press London.
37. Martinez, P., A. Sanjuan, and A. Guerra, Identification of *Illexcoindetii*, *I. illecebrosus* and *I. argentinus* (Cephalopoda: Ommastrephidae) throughout the Atlantic Ocean; by body and beak characters. *Mar Biol*, 2002. 141(1): p. 131-143.
38. Groger, J., U. Piatkowski, and H. Heinemann, Beak length analysis of the Southern Ocean squid *Psychroteuthis glacialis* (Cephalopoda: Psychroteuthidae) and its use for size and biomass estimation. *Polar Biol*, 2000. 23: p. 70-74.
39. Tan, Y., et al., Infiltration of chitin by protein coacervates defines the squid beak mechanical gradient. *Nat Chem Biol*, 2015. 11(7): p. 488-95.
40. Miserez, A., D. Rubin, and J.H. Waite, Cross-linking chemistry of squid beak. *J Biol Chem*, 2010. 285(49): p. 38115-24.
41. Zhou, F., et al., Sexual asynchrony in the development of beak pigmentation for the neon flying squid *Ommastrephes bartramii* in the North Pacific Ocean. *Fisheries Science*, 2016. 82(5): p. 737-746.

42. Kumar, M.N.V.R., A review of chitin and chitosan applications. *Reactive and Functional Polymers*, 2000. 46(1): p. 1-27.
43. Al Sagheer, F.A., et al., Extraction and characterization of chitin and chitosan from marine sources in Arabian Gulf. *Carbohydrate Polymers*, 2009. 77(2): p. 410-419.
44. Synowiecki, J. and N.A. Al-Khateeb, Production, Properties, and Some New Applications of Chitin and Its Derivatives. *Critical Reviews in Food Science and Nutrition*, 2003. 43(2): p. 145-171.
45. Hayes, M., et al., Mining marine shellfish wastes for bioactive molecules: chitin and chitosan--Part B: applications. *Biotechnol J*, 2008. 3(7): p. 878-89.
46. Rejinold, N.S., et al., Multifunctional chitin nanogels for simultaneous drug delivery, bioimaging, and biosensing. *ACS Appl Mater Interfaces*, 2011. 3(9): p. 3654-65.
47. Abdel-Mohsen, A.M., et al., Novel chitin/chitosan-glucan wound dressing: Isolation, characterization, antibacterial activity and wound healing properties. *Int J Pharm*, 2016. 510(1): p. 86-99.
48. Husain, S., et al., Chitosan Biomaterials for Current and Potential Dental Applications. *Materials (Basel)*, 2017. 10(6).
49. Sharp, R.G., A Review of the Applications of Chitin and Its Derivatives in Agriculture to Modify Plant-Microbial Interactions and Improve Crop Yields. *Agronomy*, 2013. 3: p. 757-793.
50. Aranaz, I., et al., Cosmetics and Cosmeceutical Applications of Chitin, Chitosan and Their Derivatives. *Polymers*, 2018. 10(2): p. 213.
51. Correia, C.R., et al., Chitosan scaffolds containing hyaluronic acid for cartilage tissue engineering. *Tissue Eng Part C Methods*, 2011. 17(7): p. 717-30.

52. Costa, F., et al., Characterization of hLF1-11 immobilization onto chitosan ultrathin films, and its effects on antimicrobial activity. *Acta Biomater*, 2014. 10(8): p. 3513-21.
53. Yang, C.C., et al., Vancomycin-chitosan composite deposited on post porous hydroxyapatite coated Ti6Al4V implant for drug controlled release. *Mater Sci Eng C Mater Biol Appl*, 2013. 33(4): p. 2203-12.
54. Clark, G.L. and A.F. Smith, X-Ray Diffraction Studies of Chitin, Chitosan, and Derivatives. *The Journal of Physical Chemistry*, 1935. 40(7): p. 863-879.
55. Heux, L., et al., Solid state NMR for determination of degree of acetylation of chitin and chitosan. *Biomacromolecules*, 2000. 1(4): p. 746-51.
56. Kumirska, J., et al., Application of spectroscopic methods for structural analysis of chitin and chitosan. *Mar Drugs*, 2010. 8(5): p. 1567-636.
57. Minke, R. and J. Blackwell, The structure of alpha-chitin. *J Mol Biol*, 1978. 120(2): p. 167-81.
58. Arakane, Y. and S. Muthukrishnan, Insect chitinase and chitinase-like proteins. *Cell Mol Life Sci*, 2010. 67(2): p. 201-16.
59. Boraston, A.B., et al., Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J*, 2004. 382(Pt 3): p. 769-81.
60. Duan, C.J., et al., Identification of a novel family of carbohydrate-binding modules with broad ligand specificity. *Sci Rep*, 2016. 6: p. 19392.
61. Beguin, P. and J.P. Aubert, The biological degradation of cellulose. *FEMS Microbiol Rev*, 1994. 13(1): p. 25-58.
62. Gilbert, H.J., J.P. Knox, and A.B. Boraston, Advances in understanding the molecular basis of plant cell wall polysaccharide recognition by carbohydrate-binding modules. *Curr Opin Struct Biol*, 2013. 23(5): p. 669-77.

63. Herve, C., et al., Carbohydrate-binding modules promote the enzymatic deconstruction of intact plant cell walls by targeting and proximity effects. *Proc Natl Acad Sci U S A*, 2010. 107(34): p. 15293-8.
64. Asensio, J.L., et al., NMR investigations of protein-carbohydrate interactions: studies on the relevance of Trp/Tyr variations in lectin binding sites as deduced from titration microcalorimetry and NMR studies on hevein domains. Determination of the NMR structure of the complex between pseudohevein and N,N',N''-triacetylchitotriose. *Proteins*, 2000. 40(2): p. 218-36.
65. Guillen, D., S. Sanchez, and R. Rodriguez-Sanoja, Carbohydrate-binding domains: multiplicity of biological roles. *Appl Microbiol Biotechnol*, 2010. 85(5): p. 1241-9.
66. Gao, P.J., et al., Non-hydrolytic Disruption of Crystalline Structure of Cellulose by Cellulose Binding Domain and Linker Sequence of Cellobiohydrolase I from *Penicillium janthinellum*. *Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai)*, 2001. 33(1): p. 13-18.
67. Wang, L., Y. Zhang, and P. Gao, A novel function for the cellulose binding module of cellobiohydrolase I. *Sci China C Life Sci*, 2008. 51(7): p. 620-9.
68. Rebers, J.E. and J.H. Willis, A conserved domain in arthropod cuticular proteins binds chitin. *Insect Biochem Mol Biol*, 2001. 31(11): p. 1083-93.
69. Rebers, J.E. and L.M. Riddiford, Structure and expression of a *Manduca sexta* larval cuticle gene homologous to *Drosophila* cuticle genes. *J Mol Biol*, 1988. 203(2): p. 411-23.
70. Cornman, R.S., et al., Annotation and analysis of a large cuticular protein family with the R&R Consensus in *Anopheles gambiae*. *BMC Genomics*, 2008. 9: p. 22.
71. Togawa, T., H. Nakato, and S. Izumi, Analysis of the chitin recognition mechanism of cuticle proteins from the soft cuticle of the silkworm, *Bombyx mori*. *Insect Biochem Mol Biol*, 2004. 34(10): p. 1059-67.

72. Suzuki, M., et al., Structural and functional analyses of a strong chitin-binding protein-1 (SCBP-1) from the exoskeleton of the crayfish *Procambarus clarkii*. *Biosci Biotechnol Biochem*, 2013. 77(2): p. 361-8.
73. Andersen, S.O., Amino acid sequence studies on endocuticular proteins from the desert locust, *Schistocerca gregaria*. *Insect Biochem Mol Biol*, 1998. 28: p. 421-434.
74. Andersen, S.O., Studies on proteins in post-ecdysial nymphal cuticle of locust, *Locusta migratoria*, and cockroach, *Blaberus craniifer*. *Insect Biochem Mol Biol*, 2000. 30: p. 569-577.
75. Elvin, C.M., et al., Characterization of a major peritrophic membrane protein, peritrophin-44, from the larvae of *Lucilia cuprina*. cDNA and deduced amino acid sequences. *J Biol Chem*, 1996. 271(15): p. 8925-35.
76. Quijcho, F.A., Carbohydrate-binding proteins: tertiary structures and protein-sugar interactions. *Annu Rev Biochem*, 1986. 55: p. 287-315.
77. Weis, W.I. and K. Drickamer, Structural basis of lectin-carbohydrate recognition. *Annu Rev Biochem*, 1996. 65: p. 441-73.
78. Peters, T. and B.M. Pinto, Structure and dynamics of oligosaccharides: NMR and modeling studies. *Curr Opin Struct Biol*, 1996. 6(5): p. 710-20.
79. Jimenez-Barbero, J., et al., Free and protein-bound carbohydrate structures. *Curr Opin Struct Biol*, 1999. 9(5): p. 549-55.
80. Asensio, J.L., et al., NMR investigations of protein-carbohydrate interactions: refined three-dimensional structure of the complex between hevein and methyl beta-chitobioside. *Glycobiology*, 1998. 8(6): p. 569-77.
81. Zuiderweg, E.R., Mapping protein-protein interactions in solution by NMR spectroscopy. *Biochemistry*, 2002. 41(1): p. 1-7.

82. Williamson, M.P., Using chemical shift perturbation to characterise ligand binding. *Prog Nucl Magn Reson Spectrosc.*, 2013. 73: p. 1-16.
83. Farina, B., et al., Structural and biochemical insights of CypA and AIF interaction. *Sci Rep*, 2017. 7(1): p. 1138.
84. Baxter, N.J. and M.P. Williamson, Temperature dependence of ¹H chemical shifts in proteins. *J Biomol NMR*, 1997. 9(4): p. 359-69.
85. Dyson, H.J., et al., Folding of immunogenic peptide fragments of proteins in water solution. I. Sequence requirements for the formation of a reverse turn. *J Mol Biol*, 1988. 201(1): p. 161-200.
86. Jiminez, M.A., et al., A study of the NH NMR signals of Gly-Gly-X-Ala tetrapeptides in H₂O at low temperature. *J. Mol. Biol.*, 1986. 143: p. 435-438.
87. Cierpicki, T. and J. Otlewski, Amide proton temperature coefficients as hydrogen bond indicators in proteins. *J Biomol NMR*, 2001. 21(3): p. 249-61.
88. Haselhorst, T., A.C. Lamerz, and M. Itzstein, Saturation transfer difference NMR spectroscopy as a technique to investigate protein-carbohydrate interactions in solution. *Methods Mol Biol*, 2009. 534: p. 375-386.
89. Angulo, J., P.M. Enriquez-Navas, and P.M. Nieto, Ligand-receptor binding affinities from saturation transfer difference (STD) NMR spectroscopy: the binding isotherm of STD initial growth rates. *Chemistry*, 2010. 16(26): p. 7803-12.
90. Mayer, M. and B. Meyer, Characterization of Ligand Binding by Saturation Transfer Difference NMR Spectroscopy. *Angew Chem Int Ed Engl*, 1999. 38(12): p. 1784-1788.
91. Meyer, B. and T. Peters, NMR spectroscopy techniques for screening and identifying ligand binding to protein receptors. *Angew Chem Int Ed Engl*, 2003. 42(8): p. 864-90.

92. Vogtherr, M. and T. Peters, Application of NMR Based Binding Assays to Identify Key Hydroxy Groups for Intermolecular Recognition. *J. Am. Chem. Soc.*, 2000. 123: p. 6108-6117.
93. Viegas, A., et al., Saturation-Transfer Difference (STD) NMR: A Simple and Fast Method for Ligand Screening and Characterization of Protein Binding. *Journal of Chemical Education*, 2011. 88(7).
94. Bhunia, A. and S. Bhattacharjya, Mapping residue-specific contacts of polymyxin B with lipopolysaccharide by saturation transfer difference NMR: insights into outer-membrane disruption and endotoxin neutralization. *Biopolymers*, 2011. 96(3): p. 273-87.
95. Mayer, M. and B. Meyer, Group epitope mapping by saturation transfer difference NMR to identify segments of a ligand in direct contact with a protein receptor. *J Am Chem Soc*, 2001. 123(25): p. 6108-17.
96. Friesner, R.A., et al., Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem*, 2006. 49(21): p. 6177-96.
97. Jorgensen, W.L., D.S. Maxwell, and J. Tirado-Rives, Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 1996. 118: p. 11225-11236.
98. Maier, J.A., et al., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput*, 2015. 11(8): p. 3696-713.
99. Wang, J., et al., Antechamber, An Accessory Software Package For Molecular Mechanical Calculations. *J. Am. Chem. Soc.*, 2000(222).
100. Guruprasad, K., B.V.B. Reddy, and M.W. Pandit, Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng*, 1990. 4(2): p. 155-161.
101. Omenetto, F.; Kaplan, D. From silk cocoon medical miracle. *Scientists are crafting*

arteries, ligaments, circuitry and holograms from worm yarn. *Sci. Am.* 2010, 303 (5), 76–7.

102. Wenk, E.; Merkle, H. P.; Meinel, L. Silk fibroin as a vehicle for drug delivery applications. *J. Controlled Release* 2011, 150 (2), 128–41.

103. Omenetto, F. G.; Kaplan, D. L. A new route for silk. *Nat. Photonics* 2008, 2 (11), 641–643

104. Iconomidou, V. A.; Chrysikos, G. D.; Gionis, V.; Willis, J. H.; Hamodrakas, S. J., "Soft"-cuticle protein secondary structure as revealed by FT-Raman, ATR FT-IR and CD spectroscopy. *Insect Biochem Mol Biol* **2001**, 31 (9), 877-85.

105. Hamodrakas, S. J.; Willis, J. H.; Iconomidou, V. A., A Structural Model of the Chitin-Binding Domain of Cuticle Proteins. *Insect Biochemistry and Molecular Biology* **2002**, 32, 1577-1583.

106. Wagner, G., Pardi, A. and Wuthrich, K. (1983) *J. Am. Chem. Soc.*, 1983, 5948–5949.

Chapter 3 Experimental Methodology

This project employed experimental approaches from the areas of molecular biology, structural biology, biophysics and to a certain extent bioinformatics to investigate the structural and biophysical properties of Dosidicus gigas chitin binding beak proteins. The primary experimental techniques used include: (i) construct design of recombinant proteins, expression and purification; (ii) NMR spectroscopy, (iii) CD spectroscopy, surface plasmon resonance spectroscopy and analytical FPLC and (iv) bioinformatical tools. In this chapter, these experimental techniques, and to a certain extent, their theoretical background, will be introduced.

3.1 Rational

To characterise the structure and function of CBD- γ in detail, a range of experimental procedures was applied in this project. Firstly, to derive to a sufficient amount of CBD- γ for analytical purposes, a recombinant protein expression had to be optimised. Secondly, bioanalytical methods were used to verify the proper folding of the protein and to qualitatively study its activity. Thirdly, structural characterisations were conducted on CBD- γ in the absence and presence of derivatives of the native ligand chitin in the solution state. Finally, additional analytical methods on an atomic level were conducted to quantify the binding of CBD- γ .

3.2 Construct Design, Cloning & Analytical Small-Scale Expression of DgCBPs

The sequences for *D. gigas* DgCBPs were obtained from previous (unpublished) studies done in our laboratory. In total, 85 constructs of the four DgCBPs comprising either the full-length protein sequence or the CBDs only were submitted to the Protein Production Platform (PPP) at in the School of Biological Sciences of NTU. The cloning was performed in 96-well format using ligation-independent cloning (LIC) technology in *Escherichia coli* (*E.coli*) or in *Spodoptera frugiperda* Sf21 cells (SF9 cells) [102-104]. All target constructs were designed in various vectors, including pNIC28-Bsa4, pNIC-CH2, pSUMO-LIC and pNIC-GST for expression in *E.coli* and pFB-LIC-Bse and pFB-LIC-SUMOStar for expression in Sf9 cells. Apart from varying the vectors, the constructs were also expressed in various strands of *E.coli*, including BL21(DE3) Rosetta T1R and BL21(DE3) pFH255.

The clones were screened in a small-scale analytical expression in cell cultures of 1 mL Terrific Broth (TB) to assess the likelihood of successful large-scale protein expression and purification. The bacteria were grown in 96-deep well blocks at 37°C to an optical density at 600nm of OD₆₀₀ 2 - 3, the temperature lowered to 18°C and protein expression induced with 1mM IPTG. Preliminary expression was verified by using SDS-PAGE.

3.3 Construct Optimisation for DgCBP-3

The initial construct His₆-TEV-DgCBP-3^{ΔR141-V201} (His₆-TEV-CBD-γ) was used as a starting point to design further constructs. Two constructs, GST-PreSci-His₆-TEV- CBD-γ (in pNIC-GST) and CBD-γ-His₆ (in pNIC-CH2), were designed and expressed in BL21(DE3) Rosetta-gami, a strain commonly used for high-level expression that carries additional tRNAs to overcome the effects of codon bias. While BL21(DE3) Rosetta gami carries a natural chloramphenicol resistance, both pNIC vectors are designed to carry kanamycin resistance.

Construct	Molecular Weight	
His ₆ -TEV-DgCBP-3 ¹⁴¹⁻²⁰¹	9.4 kDa
GST-PreSci-His ₆ -TEV-DgCBP-3 ¹⁴¹⁻²⁰¹	36.5kDa
DgCBP-3 ¹⁴¹⁻²⁰¹ -His ₆	7.7 kDa

Figure 3-1 | Construct Optimisation of His₆-TEV-CBD-γ

The initial construct of CBP-3 comprising the third carbohydrate-binding domain was cloned into two vectors to optimise the construct design. While pNIC-GST adds an N-terminal GST-PreSci-His₆-TEV segment, pNIC-CH2 only carries a C-terminal His₆-tag.

3.4 Expression of DgCBP-3

Prior to large scale, a pre-culture of DgCBPs-carrying bacteria were inoculated in LB medium, supplemented with kanamycin (50 $\mu\text{g mL}^{-1}$) and chloramphenicol (30 $\mu\text{g mL}^{-1}$) antibiotics, overnight at 37°C under continuous shaking. Each 1L of 2x LB medium was inoculated with 1:100 pre-culture and supplemented with antibiotics. The cells were grown under continuous shaking at 37°C until an OD₆₀₀ of 0.7 was reached, chilled to 18°C, and 0.5 mM of IPTG added to induce the over-expression at reduced shaking for 18 hours. The cells were harvested at 2500 g at 4°C (J-Lite JLA-8.1000 fixed-angle rotor, Beckman Coulter, Brea, CA, USA) and subsequent used for protein extraction. Immediate protein extraction was assessed as the preferred method as storing the harvested cells at either -20°C or -80°C yielded in a much lower amount of protein after purification.

3.5 Protein Purification of GST-PreSci-His6-TEV- CBD- γ

The harvested cell pellets containing GST-PreSci-His6-TEV-DgCBP-3_{R141-V201} were resuspended in 10 times the pellet weight in lysis buffer (20mM HEPES, pH 7.0, 500mM NaCl, supplemented with cOmplete™ Protease Inhibitor Cocktail tablets (Roche, Basel, Switzerland)) and homogenized in three circles at the 25,000 psi using high-pressure homogenisation (Model: LM20 Microfluidizer High Pressure Homogeniser, Microfluidics, Boston, MA, USA) on ice. The resulting homogenate was sedimented at 13,000 $\times$ g (JA-30.50 Ti Fixed Angle Titanium Rotor, Beckman Coulter, Brea, CA, USA) at 4°C for 60 minutes to separate cell debris and the supernatant collected for subsequent protein purification.

Each 1 mL of lysis buffer-equilibrated Gluthatione Sepharose 4 Fast Flow resin (GE Healthcare Life Sciences, Westborough, MA, USA) was added to each 20mL of supernatant and incubated at 4°C under continuous stirring for at least 4 hours. The resin beads were washed with 20 column volumes of wash buffer (20mM HEPES, pH 7.0, 500mM NaCl) and 2 column volumes of wash buffer added to the resin. The cleavage of the GST-PreSci-His6-TEV from the target CBP-3_{ΔR141-V201} construct was induced by adding 4mg of in-house purified Tobacco Etch Virus (TEV) protease for each 1mL of Gluthatione Sepharose 4 Fast Flow resin and performed at 4°C under continuous rotation over-night.

Standardised, the most used recognition site for TEV protease is the seven-amino acid sequence “ENLYFQS”, which in fact is just one version of the linearised epitope E-X-X-Y-X-Q-(G/S), with X tolerating multiple amino acids without any loss of activity [105]. TEV protease recognises the sequence and cuts specifically between Q and S (or Q and G). Performing the cleavage “on-resin” ensures that the GST-PreSci-His₆-TEV-protein tag remains on the Gluthatione Sepharose 4 Fast Flow resin while the supernatant contains the targeted CBP-3_{ΔR141-V201} protein construct.

The supernatant was separated from the resin, the beads flushed with 2 column volumes of wash buffer and both, the supernatant and flow-through, pooled and concentrated using a protein concentrator with a molecular weight cut-off of 3 kDa (Pierce™ Protein Concentrator PES, 3K MWCO, Thermo Fisher Scientific, Waltham, MA, USA) to a final volume of 5mL.

The protein concentrate was loaded onto a Superdex S75 16/600 (GE Life Sciences) for further cleaning using an NGC™ Chromatography Systems (Bio-Rad Laboratories, Hercules, CA, USA) using gel filtration buffer (20 mM HEPES, pH 6.7, 300 mM NaCl) at 4°C.

Finally, DgCBP-3_{R141-V201} was concentrated to the desired concentrations for analytics and homo-nuclear NMR experiments.

Table 3-1 | Protein Purification Buffers

	Buffer	Salt	Supplement
Lysis Buffer	20mM HEPES, pH 7.0	500 mM NaCl	cOmplete™ Protease Inhibitor Cocktail
Wash Buffer	20mM HEPES, pH 7.0	300 mM NaCl	
Gel Filtration Buffer	20 mM HEPES, pH 6.7	300 mM NaCl	

3.6 Expression of DgCBP-3 Constructs for Nitrogen-15 NMR Spectroscopy

For analysis using N₁₅ – labelled CBP-3_{R141-V201} the expression protocol as outlined in “3.4 *Expression of DgCBP-3 constructs*” was followed but adjusted in order include the labelling of the protein as followed.

Instead of inoculating the overnight culture 2x LB medium, M9 minimal medium, a salt base formulation that can be supplemented with alternative carbon, nitrogen and amino acid sources, was used [106]. Each 20 mL of LB overnight culture were inoculated in 1L of M9 minimal medium, supplemented with 1 g L⁻¹ of ¹⁵NH₄Cl (Cambridge Isotope Laboratories, Inc., Tewksbury, MA, USA). The culture was grown to an OD₆₀₀ of 0.8 – 1.0 at 37°C under shaking, chilled to 17°C and rested for 1 hour before inducing the overexpression of the target protein by adding IPTG with a final concentration of 0.5 mM for 18 hours. The cells were harvested at 2500 *x g* at 4°C (J-Lite JLA-8.1000 fixed-angle rotor, Beckman Coulter, Brea, CA, USA) and subsequent used for protein extraction following the same protocol as described under section “3.5 *Protein purification of GST-PreSci-His6-TEV- DgCBP-3_{R141-V201}*”.

Further, protein expression using Nitrogen-15 and Carbon-13 in M9 minimal medium was carried out but did not yield in an appropriate amount of protein to conduct NMR experiments. All subsequent heteronuclear NMR experiments were therefore performed with Hydrogen-1 and Nitrogen-15, only.

3.7 Analysis of DgCBP-3 Constructs

A commonly used method to analyse the components of a protein mixture is Sodium Dodecyl Sulphate PolyAcrylamide Gel Electrophoresis (SDS-PAGE). SDS-PAGE combines the usage of sodium dodecyl sulphate (SDS) to denature the proteins and a discontinuous polyacrylamide gel (PAGE) as a support medium to separate the protein mixture by size to identify proteins, estimate the molecular weight and assess the purity of the sample. Detection of the proteins is commonly been done via colorimetric detection using Coomassie Brilliant Blue, a stain based on triphenylmethane.

Due to low molecular weight of CBP-3^{R141-V201} primarily 4 – 20 % Mini-PROTEAN® TGX™ Precast Protein Gels (Bio-Rad Laboratories, Inc) were used. 2 µL of Laemmli Sample Buffer (4x) was mixed with 8 µL of protein to prepare the samples for denaturing gel electrophoresis.

Samples were heated at 90°C for 10 minutes under continuous shaking, centrifuged, and loaded onto the gel. The gels were in 1x Tris/Glycine/SDS (2.5 mM Tris HCl, 19.2 mM glycine, 0.01% (v/v) SDS, pH 8.3) for separation. Graded 4 – 20 % were used to create a good separation of the containing protein bands. As a reference for the identification of the molecular weight of the proteins, PageRuler™ Prestained Protein Ladder was used. Unlike house-made gels, pre-cast gels have a higher heat resistance. Therefore, gel electrophoresis was run at 210 V for 35 min.

For visualisation of the protein patterns on the gel, the gel was stained in Coomassie Brilliant Blue for 30 minutes and destained in 70% v/v ethanol until background transparency got acquired.

3.8 Circular Dichroism Spectroscopy

To assess the structural integrity of the proteins and to examine the secondary structure of proteins, circular dichroism (CD) spectroscopy is a frequently used bioanalytical technique. CD spectroscopy is based on the proteins' differential absorption of left and right circularly polarized light, which arises from chiral (optical active) molecules [6]. In the far ultraviolet (UV) spectrum proteins express highly sensitive to their adopted molecular conformation [7] with the far UV spectrum, the peptide bond is one of the main chromophores to be monitored in the CD spectrum [8]. Other prominent secondary structures in the far UV region are α -helices and β -sheets. α -helical proteins display two negative bands at 222 nm and 208 nm and a positive band at 193 nm, β -pleated sheets contain a minimum at 218 nm and maximum at 195 nm, whereas random-coiled or unstructured proteins (Figure 3-2; green), only contain a low ellipticity at around 215 nm and a minimum at around 195 nm [9-11].

Figure 3-2 | Reference Circular Dichroism Spectroscopy Spectra.

Reference spectra for a (red) α -helical, (blue) β -sheet, (green) random-coiled or unstructured protein

The far-UV CD spectra were measured using a chriscan spectropolarimeter (Model 420, AVIV Biomedical Inc.) using Quartz supracil cells (0.2 mm path length; Hellma Analytics). The spectra were recorded as triplicates at 18 °C, at wavelengths between 240 nm and 190 nm incrementally, using 1 nm as step size and a 1 nm bandwidth, with instrument dynode voltage less than 600. The baseline was acquired using the protein sample buffer. All spectra were smoothed using the Savitzky-Golay filter using MATLAB and the acquired spectra adjusted by subtracting the smoothed, mean averaged baseline spectrum from the smoothed, mean averaged protein sample spectrum [12].

For the protein – ligand interaction studies, the protein was mixed at molar ratios of 1:1, 1:2.5 and 1:5 with pentaacetyl-chitopentaose (PCP) prior to the measurement and incubated for 30 min at room temperature before loading into the quartz cuvette (0.2 mm path length; Hellma Analytics). Spectra were again acquired as triplicates at 18 °C, at wavelengths between 250 and 190nm incrementally using the same specifications. Spectra were again smoothed using the Savitzky-Golay filter and baseline adjusted using the recorded buffer spectrum.

3.9 Analytical FPLC

Analytical fast protein liquid chromatography (FPLC) is a commonly used method to evaluate the content of a sample. A variation of the method can be used to assess quality of binding of the protein to its ligand.

For the assessment, 5 µg (in 200 µL) of CBD-γ was either loaded onto 100 µL of Chitin Magnetic Beads (E8036L, New England BioLabs) or pipetted in an empty tube and both samples incubated at 4°C for 6h under permanent rotation. The samples were centrifuged through a 96-well filter plate (0.2 µm pores) and loaded in sequence onto a Gel Filtration buffer-equilibrated Superdex 75 10/300 GL column (17-5174-01 GE Life Sciences) and run at a flow rate of 1 mL min⁻¹ at 4 °C.

The chromatograms were exported, overlaid and compared to assess the binding activity of CBP-3_{AR141-V201} to the Chitin Magnetic Beads.

3.10 Fluorescence-based Melting Temperature Determination

Fluorescence-based differential scanning, also known as Thermal Shift Assays (TSA), is a commonly used method to monitor the temperature induced unfolding of a protein in presence of a fluorescent dye. With unfolding of the protein chain, the hydrophobic backbone of the protein becomes exposed and the fluorescence signal increases until the protein is completely denatured [13]. A frequently used TSA assay based on the fluorescence dye SYPRO Orange is performed in a real time polymerase chain reaction, and is widely applied to provide information about the melting temperature T_m of a protein or to detect ligand-induced shifts of the melting temperature as a result of its stabilising or destabilising effect [14, 15].

To determine the melting temperature of CBD- γ a serial dilution ranging between 1 μ g and 0.1 μ g of protein (absolute) in 12.5 μ L setup and screened against 12.5 μ L of a serial dilution of SYPRO Orange of 10 x to 1x (final concentration) in a 8x8 matrix (including a negative protein and a negative SYPRO Orange sample) and run in a CFX96 Touch™ Real-Time PCR Detection System (Bio-Rad). Temperature-induced unfolding has been recorded between 30°C and 85°C at a ramp of 0.5°C per second. Detected signals were integrated to determine the T_m of the protein.

3.11 Structural Analysis of CBD- γ

While the amino acid sequence of a protein can often be used to predict the secondary structural elements of a protein, such α -helices, β -sheets or disulphide bonds, presently it is impossible to deduce reliably the tertiary structure of a novel protein solely on the basis of its primary sequence [16]. Frequently applied techniques in protein research used to acquire and study the tertiary structures of proteins are x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. Though x-ray crystallography was applied as initial method to study the three-dimensional structure of CBD- γ , the final structure of the protein was eventually assessed using NMR spectroscopy.

3.11.1 Protein X-ray Crystallography

Protein crystallography allows the determination of the three-dimensional structure of proteins at near atomic resolution. Today, it is one of the most widely preferred methods in drug discovery research, not only aiming to design new lead candidates but also to improve existing drugs [16]. In short, X-ray beams are electromagnetic radiations of short wavelength. If directed at a well-ordered protein crystal, a small fraction of the X-ray beam is being scattered and recorded as Bragg diffraction spots by a detector. The diffraction pattern contains the intensities relation to structural information of the atoms in the protein crystal which can be translated to the protein's structure, once the phase problem has been solved.

For this study purified CBD- γ was concentrated to final concentrations of 2.8 mM (~20 mg mL⁻¹) or 1.4 mM (~10 mg mL⁻¹) and screened using the commercially available protein crystallisation kits *Index*, *Crystal Screen*, *Crystal Screen 2*, *PEG/Ion*, *PEG/Ion 2*, *Morpheus* and *JCSG+* (all Hampton Research). Screens were performed as sitting drop vapor diffusion setup in 96 Well 3 Drops Sitting Drop Crystallization Plates (Corning) using TTP Labtech's mosquito® Nanolitre Protein Crystallization Robot and continuously monitored in a Rock Imager (Formulatrix) over one month. Further analyses have been performed, including screening of uncleaved GST-PreSci-His6-TEV-CBD- γ , at higher concentration of protein at 5.6 mM (~40 mg mL⁻¹) and at different salt concentrations and pH values using the SaltRx 1 & SaltRx 2 screening kits (Hampton Research).

Also, Lipid Cubic Phase (LPC) experiments, commonly used for membrane proteins, were conducted to determine if the crystallisation of CBD- γ is more favourable in a hydrophobic environment. LPC was setup by mixing lipids with *Index*, *Crystal Screen*, *Crystal Screen 2*, *PEG/Ion*, *PEG/Ion 2*, *Morpheus* and *JCSG+* crystal screens containing freshly purified CBD- γ at 1.4 mM.

Ultimately, analysis using X-ray crystallography was put on hold because of the difficulty in obtaining well-ordered crystals. I chose instead to focus on the analysis of the protein using NMR spectroscopy.

3.11.2 Nuclear Magnetic Resonance Spectroscopy

Unlike X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy is a unique tool to determine the structure, dynamics, and interactions of proteins in solution. In contrast to protein x-ray crystallography, which can determine structures of proteins of any size, the spectral quality in NMR is depending on the molecular tumbling rate of the molecule. For larger proteins (40 kDa and above), the resonance line-width broadens primarily due to the decreased tumbling rate which makes it difficult to determine the structure [17]. A number of methods have been introduced to determine the structure of larger macromolecules, including to run the experiment at elevated temperature to increase the tumbling rate, to dilute the ^1H spins through uniform deuteration and the introduction of TROSY (Transverse Relaxation Optimized Spectroscopy) NMR, which performs a spin manipulation to directly overcome the issue [18-20]. One of the major advantages of NMR spectroscopy however is the ability to determine partially folded structures from proteins which are commonly difficult to crystallize.

In this study, NMR spectroscopy enabled to capture the structural information of CBD- γ , one of the active regions of DgCBP-3 which is critical in order to understand the fundamental structure and functional mechanism in the squid's beak. Various NMR spectroscopy experiments with CBD- γ were executed to understand the structure of the active region. Further, studies of CBD- γ in the presence of PCP were performed to study the role of carbohydrate on the structural folding of the protein and to study the protein-carbohydrate binding mechanisms. All NMR spectroscopy experiments were carried out in Bruker 600/700/800 MHz spectrometers equipped with cryoprobe at 278 K.

3.11.2.1 1-Dimensional Proton NMR

A One-Dimensional (1D) Proton-NMR spectrum is generally the first spectrum to be recorded. It detects the resonance signals for each proton (^1H) in a biomolecule [21]. Generally, 1D ^1H -NMR spectroscopy provides insights about the overall integrity and folding of a polypeptide. While a folded protein demonstrates high dispersion triggered by inner molecular interactions between sidechain interactions within the protein, an unfolded protein commonly establishes poor dispersion as all containing amino acids are exposed to the solvent.

To assess the structural integrity of CBD- γ in this study, 700 μM of protein in gel filtration buffer (20 mM Na HEPES, 500 mM NaCl, pH 6.8), prepared with 10% D_2O and substituted with 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) for signal reference were assessed and reported. The experiments were carried out using a Bruker 700 MHz spectrometer.

3.11.2.2 2-Dimensional Proton NMR

The standard 1D-NMR spectroscopy only provides one frequency dimension which leads to problems due to spectral proton peak overlaps [22]. It can therefore only be used for qualitative assessment of the integrity of a protein. A frequently used tool for protein structure elucidation is the homonuclear ^1H - ^1H correlation spectroscopy (COSY) [23].

In the following, in complement to the original COSY, total correlation spectroscopy (TOCSY), and nuclear Overhauser spectroscopy (NOESY) were acquired to obtain additional information about the protein [24, 25]. Without providing additional details, TOCSY acquisition is using a special mixing and pulse protocol to transfer the magnetisation further along the sequence of coupled protons to obtain a sub-spectrum of coupled protons in a molecule. These protocols are well catalogued and easily applicable [23]. NOESY on the other hand is a powerful tool to identify pairs of protons which are spatially close, even if separated over a long distance. NOESY spectra are acquired with mixing times to build up Overhauser enhancements (NOEs) [26]. While NOESY and TOCSY can increase the resolution of homonuclear spectra, the size of the polypeptide remains a limiting factor. Even though NOESY and TOCSY can be applied for smaller

peptides, at molecular sizes beyond 10 kDa a poor signal-to-noise ratio and decreased number of sensitivity may limit the analysis [22].

In this study, two dimensional ^1H - ^1H TOCSY and ^1H - ^1H NOESY spectra were acquired at 600 MHz with 80 ms and 200 ms mixing time, respectively to determine the protein structure of CBD- γ . Further, the (soluble) chitin derivative PCP ligand was titrated to the protein at molar ratios between 1:0.6 and 1:3.4 to mimic the effect of the presence chito-carbohydrates and to determine the structure of CBD- γ in the protein-carbohydrate complex.

3.11.2.3 Heteronuclear Correlation Spectroscopy

As outlined above, homonuclear NMR spectroscopy has its limitations when assigning amino acid peaks caused by overlapping when correlating only proton isotopes, leading to poor signal-to-noise ratio and dizzying the number of proton cross-peaks.

To overcome the problem, additional isotopic labelling of proteins is a regular method which significantly enhances the sensitivity of NMR spectroscopy. Apart from protons, isotopic labelled proteins typically contain additional magnetic active nuclei, including ^{13}C and ^{15}N among other less frequently used isotopes which can be used for heteronuclear correlation experiments [27, 28]. Here, the additional nuclei support the structure determination process by concatenating additional magnetisation transfers required to determine larger proteins. Unlike protons which are naturally abundant (99.98%), the natural abundance of the nitrogen (0.37%) and carbon (1.108%) isotopes is lower and labelling of the protein requires the protein expression media to be supplemented with isotropically enriched components, namely ^{15}N -ammonium chloride (^{15}N - NH_4Cl) or ^{13}C -glucose [28]. The more isotopes are included in the analysis the more frequency dimensions can be correlated to each other. In the following we included ^{15}N as second magnetic active nuclei to measure the correlation increase the sensitivity of the NMR spectroscopy.

3.11.2.4 2-Dimensional ^1H - ^{15}N -HSQC

^1H - ^{15}N -HSQC (^1H - ^{15}N -Heteronuclear Single-Quantum Coherence) is one of the more frequently heteronuclear NMR spectroscopic methods. It measures the magnetisation that

is transferred from the amide proton (H_N) to the covalently bonded nitrogen. In general, the ^{15}N -HSQC spectrum visualises the correlations of H-N bonds of all amino acids (except Proline) with an additional visibility of tryptophan's N_ϵ - H_ϵ groups and the N_δ - $H_{\delta 2}$ / N_ϵ - $H_{\epsilon 2}$ carboxamides sidechains of asparagine and glutamine [22].

For the following study, CBD γ was expressed in M9 minimal media enriched with ^{15}N - NH_4Cl as outlined in Chapter 3.6 *Expression of DgCBP-3 constructs for Nitrogen-15 NMR Spectroscopy* [5]. Acquisition of the 1H - ^{15}N -HSQC spectrum was performed at 600 MHz to determine the integrity and folding of the protein.

3.11.2.5 3-Dimensional 1H - ^{15}N NOESY-HSQC and 1H - ^{15}N TOCSY-HSQC

Three dimensional 1H - ^{15}N NOESY-HSQC and 1H - ^{15}N TOCSY-HSQC spectra were also acquired at an equimolar protein-to-ligand concentration to facilitate spectral assignment and structure calculations. Amide temperature coefficients were also determined by recording 2D 1H - ^{15}N HSQC spectra of either CBP or CBP with sugar at 278K, 283K, 288K, 293K and 298 K. Amide proton chemical shift deviations were fitted linearly against temperature and the temperature coefficients were calculated as $\sigma\delta_{HN}/\Delta T$ (ppb K $^{-1}$).

3.11.2.6 NMR Spectra Assignment

The NMR spectra were assigned manually in NMRFAM Sparky with the amino acids of CBD- γ identified using their respective spin system in the 2D 1H - 1H TOCSY spectrum [29]. Using the assigned peaks from the 2D 1H - 1H TOCSY spectrum, the 2D 1H - 1H NOESY spectrum was subsequently assigned in NMRFAM Sparky.

The peak list was exported, and the structure calculated using the structure calculation program CYANA 3.1. In order to achieve conformational freedom when calculation the structure, the dihedral angles ϕ and ψ were set to -120° to 120° and 20° to -120° , respectively. Several rounds of structure calculations were performed and refined by manually incorporating additional hydrogen bonds. Out of 100 CBD- γ structures calculated by CYANA 3.1, the 10 lowest energy structures were exported and used for subsequent analysis. All structure representations were made using *PyMOL Molecular Graphics*

System. *Procheck* was used to validate the geometry of the calculated NMR structures by calculating the Ramachandran plot [30-32]. The homo-nuclear assignment of CBD- γ was used to support the assignment of the 3D ^1H - ^{15}N NOESY-HSQC and ^1H - ^{15}N TOCSY-HSQC assignment as outlined in chapter 3.11.2.5 *3-Dimensional ^1H - ^{15}N NOESY-HSQC and ^1H - ^{15}N TOCSY-HSQC*.

3.11.2.7 Chemical Shift Perturbation

In NMR spectroscopy, the addition of an interacting ligand to the protein should result in a complexation-induced change of the chemical shift, also known as chemical shift perturbation (CSP) [33]. The CSP is generally used to determine the location of the active binding site in a protein and can support the structure determination process of the protein-ligand complex. Under good conditions and provided the measurements have been made accurately, the observed shifts of the population-weighted average of the free and bound protein also allow to provide measurements of the K_D for every individual amino acid residue [34].

In this study, multiple ^1H - ^{15}N NOESY-HSQC spectra were recorded for CBD- γ at a concentration of 0.7 mM at 600 MHz over an incremental concentration of PCP ranging between 0.05 mM and 1.2 mM. The difference in exchange rates ($\Delta\delta_{\text{total}}$) is based on the chemical shift dispersion of HN to N and calculated between the free protein ($\Delta\delta_{\text{free}}$) and the protein in complex with PCP ($\Delta\delta_{\text{bound}}$) at a concentration of 0.7 mM. The CSPs were plotted to identify and map the amino acid residues that are in exchange with the chito-oligosaccharide using the formula (Eq. 3):

$$\Delta\delta_{\text{total}} = [(\delta_{\text{bound}})^2 + (0.152 \times \delta_{\text{free}})^2]^{1/2} \quad (3)$$

The weighing factor 0.152 was calculated as ratio of the range of the chemical shift dispersion between HN and N.

For the analysis, only residues for which the weighted shift change is greater than the average $\pm$ standard deviation were taken into account for being actively in exchange with PCP. In addition, those residues for which the weighted shift change is greater than average (mean) $\pm$ twice the standard deviation, were highlighted to be in exchange with PCP.

3.12 Saturation Transfer Difference NMR

Besides other regularly used tools, such as surface plasmon resonance spectroscopy (SPR) or isothermal titration calorimetry (ITC), Saturation Transfer Difference (STD) NMR titration is another method that can be used to map ligand strength. Moreover, STD NMR also provides information that can be used for group epitope mapping (GEM), a method that provides insights in the nature of protein-ligand binding interactions at the atomic level [35]. GEM identifies which proton of a ligand is positioned in the nearest proximity of the protein; an information that is valuable to understand the fundamentals of the binding mechanism.

STD NMR assays are run with an excess of ligand and are measured from the difference of a ^1H NMR spectrum. The difference spectrum is measured with two separate spectra: a) an *on*-resonance spectrum recorded at irradiations on the protein signals at frequencies where no ligand protein resonates to avoid saturation of the ligand, and b) an *off*-resonance spectrum with an irradiation frequency that is significantly different from both the protein's and ligand's frequency region, which represents the unsaturated protein [36].

For CBD- γ , STD NMR titration assays were carried with 50 μM of protein. Different chitin derivatives of different length were titrated onto the protein samples from concentrated stock solutions to minimise the dilution effects. The standard *stdiffesgp.3 pulse sequence* was used with saturation length of 50 ms at a power of 50 db. The on-resonance and off-resonance irradiations were applied at a chemical shift of -1.2 ppm and 40 ppm, respectively. In order to monitor the evolution of the STD amplification factors (STD-AF), saturation times (D20) ranging from 0.5 s to 5.0 s were used. The D1 relaxation delay was set to 5.5 s. Spectra were processed using *stdsplit*. As defined by Mayer & Meyer, the STD intensity (η_{STD}) of a given proton is directly proportional to the fraction of interacting ligand, and hence dependent according to (Eq. 4) [35].

$$\text{STD-AF} = \frac{I_0 - I_{\text{sat}}}{I_0} \times \text{ligand excess} \quad (4)$$

The dissociation constant k_D of the individual protons was obtained by titrating the chitin derivative acetyl-oligosaccharides tetraacetyl-chitotetraose (TCT), pentaacetyl-chitopentaose (PCP), hexaacetyl-chitohexaose (HCT), and heptaacetyl-chitopeptaose (HpCT) to 50 μM of CBD- γ protein with a saturation time D20 of 2 s.

The dissociation constant k_D and the dimension-less scaling factor α_{STD} were yielded by plotting the STD-AF against the increasing ligand concentration $[L]$ followed by mathematical fitting of the resulting curve following a Michaelis-Menten hyperbolic dose-response curve as described by equation 5 (**Eq. 5**) [37]:

$$\text{STD} - \text{AF} [L] = \frac{\alpha_{\text{STD}}[L]}{K_D + [L]} \quad (5)$$

NMR data processing and analysis were carried out using Topspin (Bruker) and NMR FAM-Sparky [43], respectively.

Group epitope mapping was conducted at the highest concentration of PCP at a protein-to-carbohydrate molar ratio of 1:250 at various saturation times (D20) ranging from 0.5s to 5.0s. The D1 relaxation delay was set to 5.5s. The STD-AF of different protons such as H1, H2, H3, HN (H2) and H3C (H2) were derived using the above-mentioned equation (**Eq. 3**) and STD_{max} and K_{st} were calculated by fitting the STD-AF against saturation times using the following equation (**Eq. 6**):

$$\text{STD} - \text{AF} = \text{STD}_{\text{max}}(1 - e^{-k_{\text{st}} \times t_{\text{sat}}}) \quad (6)$$

Subsequently, STD_{total} was calculated using the following equation (Eq. 7):

$$STD_{Total} = STD_{max} \times K_{st} \quad (7)$$

Lastly, the proton that recorded the highest value for STD_{total} was considered as to binding 100% to PCP in accordance to the epitope mapping rule in order to derive to the binding contribution of all remaining protons of CBD- γ towards PCP.

3.13 Predictive Ligand Docking Using Schrödinger's Glide

The following experiment has been carried out by Akshita Kumar and Chandra S. Verma at the Bioinformatics Institute (A*STAR).

The 3-dimensional structure of PCP was taken from ChemSpider (CSID27445378). The structure was minimised using the MacroModel module employing OPLS 2005 force field in the Schrodinger 9.0. The parameters for PCP were generated using antechamber and further subjected to energy minimisation and molecular dynamics simulations using the ff14SB forcefield in AMBER18 to obtain its low energy conformations [38]. Clusters were generated from the molecular dynamics simulation and the conformation from the highest populated cluster selected for docking analysis.

The prepared PCP ligand was docked onto an energy minimised structure (from MD) of CBP residues (Y₁₆₄, N₁₇₀, Q₁₇₆) using Glide. A box size of 10 × 10 × 10 Å was used which corresponds to the “inner-box” in Glide. The search space was confined to the centre of the aforementioned residues for docking PCP. For grid generation, default Glide settings were used. Subsequently, a rigid receptor docking protocol was used which fixes the protein’s conformation but allowed the ligand to remain flexible. As a result, five docked conformations were identified, and each subjected to molecular dynamics simulations. The simulations were run for 100 ns using the standard protocol in AMBER18. Trajectories, in which the ligand was bound stably, were selected for further analyses. Clustering was done on three molecular dynamics trajectories in which the ligand reported was stably bound. The top clusters populated with more than 25 - 30% of sampled conformations were assessed and the interactions which are located with the closest distance to the centroid of the highly populated cluster (RMSD 0.5-0.6 Å) analysed, respectively.

3.14 Amide Proton Temperature Coefficient

Despite being easily acquired, amide proton temperature coefficients are seldom used in protein structure analyses. The amide proton temperature coefficients are calculated from the 15N-HSQC spectrum using the following equation (**Eq. 8**):

$$\frac{\Delta\sigma_{HN}}{\Delta T} \quad (8)$$

Particularly, the temperature shift may result from a decreased population of folded proteins upon heating [39]. In proteins with higher population of folded structure, the common practice suggests that the hydrogen bond is sequestered from the solvent when the $\Delta\delta_{HN}/\Delta T$ values are less negative than -4.6 ppb/°C [40].

15N-HSQC spectra at 600 MHz were recorded for 0.7 mM of CBD-γ in absence and in equimolar presence of PCP as described by Anderesen *et al* [41]. Temperature gradients were recorded at 5°C intervals between 5°C (278 K) and 30°C (303 K) in the pH adjusted

gel filtration buffer (pH 6.7) and supplemented with 10% D₂O and DSS for signal reference. The amide proton chemical shift deviations were fitted linearly against temperature and the temperature coefficients were calculated as $\sigma\delta\text{HN}/\Delta T$ (ppb K⁻¹). Temperature gradients were expressed in ppb °C⁻¹ with negative values indicating an upfield shift and positive values indicating a downwards shift, respectively.

Standardised random coil values acquired and published by Wishart *et al* were used as reference values for analysis [42]. NH shifts of each amino acid of CBD- γ in absence and presence of pentaacetyl-chitopentaose were plotted for comparison.

3.15 References

1. Structural Genomics, C., et al., Protein production and purification. Nat Methods, 2008. 5(2): p. 135-46.
2. Gräslund, S., et al., The use of systematic N- and C-terminal deletions to promote production and structural studies of recombinant proteins. Protein Expr Purif., 2008. 58(2): p. 210-221.
3. Vaughn, J.L., et al., The establishment of two cell lines from the insect *Spodoptera frugiperda* (Lepidoptera; Noctuidae). In Vitro, 1977. 13(4): p. 213-7.
4. Parks, R.E., Jr., et al., Purine metabolism in primitive erythrocytes. Comp Biochem Physiol B, 1973. 45(2): p. 355-64.
5. Miller, J.H., Experiments in molecular genetics, ed. J.H. Miller. 1972, New York: Cold Spring Harbor.
6. Wallace, B.A. and R.W. Janes, Modern techniques for circular dichroism and synchrotron radiation circular dichroism spectroscopy. Advances in biomedical spectroscopy. Vol. 1. 2009, Amsterdam: IOS Press BV.
7. Greenfield, N.J., Using circular dichroism spectra to estimate protein secondary structure. Nat Protoc, 2009. 1(6): p. 2876–2890.

8. Kelly, S.M. and N.C. Price, The use of circular dichroism in the investigation of protein structure and function. *Curr Protein Pept Sci*, 2000. 1(4): p. 349-384.
9. Venyaminov, S., et al., Circular dichroic analysis of denatured proteins: inclusion of denatured proteins in the reference set. *Anal Biochem*, 1993. 214(1): p. 17-24.
10. Holzwarth, G. and P. Doty, The Ultraviolet Circular Dichroism of Polypeptides. *J Am Chem Soc*, 1965. 87: p. 218-28.
11. Greenfield, N. and G.D. Fasman, Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry*, 1969. 8(10): p. 4108-16.
12. Savitzky, A. and M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures. *Anal Chem*, 1964. 36(8): p. 1627-1639.
13. Pantoliano, M.W., et al., High-density miniaturized thermal shift assays as a general strategy for drug discovery. *J Biomol Screen*, 2001. 6(6): p. 429-40.
14. Lo, M.C., et al., Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery. *Anal Biochem*, 2004. 332(1): p. 153-9.
15. Rupesh, K.R., A. Smith, and P.E. Boehmer, Ligand induced stabilization of the melting temperature of the HSV-1 single-strand DNA binding protein using the thermal shift assay. *Biochem Biophys Res Commun*, 2014. 454(4): p. 604-8.
16. Alberts, B., A. Johnson, and J. Lewis, *Molecular Biology of the Cell*. Vol. 4. 2002, New York: Garland Science.
17. Kleckner, I.R. and M.P. Foster, An introduction to NMR-based approaches for measuring protein dynamics. *Biochim Biophys Acta*, 2011. 1814(8): p. 942-68.
18. Foster, M.P., C.A. McElroy, and C.D. Amero, Solution NMR of large molecules and assemblies. *Biochemistry*, 2007. 46(2): p. 331-40.
19. Pervushin, K., et al., Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of

very large biological macromolecules in solution. Proc Natl Acad Sci U S A, 1997. 94(23): p. 12366-71.

20. Sattler, M. and S.W. Fesik, Use of deuterium labeling in NMR: overcoming a sizeable problem. *Structure*, 1996. 4(11): p. 1245-9.
21. Cavanah, J., et al., ¹H NMR Methods, in *Protein NMR Spectroscopy*, n. Edition, Editor. 2007, Burlington: Academic Press. p. 405-532.
22. Cavanah, J., et al., *Protein NMR Spectroscopy: Principles and Practice*. 2010, Academic Press.
23. Breton, R.C. and W.F. Reynolds, Using NMR to identify and characterize natural products. *Nat Prod Rep*, 2013. 30(4): p. 501-24.
24. Bax, A. and D. Davis, MLEV-17-based two-dimensional homonuclear magnetization transfer spectroscopy. *Journal of Magnetic Resonance*, 1985. 1969: p. 355-360.
25. Neuhaus, D. and M.P. Williamson, *The Nuclear Overhauser Effect in Structural and Conformational Analysis*. Vol. 2nd Edition. 2000, New York: Wiley-Vch.
26. Macura, S.I., et al., Two-dimensional chemical exchange and cross-relaxation spectroscopy of coupled nuclear spins. *Journal of Magnetic Resonance*, 1981. 43(2): p. 259-281.
27. Czarnik, A.W. and H.-Y. Mei, Strategy and Drug Research, in *Comprehensive Medicinal Chemistry II*, J.B. Taylor and D.J. Triggle, Editors. 2007, Elsevier Science: New York.
28. Friebolin, H., *Ein- und zweidimensionale NMR-Spektroskopie*, ed. L. Ernst. Vol. 4. 1989, Weinheim: VCH Verlagsgesellschaft.

29. Lee, W., M. Tonelli, and J.L. Markley, NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics*, 2015. 31(8): p. 1325-7.
30. Laskowski, R.A., et al., AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR*, 1996. 8(4): p. 477-86.
31. Laskowski, R.A., et al., PROCHECK - a program to check the stereochemical quality of protein structures. *J. App. Cryst.*, 1993. 26: p. 283-291.
32. Schrodinger, L., The PyMOL Molecular Graphics System. 2015: Version 2.0 Schrödinger, LLC.
33. Williamson, M.P., Using chemical shift perturbation to characterise ligand binding. *Prog Nucl Magn Reson Spectrosc.*, 2013. 73: p. 1-16.
34. Baxter, N.J. and M.P. Williamson, Temperature dependence of ¹H chemical shifts in proteins. *J Biomol NMR*, 1997. 9(4): p. 359-69.
35. Mayer, M. and B. Meyer, Group epitope mapping by saturation transfer difference NMR to identify segments of a ligand in direct contact with a protein receptor. *J Am Chem Soc*, 2001. 123(25): p. 6108-17.
36. Bhunia, A., S. Bhattacharjya, and S. Chatterjee, Applications of saturation transfer difference NMR in biological systems. *Drug Discov Today*, 2012. 17(9-10): p. 505-13.
37. Haselhorst, T., A.C. Lamerz, and M. Itzstein, Saturation transfer difference NMR spectroscopy as a technique to investigate protein-carbohydrate interactions in solution. *Methods Mol Biol*, 2009. 534: p. 375-386.
38. Case, D.A., et al., AMBER 18, ed. U.o. California. 2018, San Francisco.
39. Andersen, N.H., et al., Extracting Information from the Temperature Gradients of Polypeptide NH Chemical Shifts. 1. The Importance of Conformational Averaging. *J. Am. Chem. Soc.*, 1997. 119(36): p. 8547-8561.
40. Cierpicki, T. and J. Otlewski, Amide proton temperature coefficients as hydrogen

bond indicators in proteins. *J Biomol NMR*, 2001. 21(3): p. 249-61.

41. Andersen, N.H., et al., Conformational isomerism of endothelin in acidic aqueous media: a quantitative NOESY analysis. *Biochemistry*, 1992. 31(5): p. 1280-95.

42. Wishart, D.S., B.D. Sykes, and F.M. Richards, The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry*, 1992. 31(6): p. 1647-51.

43. Lee W., Tonelli M., Markley J.L, NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics*. 2015. 31(8): p. 1325-7

Chapter 4 Results

4.1 Introduction

As previously outlined in *Chapter 2*, *D. gigas*' Chitin-Binding Beak Proteins (DgCBPs) play an essential role in the formation of the squid beak. They have been proposed to interact and self-assemble with chitin into a porous scaffold that is subsequently permeated and filled with histidine-binding beak protein (DgHBPs) droplets. Upon permeation DgHBPs start to cross-link, leading to sclerotization and hardening of the beak. Yet, the underlying binding mechanism that is involved in the complexation of DgCBPs and chitin remains unknown.

The following project uses a bottom up approach to report the 3D structure and molecular level interactions of a chitin-binding domain (CBD- γ) located in a CBP from the beak of *D. gigas* which is believed to interact with chitin to form a scaffold network that is infiltrated with a second set of structural proteins during beak maturation. NMR spectroscopy was used to elucidate the molecular interactions between CBD- γ and PCP and find that folding of this domain is triggered upon its interaction with PCP and molecular dynamics studies performed to propose a binding model based on the NMR results.

Hitherto, none of the DgCBPs has been studied extensively. Initial studies on the Humboldt squid conducted in 2010 only uncovered the presence of DgHBPs which were believed to contain an active site to bind chitin [1]. In 2015, DgCBPs were classified to be the second family of proteins present in the squid's beak [2]. Unlike DgHBPs, they do not vary in abundance between the soft region embedded in the buccal mass of the rostrum and the hard rostrum. It has been suggested that the binding domains of DgCBPs share sequence and structural similarity to proteins containing the carbohydrate-binding module-14 (CBM-14), found in Peritrophin-A and in peritrophic matrix proteins of insect chitinases. CBM-14 reportedly contains six cysteines that form three disulphide bonds [3, 4]. Using these initial findings and assumptions the project initially focussed on the expression and purification of the CBDs of all DgCBPs of *D. gigas*. Ultimately, one construct including one of three proposed binding domains to chitin from DgCBP-3, namely CBD- γ , was

identified as being soluble and optimised. The final construct showed a medium high expression with high solubility, suitable for subsequent studies. The determined structure was firstly compared to existing structures of two proteins, tachycitin and hevein (**Figure 4-1**), but eventually solved in a de-novo approach.

Figure 4-1 | Structures of Tachycitin and Hevein.

Depiction of two previously published structures, tachycitin (**A**) and hevein (**B**) structures, containing the the proposed CBM-14 (**taken and adapted from Suetake et. al [4]**).

Lastly, the CBD- γ was examined in absence and in complex with small soluble analogues of its natural polymeric ligand, chitin, to derive a conclusive model of the chitin-to-DgCBP binding in *Dosidicus gigas*. The studies with soluble chitin derivatives allowed not only to predict the affinity between chitin and the protein but also to make an assumption on the orientation adopted by the ligand to bind to the protein.

4.2 Protein Construct Design of DgCBPs and Small-Scale Purification

Based on previous studies performed by Tan *et. al.*, 74 full-length constructs of DgCBP-1, DgCBP-2, DgCBP-3 and DgCBP-4 without their signal peptides were designed but did not express in the soluble fraction. Alternatively, constructs of their CBDs alone were designed, cloned, and expressed in various *E. coli* strains, including BL21-Rosetta T1R, BL21(DE3)pFH255, and BL21(DE3)pFH256 among others, or in insect cells, namely *Spodoptera frugiperda*'s sf9 and High Five cells. All constructs were done at the Protein Production Platform (PPP) at NTU's School of Biological Sciences [2].

Cells were grown in small batch expression, homogenised and centrifuged to separate cell fragments from the soluble protein fraction. The supernatants were analysed using SDS-PAGE to identify the soluble protein content (SDS-PAGES not

Figure 4-2 | Soluble Fraction of CBP-3 Constructs

Small batch purification analysis of 22 protein constructs of DgCBP-3 2. The construct pNIC-28-BSA- Δ His6-TEV-DgCBP-3 Δ R141-V201 with a medium high expression of protein in the soluble fraction (yellow) was further optimised by using the plasmids pNIC-GST and pNIC-CH2, which resulted in an increased yield of protein in the soluble fraction after lysis and ultracentrifugation (green).

shown). While none of the DgCBP-1, DgCBP-2 and DgCBP-4 constructs could be expressed in the soluble fraction, the construct pNIC-28-BSA-His₆-TEV-DgCBP-3_{R141-V201} showed a medium-high expression using BL21-Rosetta T1R cells (yellow bar) (**Appendix 6-3, Figure 4-2**). The construct was used as a template to further optimise the protein expression. Subsequently, the insert of DgCBP-3 Δ R141-V201 was cloned into two plasmids pNIC-GST and pNIC-CH3 to either include an N-terminal tag containing a cleavable GST-tag followed by a cleavable His₆-tag or a non-cleavable C-terminal His₆-tag, respectively. Both optimisations resulted in a significant improvement with respect to the solubility of the final DgCBP-3 Δ R141-201 (CBD- γ) eluent (**Table 4-1**).

4.3 Purification of CBD- γ

The identified constructs were expressed in “pilot” batches and purified using appropriate protein purification methods. While the pNIC-GST construct was purified using a three-step purification with additional TEV cleavage before the final polishing using gel filtration, pNIC-CH2 was purified using a two-step purification, consisting of a nickel chromatography and subsequent gel filtration. Ultimately, the final yield of CBD- γ from the purification using the pNIC-GST plasmid exceeded the pNIC-CH2 construct by more than 30% and was used for all subsequent studies (**Figure 4-3**).

Table 4-1 | Forward and Reverse Primers of the CBD- γ Insert

FWD	R141	TACTTCCAATCCATGCGCACCGAGTGCAATACTG
REV	V201	TATCCACCTTTACTGTCAGACCTTACAGCTTGGCACAT

Figure 4-3 | Construct Design of CBD- γ

(A) Segmentation of the full-length DgCBP-3 protein with its signal peptide (green) and three chitin-binding domains CBD- α - γ (B) Final construct design of CBD- γ with double cleavable N-terminal GST and His₆-tag. (C) Protein sequence of the final TEV-cleaved protein construct with cleavage fragments (QSM).

Full scale purification was performed as outlined in *Chapter 3 → Experimental Methodology*. The low content of aromatic amino acids in the protein sequence (2% of Tyrosine) resulted in a low absorbance at 280nm (**Figure 4-4; Chromatogram**) during size exclusion chromatography. The chromatogram indicated three major peaks of which the first and second responded to the GST-tag that has not been separated during re-chromatography and TEV protease, respectively. The third peak that spans the elution volume from 80 mL to 100 mL constituted CBD- γ with high purity as indicated by SDS-PAGE analysis (**Figure 4-4; SDS PAGE**). It appeared that the protein expressed was highly

unstable and had to be used within 24 hours after purification. Neither shock-freezing and storing at -80°C nor supplementing the final protein sample with 20% glycerol could improve the lifespan of the stored protein. Consequently, all subsequent experiments were conducted timely after successful purification.

Figure 4-4 | Gel Filtration Profile and SDS-PAGE of CBD- γ

(Chromatogram) The gel filtration was monitored using an NGC Quest™ 10 Chromatography System at two wavelengths, UV260 and UV280. Due to the low amount of aromatic amino acid residues in the protein, only low absorption was recorded using the UV280 filter. **(SDS-PAGE)** The SDS-PAGE confirms a high content of the full construct of GST-PreSci-His₆-TEV-CBD- γ in the sample after GST affinity chromatography, running at about 80kDa (left band). After cleavage and subsequent gel filtration the final construct of CBD- γ collected from 20mL of elution volume, appears at a molecular size of about 10kDa (based on Thermo Scientific™ Spectra™ Multicolor Broad Range Protein Ladder 10 – 260 kDa).

4.4 Primary Sequence Analysis of CBD- γ

Protein sequence analysis of CBD- γ (without the fused GST-PreSci-His6-TEV-tag) was performed to understand the protein's features using *ExPASy's* ProtParam tool [5]. According to the analysis, CBD- γ consists of 64 amino acids with a theoretical molecular weight of 7.05 kDa. With seven positively charged amino acids (Arginine, Lysine) and six negatively charged amino acids (Aspartic acid and Glutamic acid) CBD- γ has a slightly basic isoelectric point at 7.77. The protein consists to 44% of hydrophilic / polar residues with Ser (12.5%), Cys (9.4%) and Thr (9.4%) being the most abundant amino acids.

With a high instability index being calculated (70.54) the protein was classified to be highly unstable.

Under the assumption that CBD- γ shares sequence similarity with CBM-14, its sequence was compared to Peritrophin A domain of *A. duodenale* (Figure 4-5). Sequence similarity identified a high similarity in the C-terminal region between the two sequences. Additionally, CBD- γ was compared to the original R&R Consensus (*G-x(8)-G-x(6)-Y-x-A-x-E-x-GY-x(7)-GY-x(7)-P-x-P*), initially published by Rebers and Riddiford [6]. The R&R Consensus represents a conserved motif of amino acids which has been reported in a variety of insect cuticle proteins. The analysis found that CBD- γ shares a significant similarity to the extended R&R Consensus *RR-2*, that shares the consensus of *G-x(8)-G-x(6)-Y-x(6)-GF-x(7)-P* [7, 8]. In contrast to the published extended *RR-2* Consensus that is associated with the hard/rigid cuticle in insects, a QY sequence in CBD- γ that replaces the reported GF amino acid sequence (Figure 4-5).

```

R&R_consensus      - - - - G x x x x x x x G x x (x) x x x Y x x D x x G Y x x x x x x x x P
DgCBD-γ            138 SKFRTECNTARGKGHM - LIAYPGDCSQYISCD S N D Q S P Q Q C A S G T V F N S E K Q R C D F R A N V P S C K V 202
O. bimaculoides (hyp. Protein) 72 SKFRSECDVARGKGH I - L I T Y P G D C S Q Y I S C D S N D Q N Q R C A S G T V F N P D K Q R C D F R S N V P S C K L 136
O. vulgaris (obst-E) 164 SKFRSECNVARGKGH I - L I T Y P G D C S Q Y I S C D S N D Q T P Q Q C A S G T V F N P A K Q Q C D F R S N V P S C K L 229
A. duodenale (Peritrophin-A domain) 281 D D D K G V I N G C R G A G H G D L I S H E S D C Q Q F Y R C V W D R L E K M H C P S G T V F N P E L S V C D Y P A N V P Q C A H 344
H. americanus (AMP1B) 42 S N H Q G S F R F T L E D G T I A E V T Y I A D E Y G Y Q P S S D L L P V P P P A P P H V Q R L L E I A A E Q R A Q G I T F D 105
L. polyphemus (Cuticle protein 14) 50 N V K V G S Y G L T D A D G R R R V V T Y K A D A S G F N A N V H T N E P G T D S S K D P A N T L V N K A V L P T T Y Y G G Y A P 124

```

Figure 4-5 | Homology matching of CBD- γ with the R&R Consensus and Peritrophin-A Domain

Sequence comparison of CBD- γ to the original R&R Consensus [6] with two proteins with a sequence similarity >80%, the Peritrophin-A domain of *A. duodenale* and two cuticle proteins found in *H. americanus* and *L. polyphemus*. Alignments with the R&R Consensus and the Peritrophin-A domain are highlighted in green and red, respectively. Similarity analysis identified high sequence similarity of CBD- γ to the R&R Consensus in the N-terminus and middle region of the protein and to C-terminus of the Peritrophin-A domain. The two protein a similarity of >80% to CBD- γ , a hypothetical protein found in the octopus *O. bimaculoides* (KOF83921.1) and an obstructor-E-like protein found in *O. vulgaris* (XP_029637511.1), share high similarity with both the R&R Consensus and the Peritrophin-A domain. In contrast, two cuticle proteins, namely AMP1B from *H. americanus* (P81385) and the cuticle protein 14 from *L. polyphemus* (P83355), feature only similarity with the middle region of CBD- γ and the original R&R Consensus.

Additionally, BLAST sequence analysis was conducted [9]. The analysis revealed two proteins with a sequence similarity >80%, namely a hypothetical protein found in the octopus *O. bimaculoides* (KOF83921.1) and an obstructor-E-like protein found in *O. vulgaris* (XP_029637511.1). Those two proteins share not only a high similarity with the original R&R consensus sequence (**Figure 4-5; sequence highlighted in green**), but also with the Peritrophin-A domain of *A. duodenale* (**Figure 4-5; sequence highlighted in red**). In addition, two other proteins, namely AMP1B from *H. americanus* (P81385) and the cuticle protein 14 from *L. polyphemus* (P83355), only high similarity with the middle region of CBD- γ and the original R&R Consensus.

4.5 Identification of Secondary Structure Elements Integrity of CBD- γ Using Circular Dichroism Spectroscopy

The conformation stability of the recombinant CBD- γ in the absence and presence of PCP was assessed using circular dichroism (CD) spectroscopy in the following buffer: 20 mM HEPES, pH 6.8, 500 mM NaCl). The samples were measured in triplicates and the ellipticity curve smoothened using the Savitzky-Golay filter.

Figure 4-6 | Circular Dichroism of CBD- γ in the absence and presence of PCP

The secondary structure characteristics of CBD- γ were examined using circular dichroism (CD) spectroscopy in the absence (black line) and presence (red line) of PCP to assess structural changes the protein undergoes at an incremental ligand concentration. In the absence of PCP the ellipticity exposes a minimum at 200 nm and a maximum at 192 nm. In presence of PCP CBD- γ exhibits a minimum at 205 nm, a second minimum at 214 nm, and a shift of the maximum from 192 nm to 197 nm.

The smoothed CD spectrum of the free CBD- γ displayed a single minimum of the ellipticity curve at 200 nm and a single maximum at 192 nm, suggesting a protein structure that contains β -sheets. In comparison, the smoothed CD spectrum for 700 μ M CBD- γ in the presence of 500 μ M of PCP indicated a shift of the β -sheet minimum from 200 nm to 205 nm, the emergence of an additional marginal minimum at 214 nm, and a shift of the maximum from 192 nm to 197 nm, suggesting that the protein underwent a structural change to a structure that contained both α -helices and β -sheets in the complex (**Figure 4-6**).

In addition, an analysis of CBD- γ at different pH values ranging from 6.5 to 7.5 to corroborate the sensitivity of the protein during NMR spectroscopy did not result in significant changes of the spectrum, leading to the assumption that the protein did not show any significant destabilisation in a weak acidic environment.

4.6 Qualitative Binding Analysis of CBD- γ to Chitin Magnetic Beads

A qualitative binding experiment of CBD- γ to chitin magnetic beads with subsequent analytical FPLC was performed to assess its activity. Three positive samples containing 0.625 μ g, 1.25 μ g or 5.0 μ g of CBD- γ and one negative control containing 5.0 μ g of Glutathione S-transferase (GST) were incubated with 100 μ L of chitin magnetic beads for 2h at 4°C (**Figure 4-7**). The same amounts of protein were kept as references. The supernatants were collected by spinning the samples through a filtering plate and analysed using analytical FPLC.

In the presence of chitin magnetic beads, we observed a strong drop in absorption of CBD- γ in FPLC, verifying that the protein actively bound to chitin (**blue line, Figure 4-7; A, B, C**). That not all of the protein bound to the beads leads to the assumption that either the beads were saturated or that the protein was partially unfolded. In contrast, the negative control did not show any absorption drop, corroborating that the beads did not bind non-specifically to GST (**Figure 4-7 D**).

Figure 4-7 | Binding Analysis of CBD- γ Using Analytical FPLC

Binding of CBD- γ to chitin magnetic beads. A defined amount of protein was loaded onto 100 μ L of chitin magnetic beads and incubated for 2h at 4°C and the supernatant analysed via analytical FPLC using a Superdex 75 10/300 column to measure the amount of unbound protein in the sample. All positive samples (A-C) indicated that approximately 45% of protein were bound to the beads. As a reference, the same amount of protein was kept under same conditions without adding the chitin magnetic beads. 5 μ g of GST was used as a negative control, showing no binding (D).

4.7 Ligand-Receptor Binding Affinity Determination Using STD NMR

One method to quantify the binding of PCP to CBD- γ is by using saturation transfer difference (STD) NMR spectroscopy, which has the advantage over other methods, such as isothermal titration calorimetry (ITC) or surface plasmon resonance spectroscopy (SPR) that it only utilises small amounts of protein but yet provides

Figure 4-8 | Saturation Transfer Difference NMR Spectroscopy of CBD- γ in Presence of PCP

1D STD-reference and difference NMR spectra of CBD- γ :PCP ($t_{\text{sat}} = 2.0$ s) with different proton resonances (H1, H2, H3, H4, H5, NH(H2) and CH₃ (H2)) of PCP labelled. **(B)** STD amplification factor of five different proton resonances (H1, H2, H3, NH(H2) and CH₃ (H2)) as a function of increasing saturation times ($t_{\text{sat}} = 0.5$ s, 1 s, 1.5 s, 2 s, 3 s, 4 s, 5 s) to calculate STD_{max} and K_{st} from the equation, $\text{STD-AF} = \text{STD}_{\text{max}}(1 - e^{-K_{\text{st}} \cdot t_{\text{sat}}})$. **(C)** Stacked 1D STD-difference NMR spectra of acetyl group (CH₃) of the acetamide region at a saturation time $t_{\text{sat}} = 2.0$ s, for increasing PCP concentrations. (Inset: fitting of STD amplification factor (STD-AF) against increasing PCP concentrations to obtain K_D). **(D)** Group epitope mapping of one chitin subunit of PCP based on the $\text{STD}_{\text{total}}$ value calculated from the equation $\text{STD}_{\text{total}} = \text{STD}_{\text{max}} \cdot K_{\text{st}}$. The proton that exhibits the highest $\text{STD}_{\text{total}}$ value was defined as 100% and all other protons are normalized with respect to the highest $\text{STD}_{\text{total}}$.

high sensitivity and robustness [10]. Applying an equation described by Angulo *et al*, STD NMR spectroscopy titration can deliver the equilibrium dissociation constant K_D accurately if the binding isotherm is constructed as a function of the ligand concentration using the initial growth rates of the STD amplification factors (STD-AF_0) [11].

The dependence of the k_D on the saturation time of the STD-NMR experiment was analysed using 40 μM of CBD- γ between 0.5 seconds and 5 seconds (**Figure 4-8; A**). Assessing the saturation τ_{sat} is important, since the intensity of the peaks and the sensitivity of the experiment are directly correlated. The best signal-to-noise ratio was achieved at $\tau_{\text{sat}} = 2.0$ seconds, which was subsequently used for ligand titration.

For ligand titration and k_D determination tetraacetyl-chitotetraose (TCT), PCP, hexaacetyl-chitohexaose (HCH), and heptaacetyl-chitoheptaose (HpCH) were investigated. While TCT and PCP were soluble at a concentration of 10 mM, HCH and HpCH were only soluble until a molar concentration of 8 mM. Therefore, the titration was conducted for concentration between 0.5 mM and 10 mM for TCT and PCP but only up to 8 mM for HCH and HpCH. The *Stdsplit* macro was used to generate a STD-NMR spectrum and a reference spectrum.

For peak assignment, pre-defined chemical shifts of chitin protons published by *Kasaai* were used [12]. The absolute intensity of the signals at 2.1 ppm, referring to the CH_3COOH (AcCH_3) peak, the NH peak (8.5 ppm) and the H1 peak (4.65 ppm) increased with

Table 4-2 K_D of acetyl-chitooligomers determined using STD-NMR spectroscopy	
	K_D (CH_3) (μM)
(GlcNAc) ₄ (TCT)	15.3
(GlcNAc) ₅ (PCP)	15.0
(GlcNAc) ₆ (HCH)	14.7
(GlcNAc) ₇ (HpCH)	10.3

increasing ligand concentration. Next, from the STD amplification factor for the five interacting protons H1, H2, H3, NH (H2) and CH_3 (H2) of PCP at saturation times ranging from 0.5 s to 5 s, the STD total values were calculated. With STD total values of 21.5 and 20.8, respectively, H1 and H2 protons were the most exhibited, followed by NH (19.9) and CH_3 (17.6) protons of H2 and H3 (11.6) (**Fig. 4-8; B**). Next, the STD amplification factors (STD-AF) for all AcCH_3 peaks of TCT, PCP, HCH and HpCH were calculated (as described in Eq. 5 in Chapter 3) and plotted against the ligand concentrations (**Figure 4-9**). The k_D of the proteins was determined using a hyperbolic Michaelis-Menten type fit. Comparing the k_D of the AcCH_3 peaks demonstrates that the k_D decreases by 33% from 15.3 μM for TCT to 10.05 μM for HpCH (**Table 4-2**), reflecting the binding is stronger for longer

oligomers.

Lastly, group epitope mapping was done to identify the binding of PCP- γ based on the STD_{total} value calculated (**Figure 4-8; C-D**). It was possible to identify that only H1, H2, H3, HN and H₃C protons of PCP were responsible for the binding to CBD- γ , which is in line with observed protein-carbohydrate interactions reported for the hevein-chitin complex [13]. In addition, H1 and H2 exhibited the strongest STD effect of protons. This can be attributed to its interaction with the exposed polar residues of the protein [17]. In accordance to group epitope mapping principle, the H1 protons were therefore scaled to 70-100% binding and all the other protons are mapped with respect to H1 (**Figure 4-8; D**).

Figure 4-9 | K_D Value Determination of the Binding of Chitooligomers to CBD- γ

Plotting of the STD-AF vs the ligand concentrations for CH₃COOH (blue), NH₃ (red), H₁ (grey) to determine the proton K_D values. (A) TCT, (B) PCP, (C) HCH, and (D) HpCH.

4.8 One-Dimensional Homonuclear NMR

As discussed Chapter 4.7, carbohydrates exceeding the length of PCP have a solubility lower than 10 mM. Therefore, all subsequent complexation experiments were carried out using PCP.

After qualitatively and quantitatively assessing the folding and binding activity of CBD- γ to chito-oligomers, it is instructive to confirm the above findings by conducting ^1H -NMR of the protein without and with the carbohydrate to determine the structure of the protein. One-dimensional homonuclear NMR is the easiest form of NMR and provides information about the integrity of a protein. To a certain extent 1D NMR can also be used to qualitatively identify structural changes in macromolecules.

From the resonance obtained without the carbohydrate, a good dispersion with sharp peaks in the fingerprint amide region was seen, which is generally observed for proteins with structural propensity (**Figure 4-10, A**). That is because of the exposure of amino acid residues to different microenvironments based on their position in the fold of the protein. With the addition of PCP at an equimolar protein-to-carbohydrate ratio, the dispersion of the intensity of the peaks in the fingerprint amide region increased, and additional peaks appeared in the spectrum (**Figure 4-10, B**). Two peaks observed at approximately 8.5 ppm and 2.0 ppm are likely related to the additional content of NH and CH₃ in the sample arising from PCP (**Figure 4-10 B (1), (4)**). In addition, a peak split was observed between 7.0 ppm and 7.5 ppm and an additional peak appeared at 4.6 ppm. Those changes corroborate that the protein underwent a structural change upon addition of PCP. Consequently, these findings suggest that CBD- γ underwent conformational changes upon the addition of the PCP carbohydrate.

Figure 4-10 | ^1H -NMR Spectroscopy in Presence and Absence of PCP

(A) in absence of PCP, CBD- γ expresses a dispersed spectrum in the protein's fingerprint amide region that relates to the internal structure of the protein. **(B)** Upon addition of PCP, the spectrum shows changes, including an increased resonance in the NH region (1), a more intense dispersion of peaks in the fingerprint amide region (2), a new peak in the methyl region (3) and a more intense resonance of the CH₃ peak (4).

4.9 Protein Stability Assessment Using Two-Dimensional ^1H NMR

Protein stability results from the net balance of intramolecular forces, which determine the native folding of a protein. The native structure of the protein is stabilised by intramolecular interactions, including hydrogen bonding, hydrophobic interaction (van der Waals bonds) and disulphide bonds, among others [14].

From the two-dimensional ^1H NMR obtained for CBD- γ in presence of PCP, strong resonance peaks in a well dispersed spectrum were observed (**Figure 4-11, A**). Triggering unfolding of CBD- γ using 8 M Urea indicated that the proteins integrity partially remained, while parts of the protein were denaturated. A second unfolding step, initiated by the addition 20 mM of dithiothreitol (DTT) depleted most of the resonance signals, however with a few remaining peaks (**Figure 4-11, B,C**). Conducting the same experimental approach without PCP resulted in much less resonance peaks after the addition of Urea and no resonance peaks after the addition of DTT. Those results support previous assumptions that the protein is more folded in the presence of PCP

Figure 4-11 | Unfolding of CBD- γ in Presence and Absence of PCP

(A-C) In presence of PCP, CBD- γ shows strong resonances which remained distinct even after addition of 8M urea. Adding 20 mM DTT, most of the resonance peaks dispersed and only a few remained. (D-F) In the absence of the PCP carbohydrate, CBD- γ displays less resonance peaks, with many peaks disappearing upon addition of urea. With the addition of DTT, all resonance signals strongly decreased in intensity, suggesting that CBD- γ is fully denatured.

4.10 Two-Dimensional Homonuclear NMR

After acquiring one-dimensional spectra, two dimensional homonuclear NMR spectra were acquired to obtain tertiary structure information for CBD- γ . The most common form of two dimensional homonuclear NMR spectra are correlation spectra. Correlation spectra provide information on cross peaks of coupled protons which makes it a useful tool to identify larger interconnected protons in macromolecules. They further provide insights on the folding of the proteins, to identify intermolecular interactions between ligands and macromolecules, and, to a certain extent, to calculate structures of peptides and small proteins.

For the analysis of CBD- γ , 0.7 mM of protein was added to incremental concentrations of PCP ranging from 0.4 mM to 2.4 mM. During the titration, a series of 1D and 2D ^1H - ^1H TOCSY and NOESY experiments were recorded. With increased concentration of PCP, peak intensities evolved, and minor peak broadening was observed, suggesting a tight interaction between CBD- γ and the carbohydrate at a low k_D value.

Without PCP, the 2D ^1H - ^1H TOCSY of CBD- γ only showed a limited number of resonances, signalling that the protein is likely to be dynamic and therefore less likely to be measured unbiasedly in the NMR timescale (**Figure 4-10, A**). Interestingly, PCP-related proton resonance peaks appeared with saturation of CBD- γ at an equimolar protein-to-carbohydrate ratio. Additionally, 2H and 3H proton peaks resonating with GlcNAc subunits of PCP showed well resolved peaks at around 8.4 ppm. Surprisingly, those resonances displayed cross peaks with amino acid residues from the protein (**Figure 4-10, B**). Thereupon, it is indicative that the protein is interacting with and stabilised by PCP.

Figure 4-12 | 2D ^1H - ^1H TOCSY of DgCBD in Absence and Presence of PCP

(A) In the absence of PCP only a limited number resonance peak arise for CBD- γ , indicating the presence of a structure with only a few structural elements. **(B)** In presence of PCP resonance peaks of both the CBD- γ and PCP arose, including resonance cross-peaks between the carbohydrate and the protein.

Next, the ^1H - ^1H TOCSY and ^1H - ^1H NOESY spectra of CBD- γ without PCP and at an equimolar protein-carbohydrate concentration were fully assigned (**Figure 4-13, Figure 4-14**). Initial protein assignment of the free CBD- γ was limited by the number of resonance peaks in the ^1H - ^1H NOESY and ^1H - ^1H TOCSY NMR spectra (**Figure 4-13, A-B**). In contrast, the complex of CBD- γ with PCP returned intense and well-dispersed resonance peaks in both ^1H - ^1H NOESY and ^1H - ^1H TOCSY spectra and were used for protein sequence assignment (**Figure 4-14, A-B**). Comparing the ^1H - ^1H TOCSY spectra of CBD- γ with and without PCP also indicated an upfield chemical shift of resonances, suggesting conformational changes of CBD- γ in the presence of PCP (**Appendix Figure 6-6**).

Figure 4-13 | Solution ^1H ^1H NMR Characterisation of CBD- γ without PCP

Sequence specific resonance assignment of CBD- γ without PCP (A) in the fingerprint amide region of the two dimensional ^1H - ^1H TOCSY spectrum and (B) the two dimensional ^1H - ^1H NOESY spectrum. (C) NOE distribution of CBD- γ without of PCP, showing primarily short-range NOEs. (D) Histogram of the final structure with sequential (white bars), medium-long NOEs ($i, i+2 - i, i+4$) (grey) and long-range NOEs ($i, i+5$ and above) (black bars).

Figure 4-14 | Solution ^1H - ^1H NMR Characterisation of CBD- γ in Complex with PCP

Sequence specific resonance assignment of CBD- γ /PCP complex (A) in the fingerprint amide region of the two dimensional ^1H - ^1H TOCSY spectrum and (B) the two dimensional ^1H - ^1H NOESY spectrum. (C) NOE assignment of CBD- γ /PCP complex showing short- and medium-range NOEs. (D) Histogram of the final structure with sequential (white bars), medium-long NOEs (i , $i+2$ - i , $i+4$) (grey) and long-range NOEs (i , $i+5$ and above).

The assignments of the free CBD- γ and CBD- γ /PCP complex revealed structural tendencies of the protein. Calculating the chemical shift deviations (CSD) of CBD- γ without PCP revealed positive deviations between Q176 and V201, indicating the presence of an extended β -sheet conformation (Figure 4-15, A). In contrast, calculating the CSD of the protein-carbohydrate complex unveiled negative deviations in the region G159 – Q174, implying a helical protein folding, followed by a stretch of positive deviations between Q176 and V196, signifying an extended β -sheet conformation (Figure 4-15, B).

These findings support the earlier assumption of a ligand induced conformational change of CBD- γ .

Figure 4-15 | CSD of CBD- γ in Absence and at an Equimolar Protein-Carbohydrate Ratio.

(A) Chemical shift deviations (CSDs) of H α of CBD- γ residues from random coil without PCP, showing positive chemical shift deviations between Q₁₇₆ and V₂₀₁. (B) CSDs of H α of CBD- γ residues from random coil at an equimolar ratio of PCP with negative chemical shift deviations between G₁₅₉ and Q₁₇₄ and positive chemical shift deviations between Q₁₇₆ and V₁₉₆.

Ultimately, the results gathered from the homonuclear NMR experiments support the two previously made assumptions that the protein is binding to PCP and that it is undergoing a structural change which stabilises the protein. The analyses also revealed new insights about the structure of CBD- γ . Firstly, PCP, CBD- γ likely exhibits a β -sheet conformation at its C-terminus stretching from Q₁₇₆ to V₂₀₁, while the N-terminus remains primarily unstructured. Secondly, in complex with PCP, CBD- γ undergoes a structural change in its middle portion, stretching the region G₁₅₉ to Q₁₇₄ which likely exhibits an α -helix while the N-terminus continues to be folded in a β -sheet structure. Thirdly, in correlation with the proton affinity calculated, the binding between CBD- γ and PCP is assumed to be strong and in the low k_D range. Ultimately, despite not succeeding in determining the structure of the full-length protein unbiasedly, the homonuclear NMR experiments provided a wealth of new insights about CBD- γ that supported the approach of solving the three-dimensional structure of the protein in the following section.

4.11 Solution Three-Dimensional Structure of CBD- γ

Where homonuclear NMR is limited, three-dimensional heteronuclear NMR can improve structural determination beyond what can be achieved with homonuclear NMR. Heteronuclear NMR returns proton resonances based upon coupling between nuclei of two different types. To refine the protein structure of CBD- γ and to affirm the tight interaction between the protein and PCP, the protein was labelled with ^{15}N . Subsequently, 2-dimensional heteronuclear ^{15}N - ^1H HSQC and 3-dimensional ^1H - ^{15}N - ^1H TOCSY and NOESY experiments were carried out.

Carbohydrate titrations on the protein-carbohydrate complex were executed with PCP concentrations ranging from 0.1 mM to 2.4 mM at 700 μM of CBD- γ . In addition, CBD- γ was also measured without PCP to measure the proton chemical shifts between the free and the complexed protein. The CSDs also corroborated the findings from the homonuclear NMR assessment. Amino acids of CBD- γ were assigned by verifying the assignments of the of heteronuclear spectra and homonuclear spectra (from *Section 4.4*).

Best results were obtained at an equimolar carbohydrate-to-protein ratio (700 μM PCP and 700 μM CBD- γ), returning well-defined peaks in the ^1H - ^{15}N HSQC spectrum in the range 9.2 ppm - 7.2 ppm in ^1H dimension, indicative of a well-folded protein conformation (**Figure 4-16, A**). In addition, the NOESY peaks were well-defined, enabling a sequential walk-through to assign the protein (**Figure 4-16, B**).

Figure 4-16 | Heteronuclear Protein Assignment using ^1H - ^{15}N HSQC and ^1H - ^{15}N - ^1H NOESY
(A) Assignment of the fingerprint amide region of CBD- γ using ^1H - ^{15}N HSQC spectrum. **(B)** Example sequential walk-through of the folded region of CBD- γ using the ^1H - ^{15}N - ^1H NOESY spectrum.

4.12 Determination of the Structure of Free CBD- γ and in Complex with PCP

After fully assigning the heteronuclear and homonuclear spectra the assignments were used to calculate the protein structure of CBD- γ without and with PCP.

For the assignment of the free CBD- γ , about 157 NOEs that include 27 long-range NOEs (**Table 4-3**) were used to determine the tertiary structure. An ensemble of ten structures with the lowest energy structure was used to calculate the structure of CBD- γ using Cyana 2.1. Notably, the free CBD- γ was found to fold into three distinctive regions which vary in their RMSD: a “flexible”, spanning the N-terminus (R₁₄₁ - P₁₅₈), a “semiflexible” middle region between G₁₅₉ and P₁₇₄ and a “rigid” C-terminal region (Q₁₇₅ - V₂₀₁) (**Figure 4-17, A**). While the rigid, well-conserved domain is composed of three antiparallel β -strands, both the semiflexible and flexible region do not contain structural elements (**Figure 4-17, B**). Due the high content of unfolded regions in the structure, the root mean square deviation of (RMSD) remains high at 7.15 Å. In contrast, the RMSD of the folded C-terminal region alone is low at 0.44 Å.

For the determination of CBD- γ /complex, about 160 NOEs that included 48 medium range and 15 long range NOEs were included (**Table 4-3**). In contrast to the free CBD- γ , the

structure of the complex indicated that the previously mentioned semiflexible middle region (G₁₅₉ and P₁₇₄) folds and forms an ordered α -helical conformation, with its negatively charged amino acid residues exposed to the surface. This is further confirmed by the identification of more medium range NOEs in the protein-carbohydrate complex NMR spectra (**Table 4-3**).

With more structural regions the overall RMSD of the protein decreased to 3.45 Å. The RMSD of the folded region alone remains low at 0.47 Å. Markedly, the side chain Y₁₆₄, D₁₆₈ in helical region are interfacing and forming weak intramolecular interaction with S₁₇₉, Q₁₇₅ in the β_1 -sheet (**Figure 4-17, C-D**). Significantly, despite the presence of six cysteines in the amino acid sequence, no disulphide bonds within the structure of CBD- γ was found, leading to the conclusion that the structure does not share any similarity with the previously suggested proteins containing a CBM-14 motif. However, it is possible that disulphide bonds are being established between one of the other two chitin binding domains (CBD- α or CBD- β) within the same protein, which ultimately stabilise the full-length protein. Instead, CBD- γ is the first experimentally-determined structure of a domain containing an R&R Consensus sequence.

Table 4-3 | NOE restrains of CBD- γ without and in complex with PCP.

	CBD-γ	CBD-γ + PCP
Distance restraints		
sequential ($ i-j = 1$)	115	97
medium range ($2 \leq i-j \leq 4$)	15	48
long range ($ i-j \leq 5$)	27	15
total NOE constraints	157	160
Hydrogen bond restraints	12	34
Angular restraints		
Φ	61	61
Ψ	61	61
Distance constraint violations		
number of violations	45	52
average target function	17.07	29.55
Deviation from mean structure (RMSD)		
Backbone atoms ($\text{\AA}$)	7.15	3.45
Backbone atoms only folded ($\text{\AA}$)	0.44	0.47
Heavy atoms ($\text{\AA}$)	7.73	4.06
Ramachandran plot for the mean structure		
% residues in most favourable region	27.3	49.1
% residues in additionally allowed region	70.9	49.1
% residues in generously allowed region	1.8	1.
% residues in the disallowed region.	0	0

Figure 4-17 | Solution NMR structure of CBD-γ

(A) An ensemble of the backbone of ten lowest energy structures of CBD-γ without PCP. (B) A representative conformation of CBD-γ with colour gradient representing flexible, semiflexible and rigid domains of the protein in accordance with the ten lowest energy structures. (C) An ensemble of the backbone of ten lowest energy structures of CBD-γ in the presence of PCP. (D) Representative conformation of CBD-γ in the presence of PCP where the backbone atoms of the structurally defined residues are highlighted. The side chain interactions of Y₁₆₄, D₁₆₈ in the helical region interfacing with S₁₇₉ and Q₁₇₅ are represented as sticks.

Next, the surface charge of both the free CBD-γ and the protein-carbohydrate complex was analysed to identify potential binding sites of the protein where PCP interacts. PCP contains both positively and negatively charged groups and hence can interact with polar residues on the proteins surface. Analysing the surface charge of the free protein revealed that the folded region exposes primarily negative amino acid residues to the solvent (**Figure 4-18; left**). In contrast, after the addition of PCP and the triggered folding of CBD-γ, the α-helix structure exposes a negative

Figure 4-18 | Three - Dimensional Protein Structure of CBD- γ without (left) and with (right) PCP.

Surface net charge of CBD- γ alone (left) and in complex with pentaacetyl-chitopentaose (right). While the folded region of the free protein is primarily negatively charged, the folded regions of the carbohydrate – protein complex is split between a negative net charge of the α -helical region and a positive net charge of the β -sheets.

net charge to the surface, while the β -sheet containing region exposes primarily positive residues to the protein's surface (**Figure 4-18; right**). It is likely that the conformational change is mediated by the stabilizing hydrophilic interactions from the interactions of aromatic rings of Y₁₆₄ and F₁₈₃ with that of the CH₃ protons of H2 in PCP.

Conducting virtual screening using Schrödinger's Induced Fit (IFD) protocol and energetics analysis suggests that the complexation is driven primarily by Van der Waals (VdW) interactions located on the α -helix and the β -region (**Figure 4-19; B**). Both aromatic residues, Y₁₆₄ and F₁₈₃, are predicted to be in close proximity to the acetlyamine group of H2. Additionally, Y₁₆₄ is also predicted to be in close proximity to the saccharide ring system of PCP and its side chains could be important for stacking the sugar ring (**Figure**

4-19; A-B).

To further collate insights about the interactions, alanine mutants CBD- γ (Y₁₆₄→A, N₁₇₀→A, Q₁₇₆→A) were generated computationally and analysed with PCP docked onto the protein. Within 40 ns, the PCP dissociated from the mutant CBD- γ , the protein showed substantial unfolding (Figure 4-19, C-D).

Figure 4-19 | MD simulation of the CBD- γ -PCP complex and Mutant using Glide

(A) Prediction of the CBD- γ -PCP complex with the amino acid residues in proximity to the carbohydrate highlighted. F₁₈₃ is predicted to interact with the acetyl group of the acetamine of H₂, while Y₁₆₄ remains in close proximity with another C₂ domain and the saccharide ring system. Other predicted residues in interaction include polar residues located on the α -helix and β ₂-sheet. **(B)** Energy components involved in the binding between CBD- γ and PCP with both, electrostatic (ele) and Van der Waals (VdW) being predominant. **(C)** Binding mechanism of PCP to the Y₁₆₄A,

N₁₇₀A, Q₁₇₆A mutant of CBD- γ over 80ns. The MD simulation demonstrates that PCP does bind to the mutant of CBD- γ . **(D)** Distance between the centre of mass of the amino acid residues of CBD- γ and the centre of mass of PCP. While the distance of the trajectory between PCP and the wild-type of CBD- γ remains constant and is even slightly decreasing over time, the distance of the trajectory between PCP and the mutant of CBD- γ expands significantly. The distance between the centre of masses including the wild-type are initially closer than that of the mutant because of the longer side chains of Y, N and F compared to A.

To confirm the findings of the molecular dynamics analysis, an alanine mutant of CBD- γ (Y₁₆₄→A, N₁₇₀→A, Q₁₇₆→A) was generated and purified using the same protocol as for the wild-type. With the mutant a binding assay with chitin magnetic beads followed by analytical FPLC was conducted in line with the experiment outlined in Chapter 4.6. The experiment confirmed that the mutant did not have any activity to bind to chitin (**Appendix Figure. S5**).

4.13 Chemical Shift Perturbation of CBD- γ

After determining the structure of free CBD- γ and CBD- γ /PCP complex, the remaining information that is imperative to unravel is the binding mechanism is involved in the complexation of the carbohydrate to the protein. Analysing the net charge of the protein as discussed in the previous Chapter can be used to narrow down the location where the ligand binds without assessing the actual position of the binding. One method to further exploit and locate the active site of a protein is by determining the proton chemical shift perturbation (CSP), which assesses the change of the chemical shift of protons in the presence of a ligand [15]. CSP can provide useful information if the binding between protein and ligand is manifested via hydrogen interactions. The findings from CSP can be seconded by correlating the amide proton temperature coefficients with the hydrogen bonds. The assumption of the correlation is based on a previous study on the inverse proportionality between amide proton temperature coefficients and hydrogen bond lengths [16].

The CSP analysis was conducted by recording ¹H-¹⁵N HSQC spectra of free CBD- γ and of CBD- γ /PCP complex at an equimolar concentration. The total chemical shift ($\Delta\delta_{\text{total}}$) of

every individual amino acid was determined as the shift of the amino acid peaks of the protein-carbohydrate complex (δ_{bound}) and the peaks of the free protein (δ_{free}) and plotted against the corresponding residue number (**Figure 4-19, A**). Three amino acids, namely Y₁₆₄, N₁₇₀ and Q₁₇₆ in the helical folded regions were found to unbiasedly express strong perturbation in the presence of an equimolar concentration of PCP, exceeding the mean + 2 x the standard deviation. Additionally, other polar amino acid residues, including C₁₆₁, S₁₆₆, D₁₇₁, S₁₇₃, S₁₇₉, F₁₈₃, among others, showed a perturbation exceeding the mean + 0.5 x the standard deviation. All of these residues were mapped to be located either in the α -helix or the β_1 -sheet of CBD- γ , suggesting that the region spanning the α -helix and the β_1 -sheet is the active site of CBD- γ (**Figure 4-19 D**). From the mapping, two major types of interactions were identified. The first ones are the hydrophilic polar interactions between the electron-rich hydroxyl groups of the protein's polar residues (S₁₆₆, N₁₇₀, S₁₇₃, S₁₇₉) and the electron deficient acetamide groups of the carbohydrate. The second are between electronegative aromatic (π) rings of Y₁₆₄ and F₁₈₃ interacting with electropositive CH groups of H2 protons that bringing about the stronger hydrophobic effect to the binding.

Finally, the amide proton temperature coefficients were assessed from the same ¹H-¹⁵N HSQC spectra of free CBD- γ and complexed with PCP. Assessing the amide proton temperature coefficients confirmed that the same residues (Y₁₆₄, N₁₇₀, and Q₁₇₆) also exhibited the highest hydrogen bond efficiency, corroborating the findings from the CSP analysis (**Figure 4-20, B**). Even though these residues exhibit the highest perturbation, yet the weaker perturbations also continued to evolve with the further addition of PCP. At a PCP concentration of 1.2 mM, these residues were found to exchange with the ligand expressing chemical shifts which correspond to both the free and bound form (**Figure 4-20, C**). Interestingly, the CSPs of the two aromatic amino acids Y₁₆₄ and F₁₈₃ were further supported by the surface charge of the tertiary structure (**Figure 4-18**). In contrast, other residues located either in the N-terminus of the protein or in the β_2 and β_3 domains of the three-stranded β -sheet were overall less perturbed by the presence of PCP.

Figure 4-20 | CSP analysis and amide temperature coefficients of protein-carbohydrate complex

(A) The chemical shift perturbation plot of protein in complex with 0.7 mM of carbohydrate plotted against residue number. The residues that show stronger perturbations (above the mean + 2 x standard deviation (SD) line) are marked. **(B)** Amide temperature coefficients of free protein and protein-carbohydrate complex plotted against residue number. The residues that show stronger CSP show an increase in hydrogen bonding efficiency as marked. **(C)** Strips of 2D ^1H - ^{15}N HSQC of residues that show weaker perturbation in the presence of 0.7 mM PCP. As the protein exhibit tight binding with carbohydrate, at concentration as high as 1.2 mM, the peaks corresponding to slow exchange between free and bound form can be clearly seen. **(D)** Mapping of strong, medium and weakly perturbed residues (decreasing gradient of blue spheres) onto the molecular structure of the complex.

4.14 Conclusion

Several observations for the study on DgCBPs were covered in this Chapter. Firstly, the expression of recombinant DgCBPs remained a major challenge due to high insolubility and instability of the wild type proteins. Yet, a successful expression and purification of CBD- γ of DgCBP-3 was established and led to the production of sufficient amount of purified protein to conduct function and structural analyses.

Sequence analysis of CBD- γ confirmed the presence of a RR-2 Consensus, in addition to the previously reported CBM-14 motif by *Tan et al [2]*.

Subsequent qualitative characterisation of CBD- γ using CD spectroscopy, analytical FPLC, and one-dimensional NMR spectroscopy confirmed the structural integrity of the CBD- γ and the active binding to chitin and chitin derivatives. The qualitative characterisation also provided evidence on a ligand-triggered structural transition of the protein in the presence of PCP.

Next, ligand binding using STD-NMR was conducted with chito-oligomers of different lengths from (GlcNAc)₄ to (GlcNAc)₇ in order to map the effect a growing chain length on the binding affinity. It was found that the affinity increases by 33% comparing TCT and HpCH. STD-NMR was also used for group epitope mapping, and indicated that only H1, H2, H3, HN and H₃C protons bind to CBD- γ .

Further, the stability of CBD- γ in the presence of PCP was analysed in an unfolding assay with urea and DTT. The experiment demonstrated that the integrity of CBD- γ in the presence of PCP was higher compared to the free protein.

The structural characterisation using NMR spectroscopy led to the determination of two structure of CBD- γ , the free CBD- γ and CBD- γ in presence of PCP. The determination confirmed previous analytical assumptions that the protein undergoes a conformational change in the presence of PCP. The unbound protein can be segmented in three regions, namely a flexible N-terminus, a semi-flexible middle region and a rigid C-terminus. The semi-flexible middle region of CBD- γ contains an α -helix in the presence of the

carbohydrate. The characterisation further established that the conformational change also impacts the exposure of both positively and negatively amino acid residues on the surface of the folded region of CBD- γ in the presence of PCP.

Finally, the amino acids interacting with PCP were assessed using the CSP method and amide proton temperature coefficient. It was found that three residues, Y₁₆₄, N₁₇₀, and Q₁₇₆ located on α -helix and on the β 1-sheet exhibited the strongest chemical shifts and hydrogen bond efficiencies, confirming them to be interacting with PCP.

In conclusion, this Chapter provided sufficient information, required to derive to a conclusion on all hypotheses set for this project, which will be elaborate in more detail in the last Chapter.

4.15 References

1. Miserez, A., D. Rubin, and J.H. Waite, *Cross-linking chemistry of squid beak*. J Biol Chem, 2010. **285**(49): p. 38115-24.
2. Tan, Y., et al., *Infiltration of chitin by protein coacervates defines the squid beak mechanical gradient*. Nat Chem Biol, 2015. **11**(7): p. 488-95.
3. Elvin, C.M., et al., *Characterization of a major peritrophic membrane protein, peritrophin-44, from the larvae of Lucilia cuprina. cDNA and deduced amino acid sequences*. J Biol Chem, 1996. **271**(15): p. 8925-35.
4. Suetake, T., et al., *Chitin-binding proteins in invertebrates and plants comprise a common chitin-binding structural motif*. J Biol Chem, 2000. **275**(24): p. 17929-32.
5. Gasteiger, E., et al., *Protein Identification and Analysis Tool on the ExPASy Server*, in *The Proteomics Protocols Handbook*, J.M. Walker, Editor. 2005: Humana Press. p. 571-607.

6. Rebers, J.E. and L.M. Riddiford, *Structure and expression of a Manduca sexta larval cuticle gene homologous to Drosophila cuticle genes*. J Mol Biol, 1988. **203**(2): p. 411-23.
7. Andersen, S.O., *Amino acid sequence studies on endocuticular proteins from the desert locust, Schistocerca gregaria*. Insect Biochem Mol Biol, 1998. **28**: p. 421-434.
8. Andersen, S.O., *Studies on proteins in post-ecdysial nymphal cuticle of locust, Locusta migratoria, and cockroach, Blaberus craniifer*. Insect Biochem Mol Biol, 2000. **30**: p. 569-577.
9. Madden, T., *The BLAST Sequence Analysis Tool*, in *The NCBI Handbook*, J. McEntyre and J. Ostell, Editors. 2002, National Center for Biotechnology Information (US): Bethesda (MD).
10. Mayer, M. and B. Meyer, *Characterization of Ligand Binding by Saturation Transfer Difference NMR Spectroscopy*. Angew Chem Int Ed Engl, 1999. **38**(12): p. 1784-1788.
11. Angulo, J., P.M. Enriquez-Navas, and P.M. Nieto, *Ligand-receptor binding affinities from saturation transfer difference (STD) NMR spectroscopy: the binding isotherm of STD initial growth rates*. Chemistry, 2010. **16**(26): p. 7803-12.
12. Kasaai, M.R., *Determination of the degree of N-acetylation for chitin and chitosan by various NMR spectroscopy techniques: A review*. Carbohydr Polym, 2010. **79**: p. 801-810.
13. Asensio, J.L., et al., *Structural basis for chitin recognition by defense proteins: GlcNAc residues are bound in a multivalent fashion by extended binding sites in hevein domains*. Chem Biol, 2000. **7**(7): p. 529-43.
14. Gromiha, M.M., *Chapter 6 - Protein Stability*, in *Protein Bioinformatics*. 2011, Academic Press: Tokyo.

15. Williamson, M.P., *Using chemical shift perturbation to characterise ligand binding*. Prog Nucl Magn Reson Spectrosc., 2013. **73**: p. 1-16.
16. Cierpicki, T. and J. Otlewski, *Amide proton temperature coefficients as hydrogen bond indicators in proteins*. J Biomol NMR, 2001. **21**(3): p. 249-61.
17. Wimmerová, M.; Kozmon, S.; Nečasová, I.; Mishra, S. K.; Komárek, J.; Koča, J., Stacking Interactions between Carbohydrate and Protein Quantified by Combination of Theoretical and Experimental Methods. *PLOS ONE* **2012**, 7 (10), e46032.

Chapter 5 Discussion and Future Work

The following Chapter summarises and discusses the findings of the previous Chapters. In detail, it will discuss the determined structure of CBD- γ and the binding that was discovered between PCP and CBD- γ .

*The Chapter also summarises the expression and purification approach and the strategy that led to the characterisation of the chitin binding mechanism that is involved in the *D. gigas*'.*

Lastly, this Chapter also proposes the impact the determination of the RR Consensus structure may have on structural research related to insect cuticles, crustaceans and similar structures to that of the squid beak in other cephalopods, such as the octopus' beak.

5.1 General Discussion

Biomimetic research and especially the study of graded biological structures remains of high interest in the research community. The vision to permanently replace artificial materials with biodegradable, regenerative components is the belief of many who are trying to move towards a *Zero Waste* future. From vision to reality, science has proven to find solutions using the most unusual combinations of materials. As examples, Michelin's VISION tire, which is produced from a combination of organic, biodegradable materials, including orange zests, bamboo, molasses, wood, and natural rubber is designed on the template of a spiderweb, and medium density fibreboards that can be made from potatoes [1, 2]. The relationship between structure, property and design-thinking based on the design of natural models and materials is the underlying principle that is required to fabricate new materials.

The presented study has been inspired by the beak structure found in squid *D. gigas*, with particular interest in the mechanical gradient found from the buccal mass to the rostrum of the beak. The squid beak is an interesting model structure for the study of regenerative materials, with its major building blocks consisting of a mixture of two types of proteins, namely DgCBPs and DgHBPs, chitin and water.

Of special interest in this study is the function of DgCBPs, and the fundamental mechanism that enables the complexation with chitin. This is important considering that chitin is the second most abundant carbohydrate on earth with however very low industrial applications. Thus, understanding the materials where chitin is abundant and the mechanism by which it is present in natural materials provides a critical step to propose new methods on how to make use of chitin. Here, complementary methods were used in order to characterise the structure of DgCBPs and to analyse the binding mechanism involved when these proteins form a complex with chitin.

As reported, the beak of *D. gigas* contains four DgCBPs that share the presence of CBS as a common building block. Previous studies have hypothesised that DgCBPs interact with chitin to form a network in which DgHBPs infiltrate and coacervate to build the beak material. The first part of this project aimed to recombinantly express and purify DgCBPs

in native environments. Previously reported methods to directly extract the proteins from the beak were too *harsh* to study the beak proteins' natural structure. In this project, one of major the challenge was the initial design of a soluble construct of DgCBPs. Yet, one construct expressed soluble, namely CBP- γ , which is based on the third chitin binding domain of Dg-CBP-3. This construct was eventually expressed and purified in a sufficient amount for protein characterisation.

To gain more detailed information about the structure of CBD- γ , x-ray crystallography was initially attempted to solve the protein's crystal structure. Even though multiple commercially available x-ray crystallography screening kits were used, and two different crystallisation methods applied, no suitable precipitating solution conducive to crystallization could be identified. This correlates with the high mobility of the protein. Other research and the small size of CBD- γ allowed NMR spectroscopy to be used for structural determination of the protein. Subsequent isotopic labelling with ^{15}N significantly increased the resolution of the spectra, which were used to further refine the structure of CBD- γ . Additional attempts to double-label the protein with ^{13}C and ^{15}N did not yield sufficient amount of proteins for NMR measurements. Ultimately, the structure of CBD- γ was solved in the unbound state and in complex with PCP.

Finally, the nature of the binding mechanism of the protein in the presence of chitin derivatives was investigated using different NMR methods. The limiting factor of the analysis, however, is the low solubility of chito-oligomers beyond a chain length of seven GlcNAc's. The study was therefore limited to chain lengths between (GlcNAc) $_4$ to (GlcNAc) $_7$ to obtain the equilibrium dissociation coefficient K_D of the carbohydrate's groups. In addition, the amino acid residues involved in the protein-carbohydrate binding were located by mapping the perturbation of the chemical shifts and comparing them with the amide temperature coefficients.

5.2 Expression and Purification of Chitin Binding Beak Proteins

Unlike functional proteins, the majority of structural proteins forms fibres or strings making it more difficult to analyse their atomic structure. In the native state, DgCBPs are present in the squid beak of *D. gigas* in complex with other histidine-binding beak proteins, DgHBPs, chitin and water. To study the structure and function of DgCBPs, a recombinant approach is necessary. In this project, protein constructs were designed that either encompass the whole protein sequence or the active chitin-binding domains using the published sequences. Initially, only a preliminary construct of CBD- γ showed high impurities. The addition of an N-terminal GST to increase the solubility or the addition of a C-terminal His₆-tag not only helped with purification, but also acted as a protection group against degradation.

After an initial screening, the construct GST-PreSci-His₆-TEV-CBD- γ was designated and used for all subsequent protein expression. This construct contains two cleavage tags, namely PreScission Protease (PreSci) and Tobacco Etch Virus (TEV) Protease and both sites were assessed for their cleavage efficiency, ultimately leading to the conclusion that TEV cleavage produced lower content of non-CBD- γ remains after cleavage. Purity of the final CBD- γ was unaffected regardless of the cleavage.

That only one protein construct remained in the soluble fraction suggests that naturally DgCBPs require other components to become stable and active. A supporting argument for this hypothesis is the finding that CBD- γ remains very unstable and loses its structural integrity after a maximum of 24 hours following purification. Also, we have shown that the addition of ligand to the CBD- γ , stabilised its structural integrity as proven by 1D NMR spectroscopy. Comparing the unfolding of CBD- γ without and with PCP supports the theory of a higher stability of the protein-carbohydrate complex. The stability is likely to be caused by a ligand-induced conformation change of CBD- γ , which we could qualitatively observe using CD-spectroscopy.

Trying to refine the construct design of the remaining chitin-binding domains by applying the same approach in line with that of CBD- γ did not lead to any improvement of the

solubility of the constructs. Instead, we decided to focus on the structure determination of CBD- γ .

Nevertheless, it is possible that the problem of low solubility of DgCBPs could be overcome by co-expressing DgCBP with DgHBP, based on the assumption that DgCBPs and DgHBPs are actively interacting in the squid beak. Another solution is based on our findings from this study that CBDs undergo a conformational shift in the presence of chito-oligosaccharides. By supplementing the expression medium with soluble chitin derivatives during the protein expression could lead to the desired stability to express other CBDs. Finally, CBPs are Cys-rich proteins and consequently could be more soluble in other expression systems, such as insect cells or mammalian cells. However, an initial attempt to express a limited number of constructs in insect cells did not yield to a desired outcome.

5.3 Structural Characterisation of CBD- γ and the Impact of PCP

Next, CBD- γ protein structure was determined to assess the impact of the presence of chitin derivatives on the protein's integrity. The structural determination of a protein is a desired outcome in biophysics in the interest of visualising and interpreting overall findings.

X-ray crystallography was initially the preferred method to visualise ligand binding as no additional labelling of neither protein nor ligand is required. However, protein crystallisation did not result in any crystal. It is likely that the high portion of flexible protein (as later concluded from the NMR structure determination) precluded crystal growth. Due to the stabilizing role of chitin derivatives on the protein, it is possible that a carbohydrate supplementation could promote crystal growth. Despite the idea, the initial target of this project was to solve the free structure of CBD- γ before attempting to solve the structure of the protein-carbohydrate complex.

While x-ray crystallography was unsuccessful, 2-dimensional homonuclear proton NMR spectroscopy resulted in resonances that could be used to determine the tertiary structure of the protein. The low molecular weight of CBD- γ (~7.7 kDa) allowed the protein to be measured without labelling of additional isotopes. Generally, ^1H - ^1H distance restraints from 2D TOCSY and NOESY spectra provide sufficient information to calculate the tertiary

structure for non-isotope labelled small proteins (c.a. < 10 kDa) [3]. For larger proteins, a structural ^1H approach is challenging because of the severe signal overlapping and consequently uncertain assignment of resonances. Even though homonuclear proton NMR allowed the unambiguous assignment of CBD- γ in both absence and in complex with PCP, single isotopic labelling with ^{15}N was conducted to further refine the protein structure using heteronuclear NMR. A second isotopic labelling with ^{13}C , however, did not yield a sufficient amount of purified protein to include additional restraints. Nonetheless, equipped with three dimensions of correlation, the structures of free CBD- γ and CBD- γ in complex with PCP was successfully and unambiguously determined for these conditions.

The structure of free CBD- γ showed a large flexible N-terminal portion followed by a semiflexible middle portion and by a three-stranded β -sheet before the C-terminus. In line with the reported ligand-induced conformational change, we observed the formation of an α -helix from the semi-flexible middle portion of CBD- γ in the presence of PCP. Interestingly, the semi-flexible region contained the RR2 consensus sequence that is described to play a critical role in the chitin binding in insect cuticles. The conformational change that forms the α -helix in CBD- γ is therefore likely to be triggered by the additional polar moieties introduced by PCP, which actively interact with the polar residues positioned in the semi-flexible region and the β -sheets of CBD- γ . Molecular dynamics studies and docking suggested that the binding of PCP to CBD- γ is primarily driven by VdW interactions. It is believed that while the aromatic ring protons of Y164 and F183 could interact with the saccharide ring CH groups of PCP, the polar residues are likely to bind to the acetamide protons of H2 of the chitin side chains. MD analysis indicated that along with the aromatic residues, polar residues act as anchoring residues at the edges of the putative binding surface and therefore add specificity to the interaction. The interaction of Y164 among two other polar residues, N170, and Q176, with the carbohydrate was also observed when plotting the chemical shift perturbation and by plotting the amide temperature coefficients of free CBD- γ versus the complex. It is believed that three amino acids, namely Y164, N170, and Q176 play a crucial role in the binding of CBD- γ to PCP. The involvement of those three amino acids in the binding to PCP was supported by performing a computational simulation of a mutant variant of CBD- γ which predicted that

PCP detaches from CBD- γ and that the protein's α -helix starts to unfold. The assumption was supported by performing binding studies to chitin magnetic beads with alanine mutants CBD- γ (Y₁₆₄→A, N₁₇₀→A, Q₁₇₆→A) to further collate insights about the interactions. The mutant showed no binding to the beads when compared to its wild-type.

STD-NMR was used to identify the protons of chito-oligomers involved in the complex. Titrating chito-oligomers of different lengths against CBD- γ revealed H1, H2 and H3, as well as the acetyl group (CH₃COOH) and the amine group (NH₂) of H2 to be involved in the binding. STD-NMR confirmed that the length of the carbohydrate has an impact on the binding. As an example, HpCH with a chain length of 7 GlcNAc monomers showed a K_D value of the CH₃ protons of 10.3 μ M. Compared to the K_D of the same protons in PCP, HpCH demonstrated a decrease of 33% over a chain extension of 3 GlcNAc monomers. Because (GlcNAc)₇ only mimics a fraction of the actual length of chitin in the natural beak structure and CBD- γ is only one of three chitin-binding domains of DgCBP-3, it is likely that the actual native binding affinity between chitin and full-length DgCBP-3 is in the low micro-molar or even pico-molar range.

In summary, the NMR studies confirmed the conformational change of CBD- γ in the presence of its ligand as hypothesised from the initial CD spectroscopy. Interestingly, the change was found in the semi-folded region containing the RR2 consensus motif, which is related to chitin binding across species. Additionally, the binding of PCP was found to stabilise the protein due to the formation of additional intramolecular hydrogen bonds between carbohydrate and CBD- γ . Finally, the NMR studies also revealed that amino acid residues from the protein as well as side-groups of the carbohydrate are involved in the complex formation.

5.4 CBD- γ - a First Protein Structure of a Structural CBM

As previously discussed, CBDs form a family of carbohydrate-binding proteins, comprising amino acid sequences within larger multi-modular proteins. Their biological function is to bind to carbohydrates, triggering an enhancement of the catalytic efficiency of proteins.

While the majority of those CBDs are shown to have enzymatic activities, some are found to occur at different stages of the insect development and in the formation of the exoskeleton of crustaceans. These proteins however, have no obvious sequence similarity to any of the Cys-containing chitin-binding domains found in chitinases and some peritrophic membrane proteins [4]. From its protein structure and the mechanical support function of CBD- γ in the squid beak, as well as the extended R&R consensus sequence that has been identified in the sequence, it can be concluded that the protein is more similar to those CBMs found in insect cuticles than to the more common class of enzymatic CBMs [5, 6].

However, structural comparison of the identified structures of both free CBD- γ and CBD- γ in complex with PCP lack the availability of comparable structures. A structural comparison of the CBD- γ with other types of chitin-binding proteins, namely tachycitin or hevein, were unable to be superimposed and did not show any sequence similarity. The only available structure that contains an R&R Consensus was a ribbon model of the cuticle protein structure of HCCP12 predicted using GRASP by Ionomidou *et al.* and using FT-Raman, ATR FT-IR and CD spectroscopy to partially validate their findings [7, 8]. It was found that the proteins share some reminiscence with respect to their C-termini, with both containing an α -helix followed by an antiparallel β -sheet.

Consequently, the presented structures of CBD- γ are the first ever experimentally solved structures of a structural CBM. Secondly, the identified aromatic residues Y₁₆₄, located on the α -helix, and F₁₈₃, were predicted to strongly interact with PCP. While F₁₈₃ was predicted to interact with the acetlyamine of PCP, Y₁₆₄ is positioned well to interact with either the acetlyamine or the saccharide rings of the PCP, a common binding mechanism that has been well defined in previous research [9, 10]. Finally, and although not structurally related, the observed binding between the polar amino acid residues of CBD- γ and the carbohydrate's H1, H2 and H3 protons are in lines with previously published results of the interaction between hevein and PCP, suggesting that chitin binding generally occurs via polar interactions [11].

Next, previous studies conducted have proposed that DgCBPs share a similar sequence to that found in the chitin-binding type 2 domain of the peritrophin-44 protein, also known as CBM-14 [12]. CBM-14 is an extracellular domain that contains six conserved Cys that form three disulphide bridges. Chitin binding to CBM-14 could only be demonstrated if two or more CBM-14 domains are available [13]. Interestingly, the present study has demonstrated that free CBD- γ showed strong binding affinity towards PCP without the presence of disulphide bonds and consequently does not belong to the group of CBM-14.

Lastly, CBD- γ has been demonstrated to appear in two distinct conformations, a semi-structured unbound conformation and a structured conformation in the presence of PCP. Even though no similar mechanism is being reported for carbohydrate binding modules, the nature of CBD- γ in the formation a protein-carbohydrate network justified the discovered mechanism. Previous research conducted on the extraction of the native proteins of the squid's beak and the binding studies of CBD- γ to chitin magnetic beads conducted in this project proved the strong and nearly irreversible binding of DgCBPs to chitin [14]. A conformational change of the protein triggered by a ligand from an open to a stable, "locked" conformation is therefore to be expected. Further, the hypothesis can be supported with the increasing affinity between CBD- γ and chito-oligomers that is being observed as the chito-oligomer chain length increases. The formation of a stable network of DgCBPs with chitin could therefore be the initial step in the formation of the squid beak.

5.5 CBD- γ in Context of the Humboldt's Beak

DgCBPs belong to one of two protein families present in the squid beak of *D. gigas*. While the second protein family, histidine-binding beak proteins (DgHBPs), have recently been studied in more detail, little was known about DgCBPs [15]. As initially outlined, earlier research had suggested that DgCBPs and chitin form a protein-carbohydrate network in which DgHBPs infiltrate and coacervate. However, the assumptions were made without a clear understanding of the binding mechanism [12]. This study presented a model, using CBD- γ and the chitin derivate PCP, that can be used to deepen our understanding of the natural fabrication process of the beak. It was found that CBD- γ actively binds to chito-oligomers, and that during complexation it undergoes a conformational change that results

in a nearly irreversible complex. The complexation step is likely to represent the initial step of the beak's formation.

Secondly, it was shown that the length of the chito-oligomer has a strong impact on the binding affinity of the carbohydrate to the protein. The affinity of soluble chito-oligomers (with a maximum length of (GlcNAc)₇) increases with the chain length, reaching K_d values in the low micromolecular range. Since chitin can consist of up to 100's units of GlcNAc, it is highly likely that the binding affinity within the beak could decrease even further in the nano-molar range.

Thirdly, CBD- γ is only one of three carbohydrate-binding modules within DgCBP-3. This study has shown that the binding of the carbohydrate to CBD- γ appears on the surface of the protein via interactions between the polar amino acid residues of the protein and primarily the H1, H2 and acetylamino protons of the sugar. With three carbohydrate-binding modules available, each DgCBP-3 protein can bind to three chitin molecules, consequently enabling cross-linking. Furthermore, with chitin consisting of numerous GlcNAc monomers, each carbohydrate can also bind to multiple DgCBPs, ultimately forming a large protein-carbohydrate network.

5.6 Future Work

A major portion of this research was spent on the design of DgCBP constructs in an attempt to express either one of the active domains or the full-length protein in the soluble fraction, in its native form. Protein expression in inclusion bodies followed by refolding was purposely shunned to avoid misfolding that may occur during the refolding process. Therefore, the primary focus of this study was on the characterisation of CBD- γ , the only protein domain that initially appeared to express in the soluble fraction. CBD- γ proved to be an ideal model domain to study the binding mechanism leading to the formation protein-carbohydrate complex in the beak of *D. gigas*. This study unravelled the role of chitin-derivatives on the stability of CBD- γ . Despite multiple conserved cysteines in the sequence, no intramolecular disulphide bonds were detected. These findings can be used to either re-examine existing constructs in an attempt to purify non-soluble constructs from inclusion bodies via re-folding, or to explore expression in the presence of soluble chitin derivatives to actively stabilise the protein during expression.

Towards obtaining the native protein structure of CBD- γ , two of the most commonly used structural biological methods were initially explored, protein x-ray crystallography and NMR spectroscopy. NMR spectroscopy enabled to retrieve structures of CBD- γ both in free form and in complex with PCP. Although NMR spectroscopy is a versatile method and is able to study protein structures in solution, it is limited by the fact that both ligand and protein are required to be labelled in order to visualise the binding mechanism in more detail. The approach may be translated to study other types of protein-carbohydrate interactions. X-ray crystallography may be re-examined based the findings of the most suitable conditions determined from the NMR experiments in this study. The presented findings may be used to tune the crystallisation setup or to evaluate co-crystallisation of the protein-carbohydrate complex.

This study also challenges the previous assumption that carbohydrate binding domains of DgCBPs belong only the class of CBM-14, and instead proposed the additional presence of an extended RR-2 consensus. R&R Consensus sequences are commonly found in cuticle proteins of insects and in the exoskeleton of crustaceans. The structure of CBD- γ can not

only be used as a template to solve the remaining chitin-binding beak proteins of *D. gigas* but it can also be applied to other cephalopods, such as octopus or cuttlefishes, and to research being conducted on other proteins containing an R&R Consensus such as the proteins in the exoskeletons of crustaceans and in insect cuticles.

Further, it was discussed that CBD- γ binds to carbohydrates via polar residues. Interestingly, a similar binding mechanism has previously been reported for hevein. Though distinct, CBD- γ and hevein have chitin as common ligand. The methods that led to the finding in both studies can be used to assess the structure of other proteins that actively bind to chitin.

Finally, sequence homology revealed two of five sequences with sequence similarity >80% as compared to CBD- γ and the presence of the CBM-14 sequence and R&R Consensus. It is novel that both sequences have been identified in a protein and similar cases should therefore be investigated further in the future.

5.7 References

1. Anzilotti, E. This New Tire Has No Air And Is 3D Printed From Biodegradable Materials. [Online News] 2017 [cited 2019 7]; Available from: <https://www.fastcompany.com/40449277/this-new-tire-has-no-air-and-is-3d-printed-from-biodegradable-materials>.
2. Dooling, B. New Biodegradable MDF Wood Made with Potatoes! [Online News] 2013 [cited 2019 7]; Available from: <https://www.onegreenplanet.org/news/new-biodegradable-mdf-wood-made-with-potatoes/>.
3. Wuthrich, K., Protein structure determination in solution by NMR spectroscopy. J Biol Chem, 1990. 265(36): p. 22059-62.
4. Rebers, J.E. and J.H. Willis, A conserved domain in arthropod cuticular proteins binds chitin. Insect Biochem Mol Biol, 2001. 31(11): p. 1083-93.
5. Rebers, J.E. and L.M. Riddiford, Structure and expression of a *Manduca sexta* larval cuticle gene homologous to *Drosophila* cuticle genes. J Mol Biol, 1988. 203(2): p. 411-23.

6. Iconomidou, V.A., J.H. Willis, and S.J. Hamodrakas, Is beta-pleated sheet the molecular conformation which dictates formation of helicoidal cuticle? *Insect Biochem Mol Biol*, 1999. 29(3): p. 285-92.
7. Nicholls, A., K.A. Sharp, and B. Honig, Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins*, 1991. 11(4): p. 281-96.
8. Hamodrakas, S.J., J.H. Willis, and V.A. Iconomidou, A structural model of the chitin-binding domain of cuticle proteins. *Insect Biochem Mol Biol*, 2002. 32(11): p. 1577-83.
9. Asensio, J.L., et al., Carbohydrate-aromatic interactions. *Acc Chem Res*, 2013. 46(4): p. 946-54.
10. Hamodrakas, S.J., et al., The crystal structure of the complex of concanavalin A with 4'-methylumbelliferyl-alpha-D-glucopyranoside. *J Struct Biol*, 1997. 118(1): p. 23-30.
11. Asensio, J.L., et al., Structural basis for chitin recognition by defense proteins: GlcNAc residues are bound in a multivalent fashion by extended binding sites in hevein domains. *Chem Biol*, 2000. 7(7): p. 529-43.
12. Tan, Y., et al., Infiltration of chitin by protein coacervates defines the squid beak mechanical gradient. *Nat Chem Biol*, 2015. 11(7): p. 488-95.
13. Shen, Z. and M. Jacobs-Lorena, A type I peritrophic matrix protein from the malaria vector *Anopheles gambiae* binds to chitin. Cloning, expression, and characterization. *J Biol Chem*, 1998. 273(28): p. 17665-70.
14. Miserez, A., et al., The transition from stiff to compliant materials in squid beaks. *Science*, 2008. 319(5871): p. 1816-9.
15. Cai, H., et al., Self-coacervation of modular squid beak proteins - a comparative study. *Soft Matter*, 2017. 13(42): p. 7740-7752.

Chapter 6 Appendix

6.1 Construct Design

Appendix Figure 6-1 | Soluble Fraction of CBP-1 Constructs

Small batch purification analysis of 21 protein constructs of DgCBP-1 with none of the constructs being in the soluble fraction after lysis and ultracentrifugation

Appendix Figure 6-2 | Soluble Fraction of CBP-2 Constructs

Small batch purification analysis of 19 protein constructs of DgCBP-1 with none of the constructs being in the soluble fraction after lysis and ultracentrifugation

Appendix Figure 6-3 | Soluble Fraction of CBP-4 Constructs

Small batch purification analysis of 12 protein constructs of DgCBP-1 with none of the constructs being in the soluble fraction after lysis and ultracentrifugation

6.2 2-Dimensional Homonuclear NMR

Appendix Figure 6-4 | Ramachandran Plot of CBD-γ

The Ramachandran plot of free CBD-γ reports 16 amino acids in most favoured, 38 amino acids in additionally allowed regions and 1 amino acid in a generously allowed region. The residues are generally dispersed over the spectrum in comparison to Ramachandran Plot of CBD-γ in presence of PCP

Appendix Figure 6-5 | Ramachandran Plot of CBD-γ in Presence of PCP

The Ramachandran plot of CBD-γ in the presence of PCP reports 27 amino acids in most favoured regions, 27 amino acids in additionally allowed regions, and 1 amino acid in a generously allowed region. The residues are well dispersed over the α -helical region between ψ of -45° and 45° and the β -sheet region between a ψ of 90° and 180° .

Appendix Figure 6-6 | Overlay of 2D ^1H - ^1H TOCSY spectra of 700 μM of CBD- γ spectra

Overlay of 2D ^1H - ^1H TOCSY spectra of 700 μM of CBD- γ in the presence (green) and absence (red) of 0.7 mM PCP. The amino acid residues of CBD- γ that displays cross peaks with H2, H3 protons of PCP are labelled.

6.3 Analytical FPLC

Appendix Figure 6-7 | Binding Analysis of CBD- γ (mut) Using Analytical FPLC

Analytical PFLC of (A) CBD- γ and (B) CBD- γ (mut) (Y₁₆₄→A, N₁₇₀→A, Q₁₇₆→A). (C) GST-fusion protein was used as a negative control. Samples were incubated either in the presence of 100 μ L of chitin-magnetic beads (blue curves) or without the addition of beads (orange curves) for 2h at 4°C. In the presence of chitin magnetic beads, CBD- γ shows a loss of absorption as compared to both the CBD- γ mutant and the negative control, suggesting that the amino acid residues Y₁₆₄, N₁₇₀ and Q₁₇₆ identified in the NMR experiments are key residues enabling the binding between CBD- γ and chitin.