7-2020

Elucidating Collagen Degradation Synergy between Col G and Col H from Hathewaya (Clostridium) histolytica and Identifying novel structural features in HPT and REC domains from VarS histidine kinase in V. alginolyticus

Perry Caviness
University of Arkansas, Fayetteville

Elucidating Collagen Degradation Synergy between Col G and Col H from *Hathewaya* (*Clostridium*) *histolytica* and Identifying novel structural features in HPT and REC domains from VarS histidine kinase in *V. alginolyticus*

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Chemistry

by

Perry Caviness
University of Arkansas
Bachelor of Science in Biochemistry

July 2020
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council

Joshua Sakon Ph.D.
Dissertation Director

Paul Adams Ph.D.
Committee Member

Roger Koeppe Ph.D.
Committee Member

Francis Millet Ph.D.
Committee Member

Mack Ivey Ph.D.
Committee Member

Abstract

In this research the mechanisms by which *Hathewaya* (*Clostridia*) *histolytica* collagenases are secreted and work together to degrade collagens are investigated. While *H. histolytica* collagenases Col G and Col H have similar multi-domain structures the difference in number of and orientation of the domains hint that the two target different regions in collagen. Study small angle x-ray scattering (SAXS) was used to give a low-resolution envelope of full-length Col G and Col H and Col G/Col H non-catalytic domains bound to a collagen-like peptide (mini-collagen). SAXS derived envelopes along with structural information was used to tease out the mechanisms by which Col G and Col H degrade collagen. Structures of collagenase:mini-collagen complexes along with acid solubilized collagen hydrolysis studies was used to hypothesis a mechanism of synergy between Col G and Col H. Using all information from this research Col G to Col H ratio of 3:1 was proposed for optimal isolation pancreatic islets.

For *V. alginolyticus,* the VarS/VarA two-component systems is responsible for regulating the expression of collagenase. All two-component systems consist of a extracellular sensing domain, a catalytic and ATP binding (CA) domain, a dimerization and histidine phosphor-transfer (dHp) domain and a response regulator. For REC domain, itasser homology model and limited proteolysis revealed a two-domain structure connected by a flexible linker. Based on B-factor of linker truncation may be necessary for crystallization of the domain. In this study we initially sought to gain structural information on VarS, a sensor histidine kinase from VarS/VarA two-component system in *V. aglinolyticus*. It was discovered that the protein forms soluble aggregates in the absence of reducing agent. Trypsin digestion coupled with mass spectrometry revealed the existence of Cys883/Cys892 disulfide bridge. Homology modeling and DALI server was used to show that Cys892 is fairly conserved in HPT domains across species. qRT-PCR

showed that in the presence of menadione VarS induced transcription of sRNA2 was slightly downregulated. This study speculates that Cys883/Cys892 disulfide bond stops collagenase expression in the presence of oxidative stress allowing *V. alginoltyicus* to turn on necessary repair pathways.

Acknowledgements

I owe a great deal of gratitude to several people who have supported me along this journey. To begin, I would like to thank Dr. Joshua Sakon. The completion of this undertaking could not have been possible without your assistance and guidance. Thank you for encouraging me and challenging me along the way. I would also like to thank the members of my committee, Drs. Paul Adams, Mack Ivey, Frank Millet, and Roger Koeppe, thank you for your valuable feedback and comments, and for volunteering your own time to support me through the research process.

I'd like to express my gratitude to Drs. Osamu Matsushita and Takehiko Mima of Okayama University in Japan. Thank you for affording me the incredible opportunity to study alongside you in Japan during the summer of 2018. The advice and assistance you provided was immeasurable.

In addition, I'd like to thank the University of Arkansas, Fulbright College, the Department of Chemistry and the Graduate School for providing and fostering an environment that promotes and supports student success. I am a better scientist, researcher, and person because of my time spent here.

Lastly, and most importantly, a very special thank you to my parents, Brad and Lori Caviness, my friends, and Mary Kate Harrison for their constant support, guidance, and reassurance. Thank you for believing in me and for encouraging me every step of the way.

Table of Contents

Chapter 1: Introduction

Hathewaya histolytica (*Clostridium histolyticum*) is a gram-positive bacterium found in soil and feces that is known to cause gas gangrene (see figure 1A). *Hathewaya histolytica's* pathogenicity is due to the different exotoxins (alpha, beta, gamma, delta, epsilon) it secretes [**1**]. The β-exotoxin of this bacteria consist of two enzymes, Col G and Col H, which seek out and degrade collagen fibers located in extracellular space. Destruction of collagen exposes new nutrient sources for the invading bacteria and helps spread the infection [**2**]. Col G and Col H have similar multi-domain structures consisting of collagen binding domains (CBDs; CDB1 and CBD2 for Col G and CBD for Col H), polycystic kidney disease like-domains (PKD for Col G and PKD1 and PKD2 for Col H) and a catalytic domain, a zinc peptidase from the M9B family that is responsible for collagen hydrolysis (see figure 1B) [**3**]. Catalytic domain contains two sub-domains; an activator domain responsible for unwinding tropocollagen and a peptidase domain containing the catalytic Zn^{2+} responsible for hydrolysis [**4**]. Helper domain is thought to help the catalytic domain fold correctly [**4**].

CBDs are responsible for anchoring enzymes to substrate. In Col G CBD2 is thought to be necessary for processive degradation of collagen fibril while CBD1 is used to feed collagen substrate into the active site [**5**]. In chapter 2, a mechanism behind Col G hydrolysis is proposed. The specific role of PKD domains for Col G/Col H collagen hydrolysis is currently unknown. Col H PKD1 and PKD2 contain surface aromatic residues located along a proposed collagen binding ridge [**6**]. These residues are not present in Col G PKD suggesting the two play different roles. In chapter 3, the role of Col H PKD domains in collagen hydrolysis is discussed. The difference in domain orientation of Col G and Col H suggest that the two target different regions in collagen to efficiently degrade substrate (see figure 1B). In chapter 4, a method for collagen

degradation synergy between Col G and Col H is proposed based on structures of the enzyme and hydrolysis of acid solubilized collagen. The ability of Col G and Col H to bind tightly to and degrade collagen fibers makes them molecules of interest for clinical research [**7**] [**8**]. Further understanding on how Col G and Col H work together would allow for the enzymes to be further optimized for clinical research.

In addition to understanding the structure and mechanism of bacterial collagenases. This research also seeks to understand the systems by which collagenase expression is regulated. Two-component systems are stimulus response mechanisms found in bacteria and some lower-level eukaryotes that allow the organisms to sense and respond to different stimulus in the environment. Two-component systems consist of a histidine kinase that initially receives a signal and a response regulator responsible for transcriptional regulation [**9**]. Briefly their mechanisms involve, a signal being received which then causes the protein to dimerize. ATP then binds to a specific region in the histidine kinase and donates a phosphate to a conserved His residue. Finally the phosphate is passed (either in a *cis* or *trans* method) to a conserved Asp residue located on the Response Regulator. Phosphate can also be passed through internal receiver (REC) and histidine phosphor-transfer (HPT) domains in what is known as a phosphor-relay system. Internalized additional domains allow for extra regulatory checkpoints along the phosphorylation pathway [**10**].

V. alginolyticus, the system described in chapters 5 and 6 uses the VarS/VarA system to regulate the expression of collagenase (see figure 2). Since *V. alginolyticus* has low pathogenicity towards humans it is considered a model system for studying two-component systems. In chapter 5, new approach for crystallization of REC domain is proposed. Crystal structure of REC domain would help in drug development of anti-biotics that specifically target

two-component systems. In chapter 6, a potential cytoplasmic disulfide bridge was discovered in VarS HPT domain. Cys residues are excellent sensors for oxidative stress. In response regulator AgrA disulfide bond is formed across the active site after treatment with H_2O_2 [**11**]. Disulfide bond formation prevents response regulator from interacting with DNA. Similarly disulfide bond formation in HPT may inhibit phosphate transfer between REC and HPT or HPT and response regulator. Investigating this potential redox sensing pathway would reveal how pathogenic bacteria are able to avoid host defense systems.

1.1 Description of tropocollagen structure and hierarchal organization of fibrillar collagen

Collagen is the main structural protein in connective tissue such as skin, bone, tendon and cartilage and its cable-like structure provides mechanical stability [**12**]. There are 29 different types of collagens and they are dividing into groups according to their structure and the role they play [**13**]. Some are fibril forming collagen responsible for ECM and bone formation, some are network forming collagens involved in basement membrane function and some are have interruptions in their triple helix and are known as fibril-associated collagens with interrupted triple-helices (FACIT) collagens [**13**]. While Col G and Col H target a wide variety of collagen substrates the most common type of collagen in humans is the fibril forming collagens (Type I, II, III, V, XI) [**14**] [**13**]. As stated above, full-length Col G and Col H are used to degrade built up fibrillar collagen in Dupuytren's contracture and Peyronie's disease [**15**]. Full-length collagenases may also be used in cell isolation for purposes of transplant [**16**] [**17**] [**18**] [**19**]. In addition to this, research into using Col G and Col H non-catalytic segments for targeted delivery of biomolecules is underway [**20**] [**21**] [**22**] [**23**] [**24**] [**25**] [**26**] [**27**]. Usage require Col G and Col H to specifically target fibrillar collagen. To improve Col G and Col H's use as therapeutics

understanding the structure of fibrillar collagen is necessary. The mechanisms behind type I collagen formation are described below.

The most common fibril collagen is type I collagen which consists of two α1 chains and one α2 chain rotated together to form a triple helical structure [**28**]. Type I collagen chains are 300 nm long and 1.5 nm thick and consist of X-Y-Gly repeats with X and Y most frequently being Pro [**28**]. All single collagen chains come together to form triple helical trimers known as tropocollagens (see figure 3A). Prior to triple helix formation Lys residues on each α-chain are hydroxylated in the endoplasmic reticulum to form hydroxylysine. Lys residues in both triple helical regions and non-triple helical regions of collagen known as the telopeptide are hydroxylated [**29**]. Disulfide bond formation within C-propeptide aligns collagen α-chains to initiate triple helix formation and is catalyzed by special subunit of prolyl 4-hydroxylases [**29**] [**30**]. Collagen triple helix forms within the cell by having the three procollagen chains supercoil around each other in a right-handed manner. Small size of Gly at every third residue in an α-chain allows chains to pack close enough together to form the triple helix [**31**]. In the cell, post translational modification of Pro residues on collagen by prolyl 4-hydroxylases to form (2S,4R)-4-hydroxyproline (Hyp) is the most common post translational modification in humans [**32**]. The presence of Hyp at Cγ stabilizes the triple helix by changing conformation of the pyrrolidine ring of proline at every Y position from endo to exo which in turn promotes packing of the α-chains to form triple helix [**33**]. Once triple helix is formed it is further stabilized by numerous amine/carbonyl hydrogen bonds between each strand [**28**].

Chaperoned by collagen dedicated Heat shock protein 47, newly formed triple helical procollagen is secreted into the extracellular space [**29**] [**34**]. N and C-propeptides are then cleaved off by specific metalloproteinases to form mature collagen [**29**]. After secretion, it is

thought that mature tropocollagen molecules begin to spontaneously assembly into a quasi-hexagonal array parallel and longitudinally staggered by a repeat distance of 67 nm known as the D period [**35**]. This assembled array of collagens is known as the micro-fibril. During micro-fibril formation, Lys and Hyl residues located at the the N and C telopeptides are oxidatively deaminated by Lysyl oxidase to form aldehydic forms of the residues [**29**] [**36**]. A series of condensation reactions then occur at these new residues to form covalent intra and inter-molecular pyridinoline cross-links in tropocollagen chains which stabilize the micro-fibril [**29**] (see figure 3B). In the N-telopeptide Lys5 from the α2 chain intramolecularly cross-links with Lys9 from one α1 chain while Lys9 from the remaining α1 chain intermolecularly cross-links with Lys946 from the triple helical region of an α1 chain in an adjacent tropocollagen [**37**]. While at the C-telopeptide Lys1046 from both α1 chains cross-link with Lys106 from the collagen triple helix [**37**]. Multiple exterior tropocollagens on micro-fibrils are cross-linked to together to form collagen fibrils. The specific number of micro-fibrils within a fibril is currently up for debate.

Small leucine-rich repeat proteoglycans (SLRPs) are known to exist on the surface of collagen fibrils [**38**]. Glycosamineglycans (GAG) chains are present on the N-terminal domains of SLRPs and bind to non-fibril FACIT collagens present in the extrafibrillar space between collagen fibrils. In addition to this, GAG chains interact with GAGs from adjacent collagen fibrils via electrostatic interactions lead to collagen fiber formation [**39**]. Collagenases Col G and Col H from *H. histolytica* were evolved to unbundle the hierarchical structure of collagen fibers, exposing tropocollagen molecules for hydrolysis. The presence of dual collagenases suggests that Col G and Col must synergistically degrade collagen. However, this mechanism of synergy is currently unknown. For us to potentially tease out this mechanism understanding the structure

of Col G and Col H is necessary. Structural and functional information from Col G and Col H domains are described in the next section.

1.2 Clinical use for full-length Col G and Col H mixtures

The ability of full-length Col G and Col H to seek out and synergistically degrade collagen fibers makes Col G and Col H molecules of interest for clinical research purposes. Santyl ointment is a mixture of Col G and Col H and is used as a wound debridement and as stated earlier in the chapter mixtures of the enzyme are used in the treatment of Dupuytren's contracture and Peyronie's disease [**15**]. Pancreatic islet transplantation was mentioned earlier in the chapter. It is an experimental treatment for type I diabetes and pancreatic cancer where pancreatic islets are isolated from the patient pancreas or from a donor using a mixture of Col G, Col H and neutral protease (normally thermolysin) (see figure 4). Newly isolated islets are cultured and then injected into the patients liver where pancreatic function is restored [**19**]. Minimizing the amount of neutral protease used in isolation is important as excess limits islet survival and viability *in vitro* [**40**]. In order to minimize the amount of protease used the optimal ratio of Col G to Col H needed for optimizing the islet yield needs to be found. Currently different ratios of Col G to Col H are being investigated on their cell isolation ability but a consensus has yet to be reached [**41**]. In chapter 5, based on what is known about Col G and Col H collagen hydrolysis the optimal ratio of Col G to Col H needed for hydrolysis of fibrillar collagen is proposed [**42**] [**43**]. For the rest of this section potential roles for *Hathewaya histolytica* collagenases are discussed.

In addition to pancreatic islet isolation, collagenases can be used for the isolation of stem cells from adipose fat tissue. Recently a method was proposed by Gentile and Sterodimas for the isolation of stem cells from adipose fat tissue using human MMP for purposes of reversing

Covid-19 lung damage [**44**]. Mesenchymal stem cell isolation process may be optimized by using bacterial collagenases instead of human MMP. Stem cells may also be isolated from synovial fluid [**45**]. In addition to treating Covid-19, collagenases could be used to isolate skeletal muscle stem cells for the purposes of treating muscular dystrophy and age associated muscle atrophy [**46**].

Collagenases may also have some applications in nano-surgery. In many orthodontic procedures minor surgery is required to remove the collagen connecting teeth to underlying bone [**47**]. An inactive version of Col G was loaded onto a calcium impermeable DMPC bilayer. Once at the oral cavities Col G is secreted from the liposomes, activated by calcium present in the oral cavity and begins to degrade collagen connecting teeth to bone. Once collagen is degraded teeth can be maneuvered into correct positions and natural tissue repair functions will regenerate the collagen to keep the tooth in its new position [**47**]. Understanding how Col G and Col H degrade collagen would potentially make them more effective as clinical tools.

1.3 Two-component systems represent novel targets for developing broad spectrum anti-biotics

Information from this research into the structures of VarS cytoplasmic domains could be used to create broad spectrum anti-biotics. Anti-biotics would target conserved structural motifs in the present in all phosphorelay sensor histidine kinases. Anti-biotic resistant bacteria are currently a severe healthcare problem for the planet. Recently developed anti-biotics all tend to target the same structures and pathways decreasing their effectiveness [**48**]. Gene expression analysis of drug resistant tuberculosis isolates show that two component response systems can enhance drug resistance [**49**]. One side effect of growing antibiotic resistance is the rising occurrence of secondary infections. Tungiasis, is a parasitic disease caused by the female sand flea and infection of lesions by a variety of anti-biotic resistant bacteria is very

common [**50**]. Broad spectrum antibiotics that target two component systems would be a way to potentially stop these types of secondary infections. Two-component system inhibitors work by binding to and inducing aggregation of the histidine kinase domain; information from this research would potentially reveal an alternative targets further downstream for anti-biotics to attack [**51**] [**52**]. Currently there are risks towards depending heavily on broad spectrum anti-biotics to fight an infection since bacteria are more likely to develop resistance to these types of anti-biotics. Broad spectrum anti-biotics that target two component systems would be more effective than current drug options as the bacteria would not have evolved any resistance to the new anti-biotics. Since these potentially novel anti-biotics are broad spectrum they would be able to effectively target a large number of bacteria.

Chapter 1: Figures

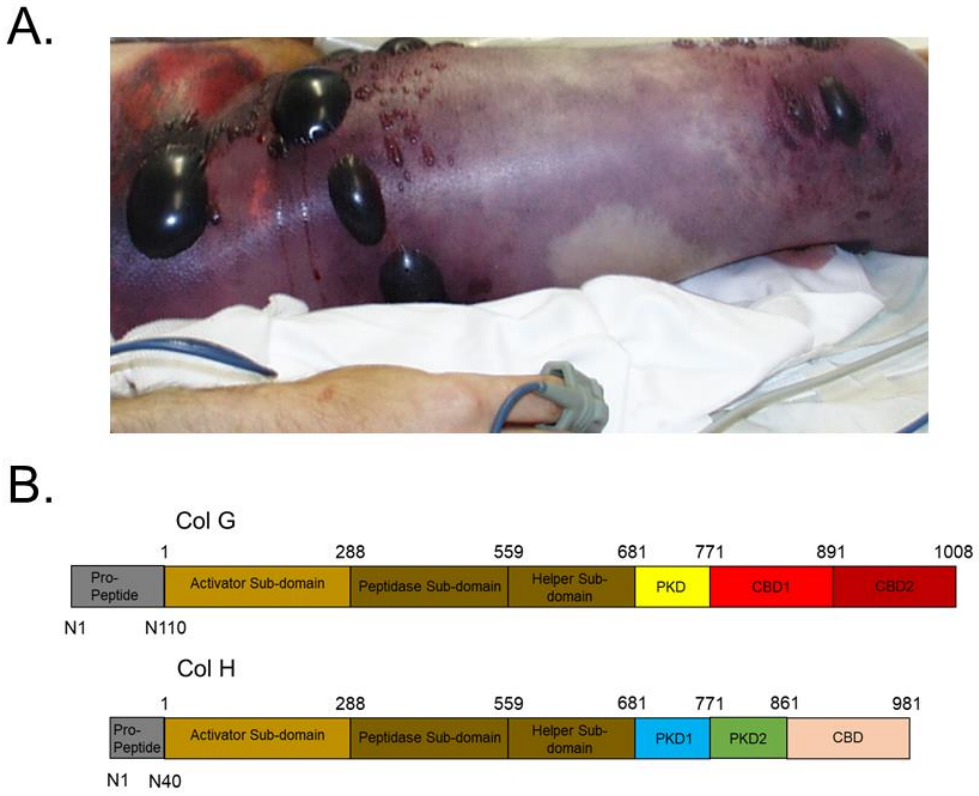


Figure 1: **A.** Photograph of a patient with Gas gangrene in their right leg before amputation. Bacterial collagenases are responsible for the spread of the disease. **B.** Domain map of Col G and Col H from *H. histolytica*. In order to activate Col G and Col H Pro-peptide (gray, Col G: N1 – N110 Col H: N1-N40) must be cleaved. Catalytic domain of Col G and Col H are composed of Activator sub-domains (light brown, residues 1 – 288) and Peptidase subdomain (dark brown, residues 289 – 559), which itself contains a helper subdomain to help the domain fold correctly (dart brown, residues 560 – 681). PKD domains (single PKD for Col G in yellow, residues 682 – 771; PKD1 and PKD2 for Col H in blue and green, residues 681 – 771 for PKD1 and residues 772 – 861 for PKD2) connect catalytic domains of both enzymes to CBDs. CBDs (CBD1 and CBD2 for Col G in red and dark red, residues 772 – 891 for CBD1 and residues 892 – 1008 for CBD2; single CBD for Col H in pink, residues 862 – 981) are responsible for binding both enzymes to collagen.

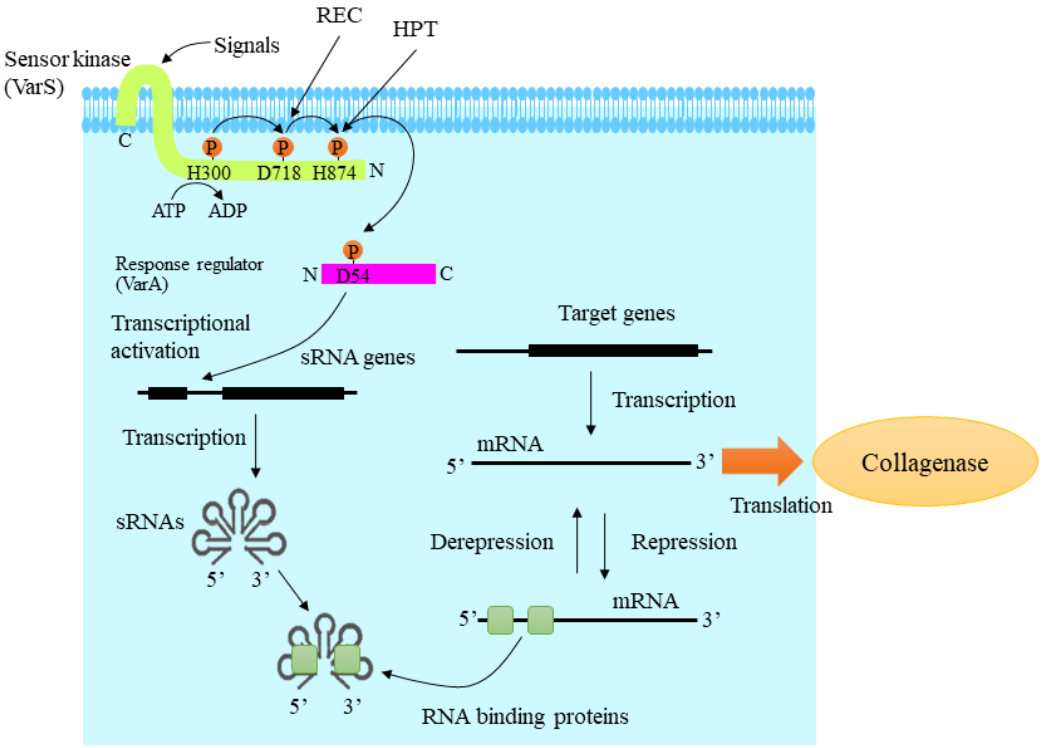


Figure 2: Cartoon representation of VarS/VarA two component system. Activation of response regulate VarA leads to transcription of sRNA which bind RNA repressor proteins and allow for transcription of Collagenase

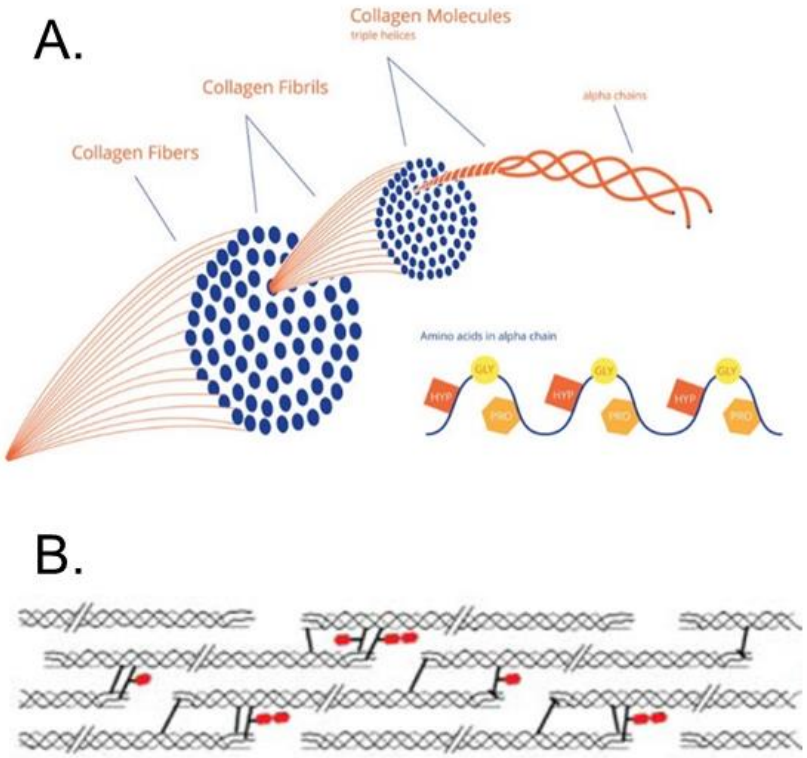


Figure 3: **A**. Cartoon representation of the hierarchical structure of fiber forming collagens. 3 α-chains with composition X-Y-Gly with X and Y most commonly being Pro and Hyp pack together to form triple helical tropocollagen. **B.** Once secreted into the extracellular space tropocollagens pack into micro-fibril. Tropocollagen is both intra and intermolecularly crosslinked at Lys residues in telopeptide and triple helical regions by LOX to give micro-fibril extra tensile strength. Figure 3B is adapted from Yamauchi et al. 2012

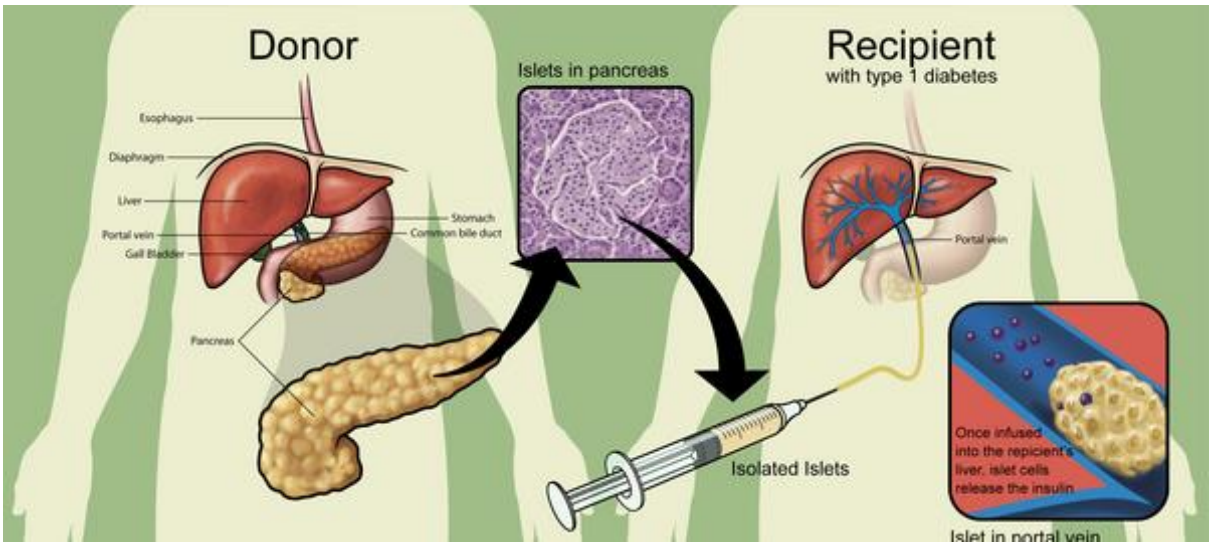


Figure 4: Cartoon outlining pancreatic islet allotransplantation. Islets are injected into liver where they begin to release insulin. Islets are isolated using mixtures of Col G and Col H. Understanding how Col G and Col H work to degrade fibrillar collagen would help to find the optimal ratio needed for isolation. Liver is chosen as site for transplantation as it allows for islets to receive sufficient amount of oxygen and nutrition [**9**]

Works Cited

1. C. L. Hatheway, *Toxigenic clostridia*, Clinical microbiology reviews **3** (1990), no. 1, 66-98.

2. A. S. Duarte, A. Correia and A. C. Esteves, *Bacterial collagenases - a review*, Crit Rev Microbiol **42** (2016), no. 1, 106-126.

3. O. Matsushita, C. M. Jung, J. Minami, S. Katayama, N. Nishi and A. Okabe, *A study of the collagen-binding domain of a 116-kda clostridium histolyticum collagenase*, The Journal of biological chemistry **273** (1998), no. 6, 3643-3648.

4. U. Eckhard, E. Schonauer, D. Nuss and H. Brandstetter, *Structure of collagenase g reveals a chew-and-digest mechanism of bacterial collagenolysis*, Nature structural & molecular biology **18** (2011), no. 10, 1109-1114.

5. P. Caviness, R. Bauer, K. Tanaka, K. Janowska, J. R. Roeser, D. Harter, J. Sanders, C. Ruth, O. Matsushita and J. Sakon, *Ca(2+) -induced orientation of tandem collagen binding domains from clostridial collagenase colg permits two opposing functions of collagen fibril formation and retardation*, Febs j **285** (2018), no. 17, 3254-3269.

6. R. Bauer, K. Janowska, K. Taylor, B. Jordan, S. Gann, T. Janowski, E. C. Latimer, O. Matsushita and J. Sakon, *Structures of three polycystic kidney disease-like domains from clostridium histolyticum collagenases colg and colh*, Acta Crystallographica Section D-Structural Biology **71** (2015), 565-577.

7. S. T. L. Philominathan, T. Koide, K. Hamada, H. Yasui, S. Seifert, O. Matsushita and J. Sakon, *Unidirectional binding of clostridial collagenase to triple helical substrates*, Journal of Biological Chemistry **284** (2009), no. 16, 10868-10876.

8. S. T. L. Philominathan, T. Koide, O. Matsushita and J. Sakon, *Bacterial collagen-binding domain targets undertwisted regions of collagen*, Protein Science **21** (2012), no. 10, 1554-1565.

9. K. Lapouge, M. Schubert, F. H. Allain and D. Haas, *Gac/rsm signal transduction pathway of gamma-proteobacteria: From rna recognition to regulation of social behaviour*, Mol Microbiol **67** (2008), no. 2, 241-253.

10. R. Gao and A. M. Stock, *Biological insights from structures of two-component proteins*, Annual review of microbiology **63** (2009), 133-154.

11. F. Sun, H. Liang, X. Kong, S. Xie, H. Cho, X. Deng, Q. Ji, H. Zhang, S. Alvarez, L. M. Hicks, T. Bae, C. Luo, H. Jiang and C. He, *Quorum-sensing agr mediates bacterial oxidation response via an intramolecular disulfide redox switch in the response regulator agra*, Proceedings of the National Academy of Sciences of the United States of America **109** (2012), no. 23, 9095-9100.

12. J. Bella, *Collagen structure: New tricks from a very old dog*, Biochemical Journal **473** (2016), no. 8, 1001-1025.

13. S. Ricard-Blum, *The collagen family*, Cold Spring Harb Perspect Biol **3** (2011), no. 1, a004978.

14. T. Toyoshima, O. Matsushita, J. Minami, N. Nishi, A. Okabe and T. Itano, *Collagen-binding domain of a clostridium histolyticum collagenase exhibits a broad substrate spectrum both in vitro and in vivo*, Connect Tissue Res **42** (2001), no. 4, 281-290.

15. H. Alipour, A. Raz, S. Zakeri and N. Dinparast Djadid, *Therapeutic applications of collagenase (metalloproteases): A review*, Asian Pacific Journal of Tropical Biomedicine **6** (2016), no. 11, 975-981.

16. A. N. Balamurugan, A. G. Breite, T. Anazawa, G. Loganathan, J. J. Wilhelm, K. K. Papas, F. E. Dwulet, R. C. McCarthy and B. J. Hering, *Successful human islet isolation and transplantation indicating the importance of class 1 collagenase and collagen degradation activity assay*, Transplantation **89** (2010), no. 8, 954-961.

17. R. C. McCarthy, A. G. Breite, M. L. Green and F. E. Dwulet, *Tissue dissociation enzymes for isolating human islets for transplantation: Factors to consider in setting enzyme acceptance criteria*, Transplantation **91** (2011), no. 2, 137-145.

18. B. J. Olack, C. J. Swanson, T. K. Howard and T. Mohanakumar, *Improved method for the isolation and purification of human islets of langerhans using liberase™ enzyme blend*, Human Immunology **60** (1999), no. 12, 1303-1309.

19. M. L. Green, A. G. Breite, C. A. Beechler, F. E. Dwulet and R. C. McCarthy, *Effectiveness of different molecular forms of c. Histolyticum class i collagenase to recover islets*, Islets **9** (2017), no. 6, 177-181.

20. H. Sekiguchi, K. Uchida, G. Inoue, O. Matsushita, W. Saito, J. Aikawa, K. Tanaka, H. Fujimaki, M. Miyagi and M. Takaso, *Acceleration of bone formation during fracture healing by poly(pro-hyp-gly)10 and basic fibroblast growth factor containing polycystic kidney disease and collagen-binding domains from clostridium histolyticum collagenase*, J Biomed Mater Res A **104** (2016), no. 6, 1372-1378.

21. W. Saito, K. Uchida, O. Matsushita, G. Inoue, H. Sekiguchi, J. Aikawa, H. Fujimaki and M. Takaso, *Acceleration of callus formation during fracture healing using basic fibroblast growth factor-kidney disease domain-collagen-binding domain fusion protein combined with allogenic demineralized bone powder*, Journal of Orthopaedic Surgery and Research **10** (2015), 6.

22. K. Uchida, O. Matsushita, K. Naruse, T. Mima, N. Nishi, S. Hattori, T. Ogura, G. Inoue, K. Tanaka and M. Takaso, *Acceleration of periosteal bone formation by human basic fibroblast growth factor containing a collagen-binding domain from clostridium histolyticum collagenase*, Journal of biomedical materials research. Part A **102** (2014), no. 6, 1737-1743.

23. H. Sekiguchi, K. Uchida, O. Matsushita, G. Inoue, N. Nishi, R. Masuda, N. Hamamoto, T. Koide, S. Shoji and M. Takaso, *Basic fibroblast growth factor fused with tandem collagen-binding domains from clostridium histolyticum collagenase colg increases bone formation*, Biomed Res Int **2018** (2018), 8393194.

24. T. Ponnapakkam, R. Katikaneni, E. Miller, A. Ponnapakkam, S. Hirofumi, S. Miyata, L. J. Suva, J. Sakon, O. Matsushita and R. C. Gensure, *Monthly administration of a novel*

pth-collagen binding domain fusion protein is anabolic in mice, Calcif Tissue Int **88** (2011), no. 6, 511-520.

25. T. Ponnapakkam, R. Katikaneni, H. Suda, S. Miyata, O. Matsushita, J. Sakon and R. C. Gensure, *A single injection of the anabolic bone agent, parathyroid hormone-collagen binding domain (pth-cbd), results in sustained increases in bone mineral density for up to 12 months in normal female mice*, Calcified Tissue International **91** (2012), no. 3, 196-203.

26. R. Katikaneni, A. W. Seymour, R. Gulati, T. Ponnapakkam and R. C. Gensure, *Therapy for alopecia areata in mice by stimulating the hair cycle with parathyroid hormone agonists linked to a collagen-binding domain*, J Investig Dermatol Symp Proc **17** (2015), no. 2, 13-15.

27. R. Katikaneni, T. Ponnapakkam, H. Suda, S. Miyata, J. Sakon, O. Matsushita and R. C. Gensure, *Treatment for chemotherapy-induced alopecia in mice using parathyroid hormone agonists and antagonists linked to a collagen binding domain*, Int J Cancer **131** (2012), no. 5, E813-821.

28. M. D. Shoulders and R. T. Raines, "Collagen structure and stability," *Annual review of biochemistry*, vol. 78, Annual Reviews, Palo Alto, 2009, pp. 929-958.

29. M. Yamauchi and M. Sricholpech, *Lysine post-translational modifications of collagen*, Essays Biochem **52** (2012), 113-133.

30. K. I. Kivirikko and T. Pihlajaniemi, *Collagen hydroxylases and the protein disulfide isomerase subunit of prolyl 4-hydroxylases*, Adv Enzymol Relat Areas Mol Biol **72** (1998), 325-398.

31. J. Bella, M. Eaton, B. Brodsky and H. M. Berman, *Crystal and molecular structure of a collagen-like peptide at 1.9 a resolution*, Science **266** (1994), no. 5182, 75.

32. K. L. Gorres and R. T. Raines, *Prolyl 4-hydroxylase*, Critical reviews in biochemistry and molecular biology **45** (2010), no. 2, 106-124.

33. J. D. Vasta and R. T. Raines, *Collagen prolyl 4-hydroxylase as a therapeutic target*, Journal of medicinal chemistry **61** (2018), no. 23, 10403-10411.

34. T. R. Dafforn, M. Della and A. D. Miller, *The molecular interactions of heat shock protein 47 (hsp47) and their implications for collagen biosynthesis*, J Biol Chem **276** (2001), no. 52, 49310-49319.

35. J. P. R. O. Orgel, T. C. Irving, A. Miller and T. J. Wess, *Microfibrillar structure of type i collagen in situ*, Proceedings of the National Academy of Sciences of the United States of America **103** (2006), no. 24, 9001-9005.

36. M. Yamauchi and M. Shiiba, *Lysine hydroxylation and cross-linking of collagen*, Methods Mol Biol **446** (2008), 95-108.

37. K. Sato, T. Ebihara, E. Adachi, S. Kawashima, S. Hattori and S. Irie, *Possible involvement of aminotelopeptide in self-assembly and thermal stability of collagen i as revealed by its removal with proteases*, J Biol Chem **275** (2000), no. 33, 25870-25875.

38. S. Kalamajski and Å. Oldberg, *The role of small leucine-rich proteoglycans in collagen fibrillogenesis*, Matrix Biology **29** (2010), no. 4, 248-253.

39. Y. X. Liu, Z. Q. Liu, L. Song, Q. R. Ma, D. Y. Zhou, B. W. Zhu and F. Shahidi, *Effects of collagenase type i on the structural features of collagen fibres from sea cucumber (stichopus japonicus) body wall*, Food Chem **301** (2019), 125302.

40. H. Brandhorst, P. R. Johnson, J. Mönch, M. Kurfürst, O. Korsgren and D. Brandhorst, *Comparison of clostripain and neutral protease as supplementary enzymes for human islet isolation*, Cell transplantation **28** (2019), no. 2, 176-184.

41. A. G. Breite, R. C. McCarthy and F. E. Dwulet, *Characterization and functional assessment of clostridium histolyticum class i (c1) collagenases and the synergistic degradation of native collagen in enzyme mixtures containing class ii (c2) collagenase*, Transplantation Proceedings **43** (2011), no. 9, 3171-3175.

42. M. F. French, K. A. Mookhtiar and H. E. Van Wart, *Limited proteolysis of type i collagen at hyperreactive sites by class i and ii clostridium histolyticum collagenases: Complementary digestion patterns*, Biochemistry **26** (1987), no. 3, 681-687.

43. M. F. French, A. Bhown and H. E. Van Wart, *Identification of clostridium histolyticum collagenase hyperreactive sites in type i, ii, and iii collagens: Lack of correlation with local triple helical stability*, J Protein Chem **11** (1992), no. 1, 83-97.

44. P. Gentile and A. Sterodimas, *Adipose-derived stromal stem cells (ascs) as a new regenerative immediate therapy combating coronavirus (covid-19)-induced pneumonia*, Expert opinion on biological therapy (2020), 1-6.

45. N. Sugita, Y. Moriguchi, M. Sakaue, D. A. Hart, Y. Yasui, K. Koizumi, R. Chijimatsu, S. Shimomura, Y. Ikeda, H. Yoshikawa and N. Nakamura, *Optimization of human mesenchymal stem cell isolation from synovial membrane: Implications for subsequent tissue engineering effectiveness*, Regen Ther **5** (2016), 79-85.

46. K. Ishii, N. Suzuki, Y. Mabuchi, I. Sekiya and C. Akazawa, *Technical advantage of recombinant collagenase for isolation of muscle stem cells*, Regen Ther **7** (2017), 1-7.

47. A. Zinger, O. Adir, M. Alper, A. Simon, M. Poley, C. Tzror, Z. Yaari, M. Krayem, S. Kasten, G. Nawy, A. Herman, Y. Nir, S. Akrish, T. Klein, J. Shainsky-Roitman, D. Hershkovitz and A. Schroeder, *Proteolytic nanoparticles replace a surgical blade by controllably remodeling the oral connective tissue*, ACS Nano **12** (2018), no. 2, 1482-1490.

48. H. B. Felise, H. V. Nguyen, R. A. Pfuetzner, K. C. Barry, S. R. Jackson, M. P. Blanc, P. A. Bronstein, T. Kline and S. I. Miller, *An inhibitor of gram-negative bacterial virulence protein secretion*, Cell host & microbe **4** (2008), no. 4, 325-336.

49. G. Yu, Z. Cui, X. Sun, J. Peng, J. Jiang, W. Wu, W. Huang, K. Chu, L. Zhang, B. Ge and Y. Li, *Gene expression analysis of two extensively drug-resistant tuberculosis isolates show that two-component response systems enhance drug resistance*, Tuberculosis (Edinburgh, Scotland) **95** (2015), no. 3, 303-314.

50. R. M. Nyangacha, D. Odongo, F. Oyieke, M. Ochwoto, R. Korir, R. K. Ngetich, G. Nginya, O. Makwaga, C. Bii, P. Mwitari and F. Tolo, *Secondary bacterial infections and antibiotic resistance among tungiasis patients in western, kenya*, PLoS neglected tropical diseases **11** (2017), no. 9, e0005901.

51. S. Francis, K. E. Wilke, D. E. Brown and E. E. Carlson, *Mechanistic insight into inhibition of two-component system signaling*, MedChemComm **4** (2013), no. 1, 269-277.

52. N. Velikova, S. Fulle, A. S. Manso, M. Mechkarska, P. Finn, J. M. Conlon, M. R. Oggioni, J. M. Wells and A. Marina, *Putative histidine kinase inhibitors with antibacterial effect against multi-drug resistant clinical isolates identified by in vitro and in silico screens*, Scientific Reports **6** (2016).

Chapter 2: Ca^{2+} dependent structural change and proposed collagenolysis mechanism of Col G

Abstract

Both Col G and Col H are dependent upon Ca^{2+} in order to degrade substrate. This chapter seeks to understand how *Hathewaya histolytica* is able to use the difference in Ca^{2+} concentration between bacteria and host to efficiently secrete active Col G. It also seeks to understand the specific Ca^{2+} domain rearrangement that must occur in Col G to fully activate it. Through structural analysis of CBD1BD2 linker as well as Small Angle X-ray Scattering (SAXS) envelopes of full-length Col G and non-catalytic Col G segments bound to mini-collagen we seek to understand how the enzyme binds to collagen. Through this we have been able to tease out a mechanism by which Col G processively degrades collagen.

2.1 Introduction

Hathewaya (*Clostridium*) *histolytica* is a gram-positive bacteria that is known to cause gas gangrene. The β-exotoxin of this bacteria consist of two enzymes, Col G and Col H, which seek out and degrade collagen fibers located in extracellular space. Destruction of collagen exposes new nutrient sources for the invading bacteria and helps spread the infection [**2**]. Both Col G and Col H have a similar multi-domain structure consisting of collagen binding domains (CBDs; CDB1 and CBD2 for Col G and CBD for Col H), polycystic kidney disease like-domains (PKD for Col G and PKD1 and PKD2 for Col H) and a catalytic domain, a zinc peptidase from the M9B family that is responsible for collagen hydrolysis (see figure 1) [**3**]. Duplicate CBD for Col G not present in Col H suggest that Col G is a much tighter binder to collagen than Col H. One of the goals of this research is to understand the role dual CBDs play in collagen hydrolysis for Col G.

CBD1 and CBD2 from Col G and CBD from Col H are homologous with each other with each CBD having approximately 120 residues and a β-sandwich "jelly roll" conformation composed of nine β-strands (see figure 2A) [**1**]. High-speed AFM video of Col G with a collagen "micro-ribbon" and SAXS experiment with disordered mini-collagen shows that the enzymes flocks to disordered regions in both collagen fibril and mini-collagen [**2**] [**3**]. It is the CBD that allows Col G and Col H to initially bind to these disordered regions in collagen. NMR and SAXS experiments with mini-collagen revealed that if disordered regions are not present then Col G CBD2 binds to the C-terminus of mini-collagen in a unidirectional manner [**4**].

Crystal structure of CBD2 revealed a hydrophobic region on the surface of Col G CBD2 [**5**]. Site-directed mutagenesis of these surface exposed hydrophobic residues alongside a collagen binding assay was able to initially map the collagen binding cleft for Col G CBD2. Study revealed that surface exposed aromatic residues are important for binding with a group of three Tyr residues; Tyr970, Tyr994 and Tyr996; being the most important [**5**]. Aromatic rings from Tyr residues may interact with the Pro residues of collagen through aromatic CαH-π interactions as demonstrated for model collagen like-peptides [**6**] [**7**]. Loss of Tyr994 reducing binding affinity to immeasurable levels [**5**]. Tyr994 may alson form a hydrogen bond with main chain of tropocollagen which may explain its huge impact on binding [**5**]. NMR titration experiments revealed the presence of other hot spot residues that help to form the protein binding interface [**4**]. Sequence alignment study with Col H revealed that the collagen binding residues are highly conserved with Tyr996 from Col G being replaced with Phe964 in Col H [**8**] [**1**].

Surface plasmon resonance results for Col G and Col H CBDs revealed that CBD2 from Col G binds to mini-collagen tighter than CBD from Col H by a factor of 10 [**9**]. Oddly enough, CBD1CBD2 from Col G does not bind to the substrate tighter than CBD2 alone suggesting that

for Col G, CBD1 binds to a different collagen substrate than CBD2 [**9**]. Fibril binding affinity for Col H CBD segment was found using a modified pull-down assay [**10**]. Col H CBD was found to have somewhat low affinity, 15.9 μM, and high capacity, 1.01 nmol/mg of collagen [**10**]. Insoluble collagen binding assay for Col G CBDs has yet to be performed however, for promoting callus formation bFGF conjugated to different non-catalytic segments from Col G and Col H was found to have the most pronounced effect when conjugated to CBD1CBD2 from Col G [**9**]. Currently, it is not entirely known why Col G non-catalytic segments have higher collagen binding affinities than Col H non-catalytic segments. Col H CBD is missing some of the collagen binding residues present in Col G CBD2 which may explain the difference in binding between Col G CBD2 and Col H CBD. The potential role of CBD1 in allowing Col G to bind to collagen tighter than Col H will be discussed further on in this chapter.

Gelatinase, either Col G or Col H missing their non-catalytic domains, was found to be able to degrade denatured collagen (gelatin) but not native collagen [**10**] [**11**]. The role of CBDs in the degradation of collagen fibrils is well reported [**8**] [**11**] [**1**]. In pancreatic islet isolation studies mixtures of Col G and Col H are used isolate pancreatic islets from donor tissue [**12**] [**13**] [**14**]. When Col G with truncated CBD2 (Col G -ΔCBD2) was used in place of wild type Col G pancreatic islet isolation was greatly inhibited [**12**]. The role of CBD2 in Col G processive collagen fibril hydrolysis will be discussed further down in this chapter.

In addition to the conserved hydrophobic residues in the collagen binding cleft, conserved Asn, Glu and Asp residues are also found located near the N-terminus of Col G and Col H CBDs [**1**]. Conserved N-terminal residues of Col G and Col H CBDs are able to chelate to two Ca^{2+} ions [**1**] (see figure 2B). In the presence of EDTA and other chelating molecules, binding affinity and collagen degradation activity for both Col G and Col H is significantly

lowered [**15**] [**5**] [**16**]. In the presence of excess Ca^{2+} CBD N-terminal linker changes from an α-helix to a β-sheet [**5**]. Change to β-sheet causes the domain to become more rigid protecting it from both chemical and thermal denaturation [**1**] [**16**]. The change from α-helix to β-sheet at the N-terminal linker is also known to prevent linker from protease degradation (see figure 2C) [**17**] [**18**]. Both PKD and catalytic domains also contain Ca^{2+} binding regions however, upon binding the change in structure for these domains is not as apparent as it is for the CBDs [**19**] [**15**]. It is speculated that *H. histolytica* exploit the difference in Ca^{2+} between host and bacteria to efficiently secrete Col G and Col H. Later on in this chapter the full-length structures of Col G in high and low Ca^{2+} concentrations are proposed and the potential domain rearrangements that occur upon Ca^{2+} binding are discussed.

2.2 Methods

2.2a. Production and Purification of Col G and Col H segments

The expression and purification of wild type Col G CBD1CBD2 and PKDCBD1CBD2 was performed at Okayama University Graduate School of Medicine Dentistry & Pharmaceutical Sciences in the Matsushita lab using previously described methods [**10**] [**20**]. Briefly, pGEX-4T-2 plasmids containing the sequences to be expressed were transformed into *E. coli* BL21 competent cells by heat shock. Cells were plated onto LB Agar plates supplemented with ampicillin and chloramphenicol and allowed to grow at 37° C. After 24 hours 10 ml of 2-YTG broth supplemented with ampicillin was inoculated with a single colony; broth was incubated at 37° C for 12 hours with agitation. 6 ml of new culture was transferred to 120 ml of 2-YTG broth containing ampicillin and the culture was incubated once again at 37° C for 12 hours with agitation. Finally, after 12 hours 25 ml of this culture was transferred to into 2 liters of 2-YTG containing ampicillin. Cells were incubated until OD_{600} of 0.7. IPTG was added to the 2-YTG

broth and expression was carried out for 4-8 hours. After addition of 20 mL of 100 mM PMSF, cells were harvested using centrifugation (7,000 RPM, 4° C, 15 min). Cell pellet was suspended in 40 ml of 1X PBS supplemented with 1 mM PMSF and French press (10,000 PSI) was used to lyse the cells into centrifugation tubes (20 ml). 1 ml of Triton X-100 was added to each tube and tubes were rotated for 30 min. (4° C). Centrifugation was performed twice to clear out cell debris (15,000 RPM, 4° C, 30 min).

The fusion proteins were purified using batch purification with glutathione sepharose beads. Junk proteins were eluted with the PBS wash. Fusion proteins were eluted with 50 mM Tris pH 8.0 containing 10 mM glutathione. Incubation with thrombin cleaves the GST tag and a second-round chromatography with the glutathione sepharose column was used to separate the protein from the tag. Q-sepharose ion exchange chromatography was used to further purify the proteins. Flow through fractions were pooled and dialyzed into 50 mM Tris, 1 mM $CaCl_2$ pH 7.5. Following dialysis, a BCA assay was used to quantify concentration for each sample.

2.2b. Identifying lower molecular weight contaminants in full length Col G and Col H from VitaCyte

Full-length Col G was purchased from VitaCyte for structural analysis. 10% SDS PAGE gels was ran for Col G to determine if contaminants are present in the samples. Gels were stained with comassie blue and destained with 10% Acetic acid/50% Methanol for 24 hrs. Analysis of SDS PAGE gel showed the presence of contaminants for Col G. ImageJ densitometry analysis of gel revealed that both Col G is sufficiently pure (>93% purity for both Col G) (see figure 3).

Bands were excised, washed twice with 25 mM ammonium bicarbonate, destained with ammonium bicarbonate/50% acetonitrile, and dried with pure acetonitrile. Then, the proteins

were reduced using 10 mM dithiothreitol in 50 mM ammonium bicarbonate for 1hr. at 55°C water bath. Proteins were alkylated by adding 10 mM iodoacetamide (in 50 mM ammonium bicarbonate). Reaction was carried out for 1 hr. in the dark. Next, gel pieces were washed with ammonium bicarbonate, and dried with pure acetonitrile. Enough trypsin gold (Promega, Mass spectrometry grade, 10 ng/µl in 25 mM ammonium bicarbonate) was added to cover the gel pieces and the mixture was incubated at 4°C for 30 min. After digestion, 100 µl of 25mM ammonium bicarbonate was added and the mixture was left overnight at room temperature. Solutions were transferred to separate Eppendorf tubes; left over tryptic peptides being extracted by vortexing with 100 µl of 50% acetonitrile/5% formic acid for 5 min. Solutions were completely evaporated and reconstituted in 50 µl of 0.1% formic acid and mass spectrometry analysis was performed.

2.2c. Mass Spectrometry analysis of Col G and Col H

Dr. Rohanna Liyanage in the mass spectrometry department at the University of Arkansas performed all trypsin digestions and mass spectrometry analysis. The tryptic digests obtained from all samples were analyzed by ESI-LC-MS/MS (Electrospray ionization liquid chromatography-tandem mass spectrometry) at the State-Wide Mass Spectrometry Facility at the University of Arkansas (Fayetteville, AR). Analysis was performed using data-dependent acquisition mode (DDA) with collision-induced dissociation (UltraScan mode, one min, 32,500 m/z per second). Agilent 1200 series micro flow HPLC system in line with Bruker Amazon-SL quadrupole ion trop ESI mass spectrometer (QIT-ESI-MS) was used for analyzing tryptic digests. Reverse-phase high performance liquid chromatography (RP-HPLC) with a Zorbax SB C18 column (150 mm x 0.3 mm, 3.5 µm particle size, 300 Å pore size, Agilent Technologies) and a solvent flow rate of 4 µL/min. was used for separating the tryptic digests. Gradient of 5%-

38% consisting of 0.1% formic acid and acetonitrile was used during elution. Auto MS mode was used to choose peaks for MS/MS analysis. Analysis was performed in positive ion mode using Bruker captive electrospray source with a dry nitrogen gas temperature of 200 °C with a flow rate of 3 L/min with enhanced scanning mode (8,100 m/z per second).

Peaks were picked in the LC-MS/MS chromatogram using default settings in Bruker Data Analysis 4.2 software. Bruker Proteinscape bioinformatics suite coupled with MASCOT 2.1 (http://www.matrixscience.com) was used to search and match LC-MS/MS data against Col G FASTA sequences in the uniprot format (Col G: Q9X721). Scaffold (version Scaffold_4.8.4, Proteome Software Inc., Portland, OR) was used to validate MS/MS based peptide and protein identifications [**21**]. Scaffold software was used to give high confident peptide identifications using identification probability and false discovery rates. Sequence coverage for each band revealed that small amounts of Col G lose C-terminal domains during protein expression (see figure 3).

2.2d. Col G SAXS studies

Small Angle X-ray Scattering (SAXS) was used to find a low-resolution envelope full length Col G as well as Col G non-catalytic segments bound to mini-collagenmini-collagens. All mini-collagens were synthesized by Dr. Takaki Koide using Fmoc-based solid-phase method on Rink amide resin [**22**]. For Col G, SAXS data was collected for PKDCBD1CBD2 complexed to $[(POG)_{10}]_3$ and for the Col G pCa series ($pCa = -\log[Ca^{+2}]$). In addition to the pCa series for Col G, SAXS data for a pCa series of Col G with truncated CBD2 was also collected (-ΔCBD2 Col G). Data for Col G CBD1CBD2:$[(POG)_{10}]_3$ complex previously collected by Dr. Ryan Bauer was processed to help position PKD domain.

For the complex, a 5 mg/ml stock of $[(POG)_{10}]_3$ was created by dissolving min-collagen powder in 5 mM Acetic Acid. The solution was incubated overnight at 4° C and circular dichroism was used to confirm that the triple helix was formed [**23**]. Then a 1:1 mole ratio of mini-collagen to protein was incubated at 4 °C overnight, and then dialyzed into 50 mM Hepes pH 7.5, 100 mM NaCl, 5 mM $CaCl_2$ [**8**]. For the Col G pCa series, Col G was exchanged into pCa 3, 4, 5, 6 and 7 buffers using Amicon Ultra centrifugal units. pCa buffer composition is listed in Tables 1 and 2. Different glycerol concentrations (0%, 0.5%, 1.0%, 1.5% and 2.0%) for each pCa were tested using continuous native PAGE gel electrophoresis to find the optimal amount of glycerol needed to prevent aggregation for each pCa buffer [**24**]. Glycerol was incorporated into pCa buffers by adding double the concentration listed into the HBS-EGTA buffer. For Col G and Col -ΔCBD2 the following concentrations of glycerol were used Col G pCa 7: 1%, Col G pCa 6: 0.5%, Col G pCa 5: 0%, Col G pCa 4: 0%, Col G pCa 3: 0%, Col G -ΔCBD2 pCa 7: 1.5%, Col G -ΔCBD2 pCa 6: 0.5%, Col G -ΔCBD2 pCa 5: 0.5%, Col G -ΔCBD2 pCa 4: 0.5% and Col G -ΔCBD2 pCa 3: 0%.

Table 1

Buffers	Composition
HBS-EGTA	10 mM Hepes pH 7.5, 100 mM NaCl, 0.4 mM EGTA
HBS-Ca	10 mM Hepes pH 7.5, 100 mM NaCl, 2.4 mM $CaCl_2$
HBS	10 mM Hepes pH 7.5, 100 mM NaCl

Table 2

Buffer	HBS-EGTA (ml)	HBS-Ca (ml)	HBS (ml)
pCa 3	50	50	0
pCa 4	50	12.493	37.507
pCa 5	50	8.681	41.319
pCa 6	50	7.732	42.268
pCa 7	50	4.541	45.459

2.3 Results and Discussion

2.3a. Small angle X-ray Scattering of Col G targeting segments

To understand Col G Ca^{2+} dependent structural change and how the enzyme processes its substrate SAXS data was collected for the following samples for full length Col G and Col G with truncated CBD2 (Col G -ΔCBD2) at different Ca^{2+} concentration and PKDCBD1CBD2 bound to mini-collagen. In addition to collecting new data, scattering data for PKDCBD2CBD2 and pCa 3 and pCa 6 and CBD1CBD2:$[(POG)_{10}]_3$ were reprocessed. Previously derived SAXS envelope for CBD1CBD2 at pCa 3 was used to help position segments in Col G both free and bound to substrate. Due to the formation of air bubbles in the 96 well plate as well as protein aggregation; data for only certain pCa levels of Col G and truncated Col G were useable (pCa = -log[Ca^{2+}]) (Col G: pCa 3, pCa 6; Truncated Col G: pCa 4, pCa 6, pCa 7).

SAXS profiles, guinier plots, pair distribution plots and Kratky and plots for Col G samples are shown in figures 4, 5, 6 and 7. Guinier plots indicate that aggregation is minimal

among samples. According to Kratky plots, all samples are folded with the exception of PKDCBDCBD at pCa 6 which appears to be somewhat unfolded (see figure 4D) [**25**]. Unfolded appearance in kratky plot could be due to lack of calcium causing N-terminus of CBD1 and CBD2 to become extended [**1**] [**5**] [**18**]. SAXS data for PKDCBDCBD at pCa will need to be recollected and reprocessed in order to full confirm its shape.

Radius of gyration (R_G) for a protein is defined as the square root of the average squared distance of each scattered x-ray beam from the center of the molecule. At low resolution x-ray scattering can be approximated by using the equation $\ln I(q) = \ln I(0) - qR_G^2/3$ with the slope of the line being used to find R_G [**26**]. High value for R_G can be used to identify aggregation in a protein [**26**]. Pair distribution plots give the various vectors found in a molecule and are created via fourier transformation of initial x-ray scattering data. D_{Max} found via pair distribution plot is the maximum diameter present in a molecule [**26**]. D_{Max} values are a way to visualize conformational changes in a sample and are used in the development of SAXS derived envelopes. Pair distribution functions when integrated across all r values can give R_G. R_G value found this way is often used as consistency check for R_G found via Guinier approximation [**26**]. Porod exponent is calculated from flexibility and volume analysis function in Scatter and is based on the rate of decay for scattering intensity in the initial I(q) vs q intensity plot. For an ideal folded molecule P = 4 and for an unfolded molecule P = 1 [**27**]. In order derive SAXS envelopes first multiple envelope calculations are averaged and then filtered. To determine if the dummy atoms used in each envelope for calculating the average agree with each other normalized spatial discrepancy (NSD) values are used. Envelopes are superimposed with each other and distance between dummy atoms is used in calculating NSD values; if value is close to

zero it indicates high agreement [**26**]. Q-range is the range of scattering values from initial scattering profile used to create pair distribution plot.

Agreement between experimental scattering profile and theoretical profile from crystal structure coordinates for each envelope was found using Crysol from Atsas online server [**28**]. Results are listed in Appendix. The R_g, D_{max}, P, NSD and Q-range values for all samples are summarized in tables 1 & 2. P values are not listed for non pCa samples due to relevance. Porod exponent (P) values for Col G and Col G ΔCBD2 decreased as the pCa increased.

Table 3

	pCa	R_g (Å) Guinier Approx.	R_g (Å) Defined by P(r)	D_{Max} (Å)	Mean NSD value	Q-range ($Å^{-1}$)
CBD1CBD2	3.0	24	24	75	0.714	0.0159 – 0.3605
CBD1CBD2	6.0	27	27	128	0.540	0.0236 – 0.2198
CBD1CBD2:$[(POG)_{10}]_3$	2.3	29	30	113	0.572	0.0187 – 0.2801
PKDCBD1CBD2	3.0	31	30	98	0.653	0.0215 – 0.2572
PKDCBD1CBD2	6.0	39	35	120	1.392	0.0151 – 0.3483
PKDCBD1 CBD2:$[(POG)_{10}]_3$	2.3	38	38	155	0.607	0.0159 – 0.1965

Table 4

	pCa	R_g (Å) Guinier Approx.	R_g (Å) Defined by P(r)	D_{Max} (Å)	P	Mean NSD value	Q-range ($Å^{-1}$)
Col G	3	47	47	150	3.2	0.881	0.0129 – 0.3654
Col G	6	46	48	170	2.8	0.776	0.0111 – 0.2690
Col G ΔCBD2	4	43	42	130	3.4	0.821	0.0129 – 0.3220
Col G ΔCBD2	6	43	44	143	2.0	0.764	0.0129 – 0.3220
Col G ΔCBD2	7	42	43	145	1.6	0.710	0.0152 – 0.2732

2.3b. Structural analysis of CBD1CBD2 linker

At physiological concentrations of Ca^{2+} SAXS revealed that Col G becomes more rigid. The solved crystal structure of CBD1CBD2 was analyzed in order to help determine how CBD1 binds. In the presence of Ca^{2+} CBD1 is related to CBD2 by a pseudo two-fold rotational symmetry. Interdomain interactions consisting of four salt-bridges and three hydrogen bonds found in the linker region may restrict freedom of movement for CBD1 and CBD2 (see figure 8) [**8**]. For CBD2 in the presence of Ca^{2+} N-terminal linker are resistant to trypsin [**18**]. However

the C-terminal linker region in CBD1 has been shown to be susceptible to trypsin like activity [**14**].

The linker between CBD1 and CBD2 is not long enough for CBD1 and CBD2 to bind to the same tropocollagen in *cis* fashion as observed for CNA from *Staphylococcus aureus* [**29**]. When CBD1 and CBD2 were hypothetically positioned on the same tropocollagen, the distance between Glu895 of CBD1 and Lys896 of CBD2 was estimated as 40 Å. When pCa is greater than 6, the linker may still not be long enough for tandem CBD to interact with one tropocollagen, but it may be dynamic enough to allow for CBD1 and CBD2 to bind to two adjacent hexagonally arranged tropocollagen molecules. When pCa is smaller than 4, the orientation of CBDs would allow for Col G to wedge between two parallel oriented collagen molecules [**8**].

2.3c. SAXS derived envelopes for Col G segments bound to mini-collagen

To construct the PKDCBD1CBD2:$[(POG)_{10}]_3$ complex, the following information was used. The rod-like difference in electron density between PKDCBD1CBD2 at pCa 3 and CBD1CBD2:$[(POG)_{10}]_3$ (green and pink, respectively) was assigned as $[(POG)_{10}]_3$. Since CBD2 binds tighter to mini-collagen than CBD1, CBD2 segment was considered to bind $[(POG)_{10}]_3$. Binding surface of CBD2 mapped by mutagenesis and collagen-binding studies was positioned to interact with $[(POG)_{10}]_3$ [**5**]. Previous SAXS and NMR work on the binding of CBD2 to mini-collagen showed that CBD2 binds asymmetrically to the C-terminus of $[(POG)_{10}]_3$ [**4**] [**3**]. In our complex, we reasoned that the asymmetric binding to be observed by CBD2 segment of CBD1CBD2. SAXS envelopes previously determined for CBD2, $[(POG)_{10}]_3$, and CBD2:$[(POG)_{10}]_3$ were also used to position CBD2 and $[(POG)_{10}]_3$ in the envelope calculated for PKDCBD1CBD2:$[(POG)_{10}]_3$ [**4**] [**3**]. In the remaining non-overlapping electron density, PKD

was positioned so that C-terminal residue (Asn770) of PKD is as close as possible to N-terminal residue (Thr773) of CBD1CBD2 (see figure 9).

2.3d SAXS derived envelopes for full-length Col G at low and high Ca^{2+} concentrations

At Ca^{2+} concentration inside bacteria (pCa 6) Col G adopts an elongated shape with a D_{Max} of 172 Å. In order to identify the position of CBD1CBD2 (PDB: 5IKU), PKD (PDB: 4TN9) and catalytic domain (PDB: 2Y3U) in Col G pCa 6 the envelopes for PKDCBD1CBD2 pCa6 and Col G ΔCBD2 pCa 6 were used. To help find the position of CBD1CBD2 and PKD in Col G pCa 6 previously collected scattering data for PKD1CBD1CBD2 at pCa 6 was reprocessed. Data points with high intensities were removed, as they correspond to radiation damage, and the new elongated envelope for PKD1CBD1CBD2 was positioned at the tail end of the envelope of Col G pCa 6. In order to create PDB file for CBD1CBD2 at pCa 6 the PDB file for CBD1CBD2 bound to Ca^{2+} was manipulated by Dr. Ryan Bauer (a former graduate student in the Sakon lab). N-terminal linkers in CBD1 and CBD2 were changed from β-sheets to alpha-helix using MiFit to resemble the potential structure of the CBD1CBD2 in the absence of Ca^{2+} [**5**]. New structure was found to fit SAXS envelope for CBD1CBD2 at pCa 6 [**8**]. Previously derived SAXS envelope for CBD1CBD2 at pCa 6 was used to help determine the position of linkers at low Ca^{2+}. CBD1CBD2 pCa 6 was positioned at the tail-end of the Col G envelope and PKD was positioned in the remaining PKDCBD1CBD2 pCa 6 electron density. To help confirm the position of CBD1CBD2 the envelope for Col G ΔCBD2 pCa 6 was superimposed with the full-length enzyme and CBD2 was found in the non-overlapping region between the two (see figure 10A).

The distance from the N-terminus of CBD1CBD2 to the C-terminus of PKD was 20 Å (see figure 10B). For CBD1CBD2 at pCa 6 20 residues in the N-terminal linker were deleted due

to overlapping with crystal structure for PKD leaving new N-terminal residue as Ala790. C-terminal residue for PKD is Asn770. 20 residues missing between Asn770 and Ala790 should be more than enough to bridge the 20 Å gap between CBD1CBD2 at pCa 6 and PKD. The catalytic domain was positioned in the remaining electron density. The distance from the N-terminus of PKD to the C-terminus of the catalytic domain was 29 Å. N-terminal residue of PKD crystal structure is Gly685 (Glys685 is Lys in sequence from uniprot: Q9X721) and C-terminal residue of catalytic domain crystal structure is Gly680 (Ala in uniprot). Missing 5 residues between PKD and catalytic domain would at most only make up 7 Å. Positioning domains via multiple SAXS envelopes only tells us their relative positions. It could be that position of PKD and catalytic domain need to be refined in order to make up the remaining distance between the two. Currently experiments are being performed using SASREF from ATSAS online server to refine positions of the domains [**30**].

At extracellular matrix Ca^{2+} concentrations (pCa 3) Col G adopts a more defined and compact structure with a D_{Max} of 150 Å (see figure 6 and 11). In order to identify the position of each segment in Col G pCa 3 the envelopes for PKDCBD1CBD2:$[GPRG(POG)_{13}]_3$, Col G ΔCBD2 pCa 4 and PKDCBD1CBD2 pCa 3 were used. To determine the location of CBD1CBD2, PKDCBD1CBD2:$[GPRG(POG)_{13}]_3$ envelope was superimposed with the envelope for Col G pCa 3. CBD2 was positioned in the overlapping region near where mini-collagen is known to be. To confirm the position of CBD2 the envelope for Col G ΔCBD2 pCa 4 was superimposed with Col G pCa 3 envelope. Previous research has shown that the orientation of segments in Col G at pCa 3 and pCa 4 should be nearly identical [**8**] [**5**] [**16**]. CBD2 was found in the non-overlapping region. The scattering data for PKDCDB1CBD2 pCa 3 was previously collected. Data points corresponding to radiation damage were removed and the newly derived

envelope for PKD1CBD1CBD2 was positioned at the end of the envelope. Envelope for PKDCBD1CBD2 pCa 3 was superimposed with Col G pCa 3 envelope. The position of CBD1CBD2 is known and PKD was positioned in the remaining electron density (see figure 11).

The distance from the N-terminus of CBD1CBD2 to the C-terminus of PKD was 37 Å. For CBD1CBD2 N-terminal residue is Thr773 and for PKD C-terminal residue is Asn770; 2 residue gap between CBD1CBD2 and PKD would not be enough to cover 37 Å. In crystal structure of CBD1CBD2 N-terminal linker is pointing away from the C-terminus of PKD. Current position of CBD1CBD2 may just be a crystallographic artifact and PKDCBD1CBD2 position is different. Finally, the catalytic domain was positioned in the remaining electron density. The distance from the N-terminus of PKD to the C-terminus of the catalytic domain was 23 Å. For PKD N-terminal residue is Asn685 and for catalytic domain C-terminal residue is Gly680; 4 residue gap between PKD and catalytic domain would not be enough to cover 23 Å. Similar to the distance between PKD and catalytic domain for Col G at pCa 6 the position of these two domains may not be optimal and further work would need to be done in order to find the optimal position for PKD and catalytic domain within the envelope for Col G at pCa 3.

2.3e. Ca^{+2} induced domain rearrangement and secretion of Col G

The difference in Ca^{2+} concentration inside *H. histolytica*, which is likely similar to the concentration inside *Escherichia coli* (0.2-0.3 μM; pCa 6.5-6.7) and extracellular matrix of host (~1.2 mM; pCa~2.9) could be exploited by *H. histolytica* to facilitate rapid secretion into the host and subsequent activation [**31**] [**32**]. Based on the size of the Col G envelope gram positive type II secretion systems are the most likely candidates to secrete the enzymes into hosts [**33**]. According to PyMol the widest portion of Col G has a measured width of 75 Å at pCa 3 and 67 Å at pCa 6. The inside diameter of a type II injectosomes for *Vibrio cholerae* has a width of 75 Å

(see figure 12) [**34**]. Based on the widths of both the enzyme and the channel it would be difficult for Col G to be efficiently secreted at pCa 3. The dimensionless Kratky plots and P values for Col G at pCa 3 and pCa 6 shows that at lower Ca^{2+} concentrations Col G is more flexible (see figure 6D and table 4). The increased flexibility and narrower structure for Col G at pCa 6 may permit the protein to be secreted through the type II secretion system.

Type II secretion system proteins are found not only in *H. histolytica* (Accession: VTQ86622) but also in *H. proteolytica* (Accession: WP_072902510) and multiple other *Clostridium* species (Accession: WP_035784777, AMN33782, KOR55552). Quorum sensing may optimize *Hathewaya histolytica's* ability for infection. Collagenases would only be secreted once a certain number of bacteria have gathered at the potential infection site or if optimal conditions for survival are present [**35**]. This would ensure that a maximum number of collagenases are secreted into host tissue in order for the bacterial infection to proceed.

In Col G there is a major difference in the position and orientation of the PKD and catalytic domains between pCa 6 and pCa 3 (see figure 13). This difference in positions may be explained by the change in the N-terminal linker from α-helix to β-sheet in CBD1 and CBD2 [**5**] [**1**]. PKD domains also chelate to Ca^{2+} ions however, no large structural change is observed when Ca^{2+} chelates to its N-terminus linker [**19**]. Transition from α-helix to β-sheet at CBD1 N-terminal linker may cause PKD domain to rotate along the z-axis approximately 90°. This hypothetical repositioning of PKD domain may then cause the catalytic domain to be rotated along the y axis until the active site cleft of the peptidase sub-domain is aligned with the collagen binding cleft of CBD1. Both Col G and Col H activities toward collagen fibrils are significantly impaired at low concentration of Ca^{2+} [**20**] [**5**] [**16**]. The primary function of the PKD domain in

Col G may be to position the catalytic domain for collagenolysis and inactivity at low Ca^{2+} concentration is due to the sub-optimal position of the catalytic domain.

2.3f. SAXS revealed that CBD2 is used to anchor enzyme to collagen and CBD1 may be used to feed substrate into active site

The opposing positions of the collagen-binding clefts on tandem CBD present unique modes for the domain to latch onto collagen fibril. Whereas collagen fibril is built from a staggered array of triple-helical tropocollagen, and is water insoluble, synthetic mini-collagen, which mimics the tropocollagen structure, is water-soluble and allows solution-based analysis of CBD-collagen interaction. Measured dissociation constants for the interaction between the individual CBDs and either mini-collagen or fibril tend to agree [**20**]. Tandem CBD is the tightest binder to collagen fibril, and can bind tighter than the sum of CBD1 and CBD2 individually [**20**] . However tandem CBD binds mini-collagen about as tightly as CBD2 alone [**20**]. Corroborating this, the SAXS-derived shapes of the CBD1CBD2:$[(POG)_{10}]_3$ complex and the PKDCBD1CBD2:$[(POG)_{10}]_3$ complex are consistent with a one-to-one complex, and CBD2 is positioned to interact with mini-collagen (see figure 9) [**8**].

The binding affinities of CBD1 and CBD2 to mini-collagen suggest CBD2 initiates collagen-binding. Furthermore, the sequence alignment of multiple CBD binding segments suggests this binding strategy is conserved amongst the collagenases possessing tandem CBDs [**19**]. While the binding clefts of both CBDs complement the curved surface of collagen, the three most critical tyrosine residues for collagen-binding in CBD2 (970, 994, and 996) are conserved amongst CBD2s, while only two tyrosine residues 851 and 875 (equivalent to 970 and 994) are conserved amongst CBD1s. Gene duplication of CBD required reduced binding affinity in the CBD1 to prevent the tandem CBD from binding too tightly to collagen fibril. The tighter

binding CBD2 could initiate binding and serves the leading role, while CBD1 plays an auxiliary, yet pivotal role in collagenolysis.

In order to tease out how collagen substrate is being fed into the catalytic domain of Col G the envelope for full length Col G was superimposed with the envelope for PKDCBDCBD:$[(POG)_{10}]_3$. Analysis of the superimposed structures reveals that CBD1 could be used to feed the substrate into the catalytic domain (see figure 14). As previously stated structural analysis of the linker between CBD1 and CBD2 even at low Ca^{2+} concentration indicated that it is highly unlikely that CBD1 could bind to the same tropocollagen molecule as CBD2 [**8**]. The orientation of CBD1 and CBD2 in the crystal structure of tandem CBD (PDB: 5IKU) suggest that Col G may be able to bind to two parallel oriented collagen molecules rather than one collagen molecule. Surface Plasmon Resonance results have indicated that CBD1CBD2 and CBD2 bind to mini-collagen with equal affinities [**9**]. The binding affinities of CBD1CBD2 and CBD2 to mini-collagen suggest CBD2 initiates collagen-binding acting as an "anchor" for Col G. If the tropocollagen bound to CBD2 were degraded Col G would most likely dissociate from collagen fibril. CBD1 with its lower affinity for mini-collagen may be better suited for "sliding" the collagen substrate into the catalytic domain. This ability to wedge between opposing tropocollagens could also explain the curious ability of CBD1CBD2 to promote fibril formation at low concentrations. Col G may promote fibril formation when it is stalled in order to provide CBD2 with more points to anchor to collagen fibril [**8**].

2.3g. Potential collagenolysis mechanism for Col G

Col G's ability to bind to parallel oriented tropocollagen molecules and degrade only the substrate corresponding to CBD1 was inferred using SAXS. High-speed atomic force microscopy has been used visualize the interplay of Col G and collagen [**2**]. In the AFM videos

Col G was seen to bind to disordered regions of collagen and continuously degrade substrate without being released until reaching the N – terminus of collagen fibril [**2**]. Below we explain the mechanism behind Col G ability to processively degrade collagen substrate.

The approximate 55 Å spacing of binding clefts in tandem CBD may exclude Col G from reaching tropo-collagen molecules within the tightly packed D-periodic structure of collagen micro-fibril. However, the spacing would allow for Col G to seek either inter-fibrillary spaces or crevices on the surface of damaged collagen fibril. The tandem CBD's role could be to intercalate into the disturbed region. High concentration of Col G may prevent fibril formation by physically keeping tropocollagen molecules from coming together. The extrafibrillar matrix is the region between collagen fibrils and consists of polysaccharides, proteoglycans and non-fibril collagens that are necessary for the formation of collagen fibers [**36**]. The unique tandem CBDs of Col G may make it better suited towards cleaving some of auxiliary, non -fibril collagens present in the extrafibrillar matrix that hold collagen fibrils together to form fibers [**37**].

Subsequent C-terminus to N-terminus movement is thought to be driven by substrate hydrolysis. Catalytic module is well suited to be located behind tandem CBD for it is a tripeptidyl carboxypeptidase [**2**]. This domain arrangement will allow for greater ratio of binding sites for substrate than product thus it processively moves towards the N-terminus of collagen. Each CBD unidirectionally binding to collagen may keep N-side of collagen apart during polarized movement. Tandem CBD utilizes collagen molecules as railings so that catalytic assembly will remain close to the substrate. As stated earlier SAXS analysis was able to determine that CBD1 is used to feed collagen substrate into the catalytic domain. After hydrolysis, Col G will begin to move towards the N-terminus. CBD1, PKD domain and catalytic domain begin to move first but tighter binding CBD2 remains anchored to its collagen fibril. It is

speculated that this exaggerated strain could be useful in dislodging the tighter binding CBD2 (see figure 15). To support this "walking" mechanism for Col G to move along collagen unpublished high speed atomic force microscopy (AFM) video of the enzyme alone shows that the enzyme is dynamic and moves as two segments: CBD1CBD2 being one and the catalytic domain being the other. It is unknown which segment the PKD domain belongs to (see video 1 in Appendix).

We speculate that PKD domain in Col G may help to move the enzyme along collagen. In SAXS envelope for fully active Col G (pCa 3 equivalent to ECM concentration of Ca^{2+} in the body) the PKD domain is rotated approximately 90° perpendicular to both CBD1CBD2 and the catalytic domain. This orientation of PKD domain in Col G may make the domain act as a torsion spring to help dislodge the enzyme from collagen. Released energy from hydrolysis may cause PKD domain to be exit from its 90° "twisted" orientation releasing stored mechanical energy. This release of built up energy may turn cause CBD1 to be pushed towards the N-terminus. Finally, enough strain is built up in CBD2 that it is eventually dislodged. B-factor analysis of Col G PKD domain shows that its midsection is dynamic in the presence of Ca^{2+} which may give support to the idea of PKD domain as a spring [**19**].

When Col G was stalled, it isolated collagen fibrils and rearranged them to thicken neighboring fibrils [**2**]. Enzymes benefits from destabilizing substrate to carry out hydrolysis, but Col G apparently can stabilize its substrate. The tandem CBD can facilitate collagen fibril formation but only at low concentration. The benefit for a collagenase's counterintuitive function is speculated. When Col G encountered tropocollagen molecules that are not oriented in parallel, tandem CBD may not be able to bind to two tropocollagen molecules to processively hydrolyze

its substrate. Col G's ability to reorient collagen to fibrillar form may allow for efficient collagenolysis. Col G's ability to stabilize collagen could become beneficial in that context.

2.3h. Structural explanation for successful applications of Col G CBD1CBD2 for localized drug delivery

Currently non-catalytic segments from both Col G and Col H have been successfully used in the localized delivery of bFGF to bone fractures in mice models [**9**] [**38**] [**39**] [**40**] [**41**]. Bone growth was found to occur quickiest when bFGF was linked to either CBD1CBD2 or CBD. In study from Sekiguchi et al. 2018, surface plasmon resonance was used to find binding information for CBD1CBD2 and CBD2 to mini-collagen. Dissociation constants for CBD1CBD2 and CBD2 were found to be virtually identical (CBD2 $K_D = 4.54 \times 10^{-5}$ M; CBD1CBD2 $K_D = 4.46 \times 10^{-5}$ M) [**9**]. In the same study, different Col G and Col H segments were linked to bFGF and their ability to heal bone fracture in mice was tested. It was shown that bFGF linked to CBD1CBD2 had the highest efficacy for bone repair [**9**]. To explain the two almost contradictory findings from Sekiguhi et al. 2018, *in vivo* CBD1CBD2 may be wedging between parallel oriented tropocollagen either within a micro-fibril or between collagen fibrils [**42**]. In surface plasmon resonance results dissociation constants are nearly identical because linker between CBD1 and CBD2 in CBD1CBD2 is not flexible enough to allow CBD1 to bind to the same tropocollagen as CBD2. The 55 Å distance between CBD1 and CBD2 may allow for CBD1CBD2 to delivery biomolecules directly to damaged collagen; in orgel structure of micro-fibril loss of three tropocollagen would give a 45 Å space for CBD1CBD2 to bind [**43**]. Anchoring growth factors to damaged regions minimizes the dosage needed and potential side effects [**40**] [**44**]. CBD1CBD2's ability to wedge between opposing collagen fibrils makes it bind tighter than CBD2 alone, making it ideal for delivery of biomolecules to promote localized

affects [**20**] [**9**]. Below we speculate on ways that would make CBD1CBD more effective at delivering bFGF.

Crystal structure of CBD1CBD2 shows that N-terminus of CBD1 is close to the binding cleft of CBD1. When wedging between collagen molecules linker between bFGF and CBD1 needs to be long enough to avoid interacting with the CBD1 binding cleft and neighboring collagen micro-fibrils. Based on collagen micro-fibril structure 45 Å should be long enough [**43**]. Repeats of Gly-Gly-Gly-Gly-Ser are commonly used to provide flexibility and help maintain stability of linkers in aqueous solvents, while Pro residues are used to prevent linkers from becoming too flexible [**45**]. Research will need to be done to determine the optimal residue content needed for linking bFGF to CBD1. Since bFGF promotes cell proliferation directly through its receptor it needs be efficiently released from CBD1CBD2 to have the greatest effect [**46**] [**47**] [**48**]. To promote the release of bFGF, the active site for HTRA1, MMP or Cathepsin K can be inserted into the linker region. These enzymes are involved in osteoclast or osteoblast regulation and are likely to be found near the graft site for bFGF-CBD1CBD2.

2.3i. Preventing loss of CBD2 is necessary for optimization of Pancreatic islet isolation

Pancreatic islet transplantation is an experimental treatment for type I diabetes and pancreatic cancer in which a blend of Col G and Col H are used to harvest healthy pancreatic islets from a deceased donor, or from the recipients own pancreas [**49**]. Islets are injected into the portal vein of the recipient's liver. Once the islets have entered the liver, they begin to release insulin, restoring pancreatic function [**49**]. Using the liver as the site for transplantation allows for an efficient supply of oxygen and nutrition to the transplanted islets [**50**]. SDS PAGE and MALDI-TOF mass spectrometry of Col G purchased from VitaCyte revealed that CBD2 is lost for some of the enzyme during expression (see figure 3). While CBD1 has some binding affinity

for collagen fibril its affinity is lower than that for CBD2 [**51**]. Intact Col G is very important in the isolation of pancreatic islets and loss of CBD2 negatively impacts the number of healthy pancreatic islets isolated [**52**] [**13**] [**12**].

Peptide cutter from Expasy was used to analyze the linker region between CBD1 and CBD2 (residue 881 to 901) revealed that the linker, contains four Lys residues and may be susceptible to cleavage by trypsin and clostripain [**8**] [**53**]. Mass spectrometry analysis of two version of truncated Col G revealed that one version is cleaved on the C-terminus side of Lys896 and the other is cleaved on the C-terminus side of Lys908 [**14**]. To explain how Col G may lose CBD2 in pancreatic islet isolation, active trypsin may be present. In addition to trypsin, another likely candidate for the loss of CBD2 is clostripain. Clostripain is used alongside Col G and Col H in pancreatic islet isolation and is responsible for trypsin like activity during isolation [**13**]. While it prefers to cleave the C-terminal side of Arg it will also cleave on the C-terminal side of Lys [**13**]. Mutation of all four Lys residues would potentially make CBD1CBD2 resistant to trypsin and clostripain and would increase yield of healthy islets.

2.4 Conclusion

SAXS derived envelopes for Col G at low and high Ca^{2+} concentrations were able to reveal the potential domain rearrangement that occurs within the enzyme necessary for activation after being secreted into the host. We speculate that the shift in the PKD domain is very important for activation of the enzyme and may help convert the chemical energy from hydrolysis into mechanical energy of movement. Col G PKDCBD1CBD2 complexed to mini-collagen showed that CBD1 and CBD2 do not bind to the same tropocollagen molecule and that PKD domain appears to be primarily responsible for positioning the catalytic domain [**51**]. SAXS derived envelope of active Col G superimposed with PKDCBD1CBD2 in a complex with mini-collagen revealed that CBD1 is used to feed collagen substrate into the active site of the catalytic domain. Knowledge of the three-dimensional structure of active Col G,

at pCa 3, was able to show that CBD2 is necessary for the enzymes polarized movement on collagen fibrils and for the processivity seen in the enzyme [**47**] [**51**]. Linker between CBD1 and CBD2 may be exposed for proteolysis. For Col G to be a more effective medical tool (either for drug delivery or collagen degradation) future research should focus on how to best strengthen the linker between CBD1 and CDB2.

Chapter 2: Figures

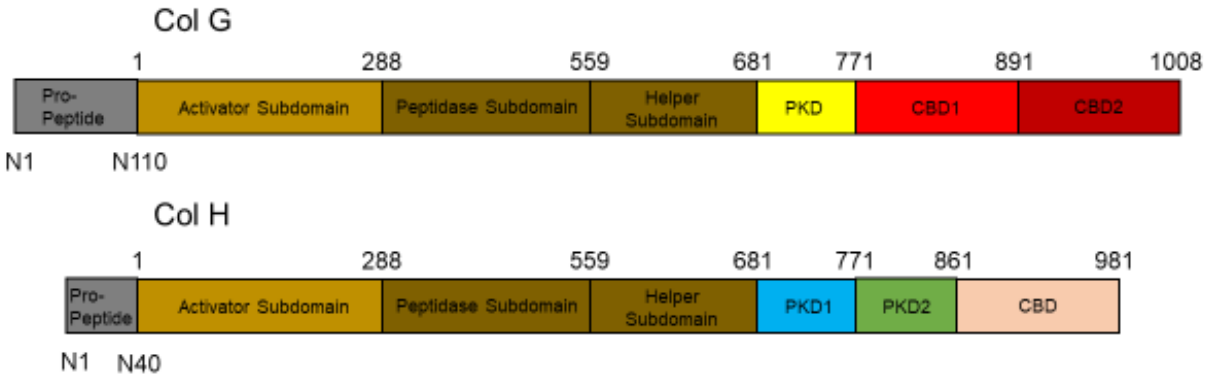


Figure 1: Domain map of Col G and Col H from *H. histolytica*. In order to activate Col G and Col H Pro-peptide (gray, Col G: N1 – N110 Col H: N1-N40) must be cleaved. Catalytic domain of Col G and Col H are composed of Activator sub-domains (light brown, residues 1 – 288) and Peptidase subdomain (dark brown, residues 289 – 559), which itself contains a helper subdomain to help the domain fold correctly (dart brown, residues 560 – 681). PKD domains (single PKD for Col G in yellow, residues 682 – 771; PKD1 and PKD2 for Col H in blue and green, residues 681 – 771 for PKD1 and residues 772 – 861 for PKD2) connect catalytic domains of both enzymes to CBDs. CBDs (CBD1 and CBD2 for Col G in red and dark red, residues 772 – 891 for CBD1 and residues 892 – 1008 for CBD2; single CBD for Col H in pink, residues 862 – 981) are responsible for binding both enzymes to collagen

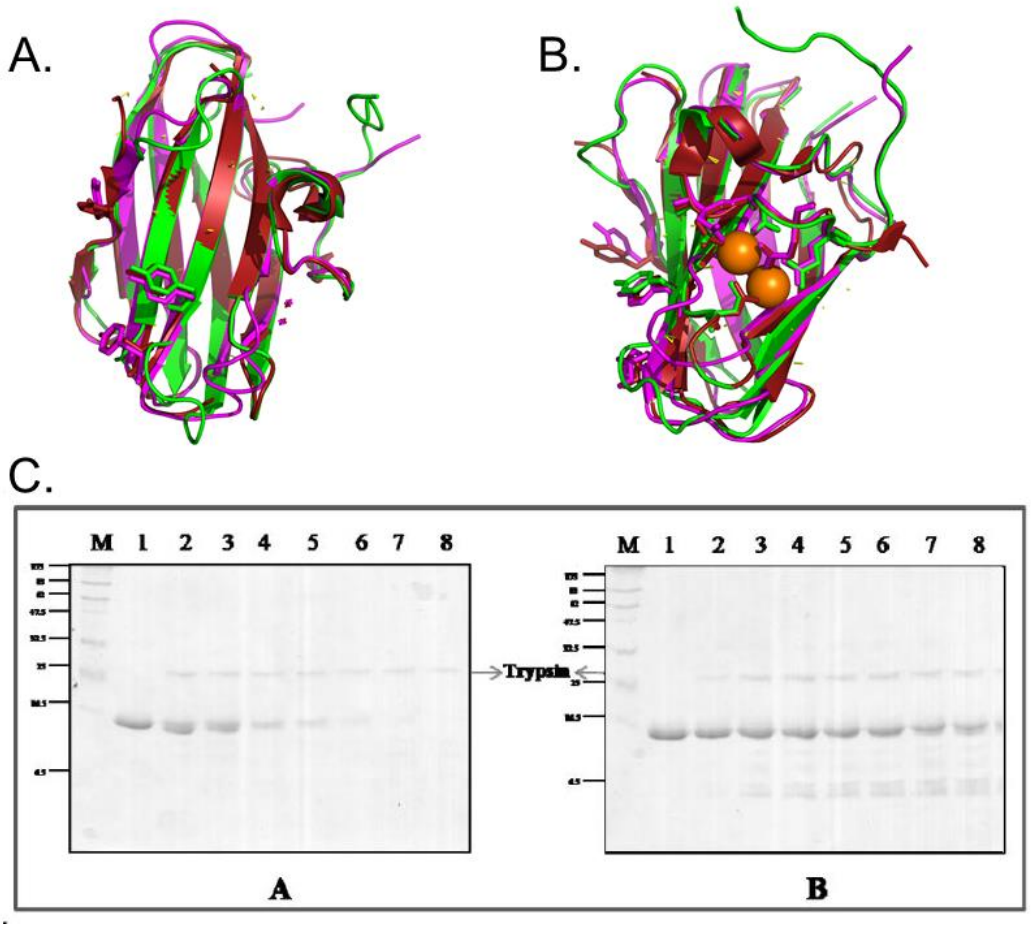


Figure 2: **A & B.** Structural alignment of Col G CBD2 (red), CBD1 (green) and Col H CBD (magenta). R.M.S.D. CBD2 to CBD1 = 0.656 Å, R.M.S.D. CBD2 to CBD = 0.609 Å, R.M.S.D. CBD1 to CBD = 0.572 Å. **A.** Key residues responsible for binding to collagen are shown as sticks. **B.** Ca^{2+} residues are shown as sticks. **C.** Proteolytic digestion of apo CBD2 (a) and holo CBD2 (b). from Philominathan et al. 2009. In the absence of Ca^{2+} CBD2 is degraded within 2.5 hrs. for holo CBD2 protein is still present after 3.5 hrs. Legend for SDS PAGE wells 1 - Control, 2 - 0.5 hrs., 3 - 1 hrs., 4 - 1.5hrs., 5 - 2hrs., 6 - 2.5 hrs., 7 - 3 hrs., 8 - 3.5 hrs.

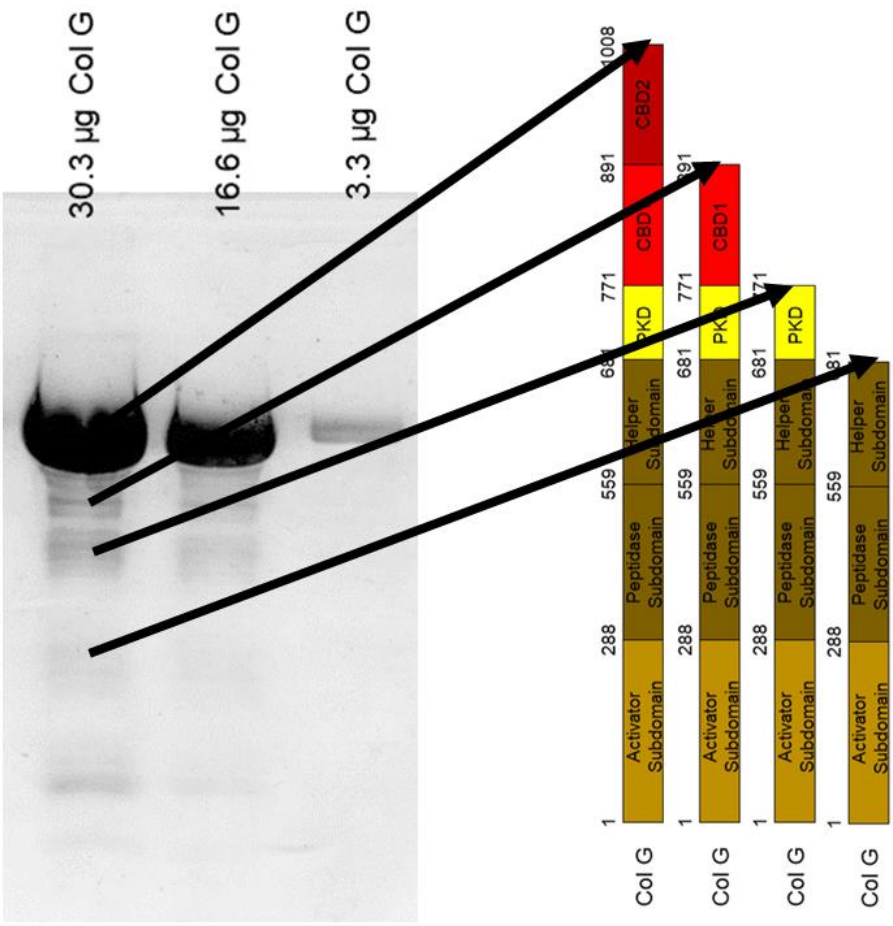


Figure 3: **A.** Greyscale image of Col G SDS PAGE gel with different amounts of protein. Not pictured, ImageJ densitometry results used to calculate purity. Rectangle was drawn around all bands for 30.3 µg well. Total intensity was calculated and intensity for each band was divided by total intensity to give purity. Purity was found to be 94%. Bands were excised for trypsin digest and mass spectrometry analysis. FASTA sequences for full-length Col G (uniprot: Q9X721) and was used with mass spectrometry data to find the residue coverage for Col G. Results show that small amounts of Col G lose C-terminal domains during expression.

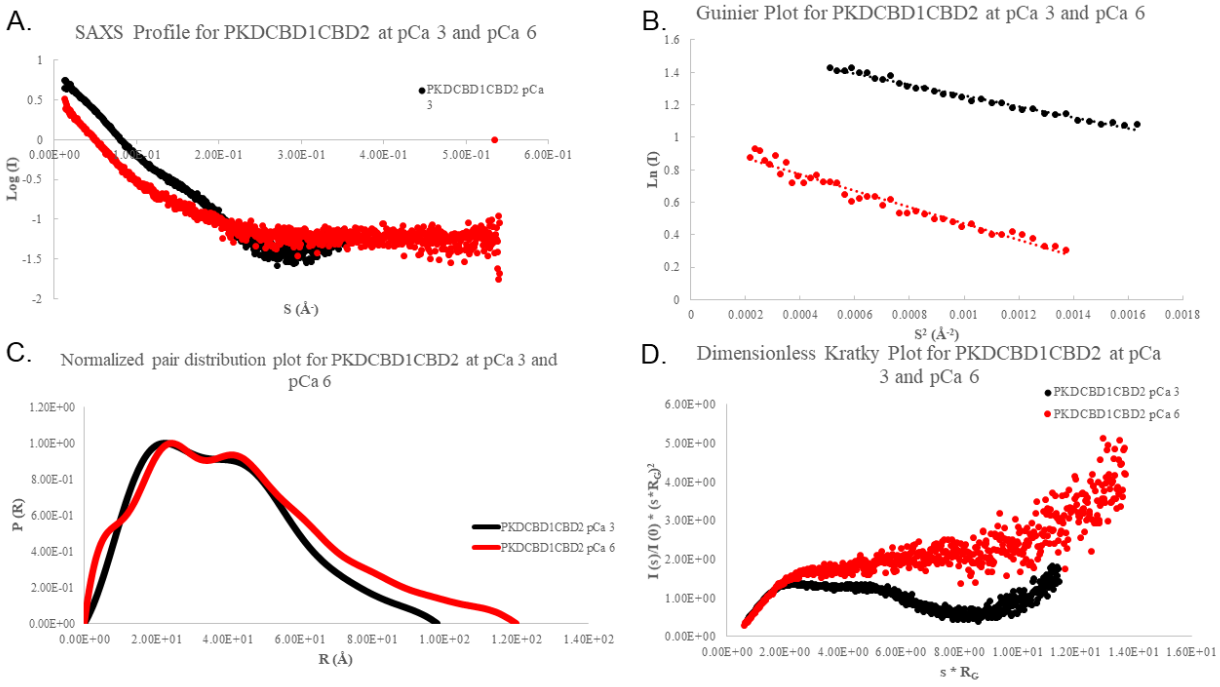


Figure 4: **A.** Experimental Scattering profile for PKDCBD1CBD2 at pCa 3 (black) and pCa 6 (red). **B.** Guinier plot for PKDCBD1CBD2 at pCa 3 and pCa 6. R_G values are calculated from the slope of Guinier region using Guinier approximation. **C.** Normalized pair distribution function for PKDCBD1CBD2 at pCa 3 and pCa 6. Extended D_{Max} in pCa 6 is due change in secondary structure at N-terminal linker for CBD1 and CBD2. **D.** Dimensionless Kratky plot for PKDCBD1CBD2 at pCa 3 and pCa 6. Plateau for pCa 6 is consistent with a protein that is unfolded [**13**]. PKDCBD1CBD2 pCa 6 envelope may need to be recalculated.

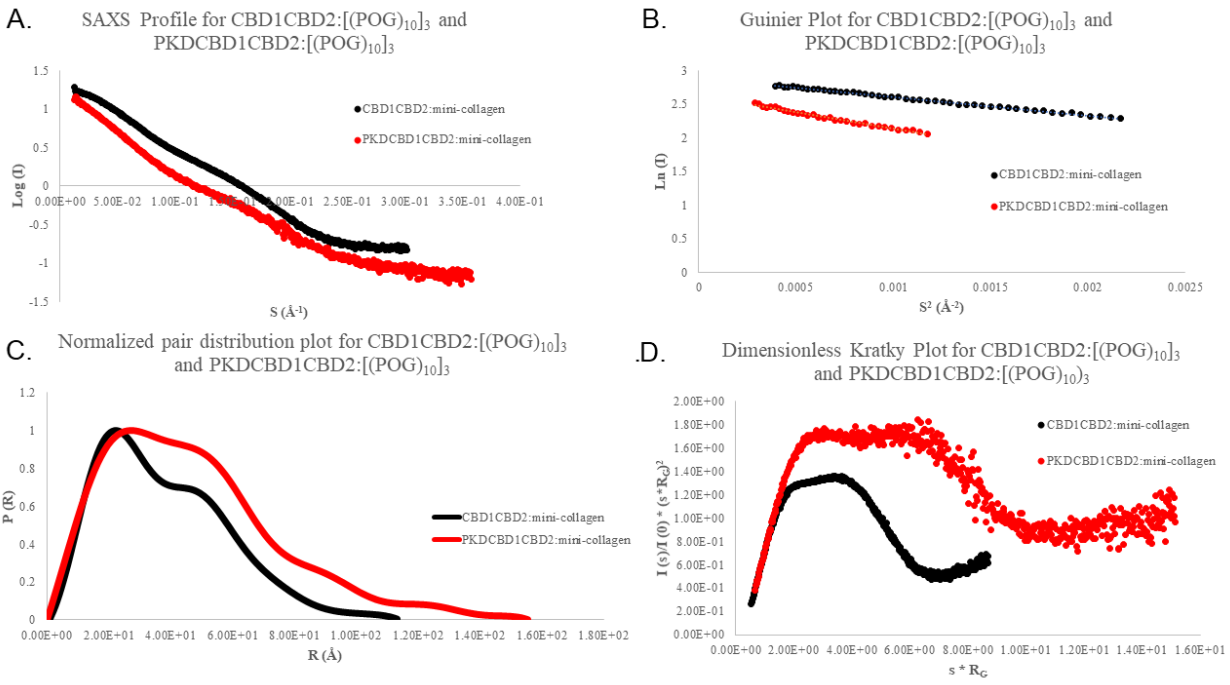


Figure 5: **A.** Experimental scattering profile for CBD1CBD2:$[(POG)_{10}]_3$ (black) and PKDCBD1CBD2:$[(POG)_{10}]_3$ (red). **B.** Guinier plot for CBD1CBD2:$[(POG)_{10}]_3$ (black) and PKDCBD1CBD2:$[(POG)_{10}]_3$ with R_G values calculated from Guinier approximation. **C.** Normalized pair distribution function for CBD1CBD2:$[(POG)_{10}]_3$ and PKDCBD1CBD2:$[(POG)_{10}]_3$. Extended D_{Max} in PKDCBD1CBD2:$[(POG)_{10}]_3$ is due to addition of PKD domain to complex (difference in D_{Max} is approximately equal to the length of PKD domain. **D.** Normalized Kratky plots for CBD1CBD2:$[(POG)_{10}]_3$ and PKDCBD1CBD2:$[(POG)_{10}]_3$ are indicative of folded proteins.

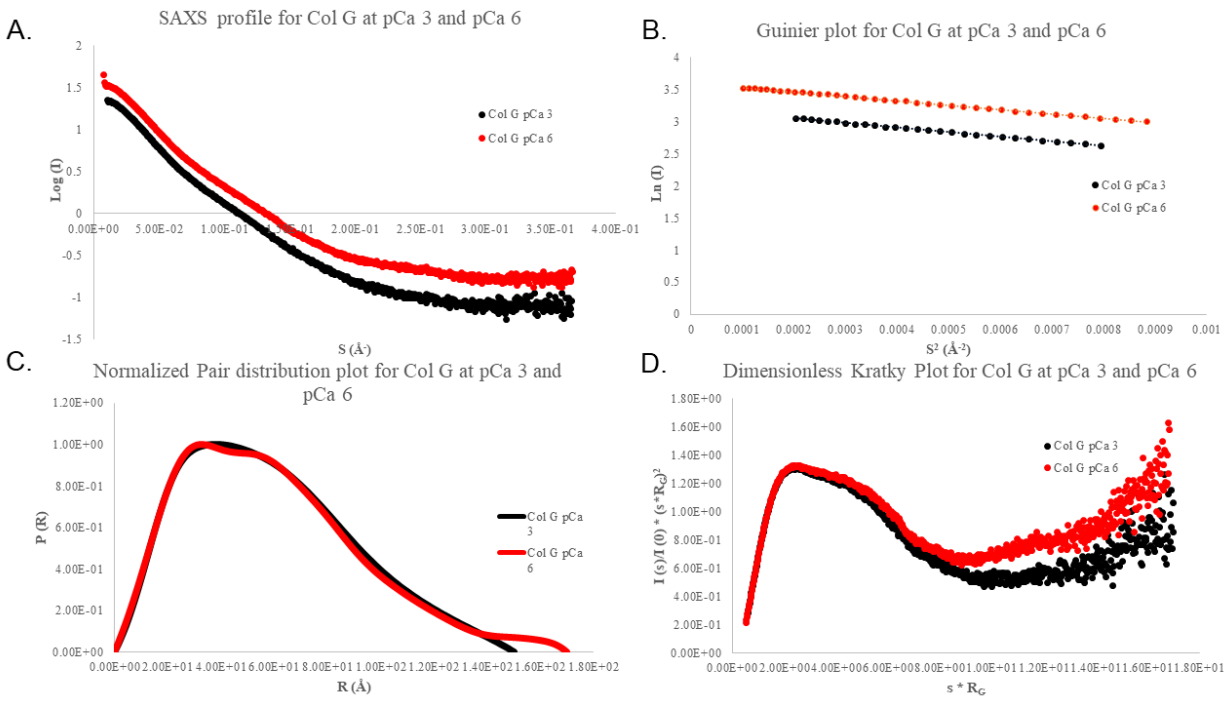


Figure 6: **A.** Experimental Scattering profile for full-length Col G at pCa 3 (black) and pCa 6 (red). **B.** Guinier plot for full-length Col G at pCa 3 and pCa 6 with R_G values calculated from Guinier approximation. **C.** Normalized pair distribution function for Col G at pCa 3 and pCa 6. Extended D_{Max} in pCa 6 is due change in secondary structure at N-terminal linker for CBD1 and CBD2. **D.** Dimensionless Kratky plot for PKDCBD1CBD2 at pCa 3 and pCa 6. After the initial "hump" y values for Col G at pCa 6 are higher than those for pCa 3. For a flexible molecule as x values approach infinity y values should start to plateau [**13**]. Movement towards plateau may indicate increase in flexibility for the enzyme at pCa 6.

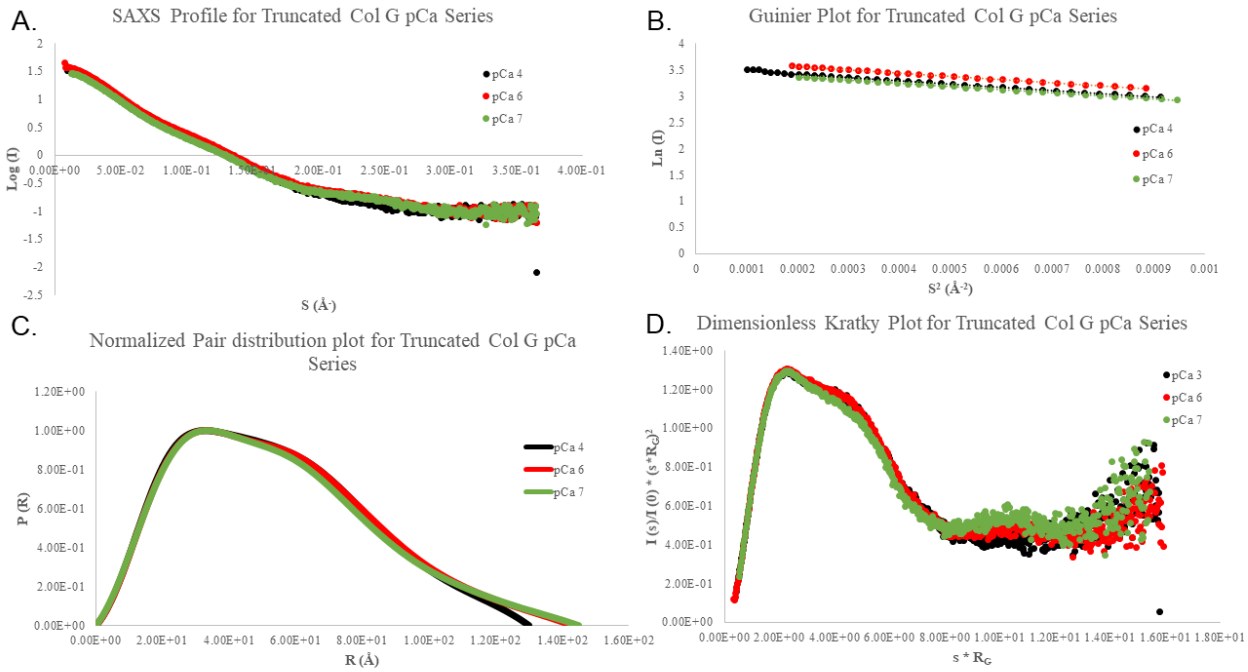
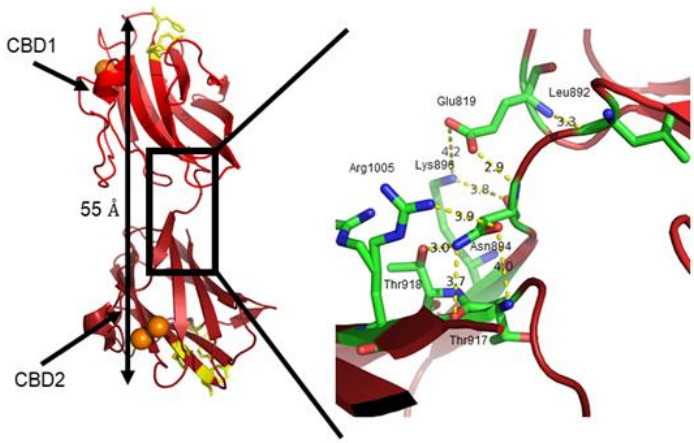


Figure 7: **A.** Experimental Scattering profile for Col -ΔCBD2 at pCa 4 (black), pCa 6 (red) and pCa 7 (green). **B.** Guinier plot for Col -ΔCBD2 at pCa 4, pCa 6 and pCa 7 with R_G values calculated from Guinier approximation. **C.** Normalized pair distribution function for Col -ΔCBD2 at pCa 4, pCa 6 and pCa 7. Extended D_{Max} at pCa 6 and pCa 7 is due change in secondary structure at N-terminal linker for CBD1 and CBD2. **D.** Dimensionless Kratky plot for Col -ΔCBD2 at pCa 4, pCa 6 and pCa 7. Plateau for pCa 6 is consistent with a protein that is unfolded [**13**]. Similar to full-length Col G the enzyme appears to become more flexible as pCa increases.



Linker Residue Bond Distances

Leu892 (-C=O)	Glu819 (-NH)	3.3 Å
Glu819 ($-CO_2$)	Asn894 (-NH)	2.9 Å
Glu819 ($-CO_2$)	Lys896 ($-NH_3^+$)	4.2 Å
Asn894 (-C=O)	Lys896 ($-NH_3^+$)	3.8 Å
Asn894 (-C=O)	Arg1005 ($-NH_3^+$)	3.9 Å
Asn894 (-C=O)	Thr917 (-NH)	4.0 Å
Thr917 (-C=O)	Asn894 ($-NH_2$)	3.7 Å
Thr918 (-OH)	Asn894 ($-NH_2$)	3.0 Å

Figure 8: Crystal structure of CBD1CBD2 (PDB: 5IKU) showing hydrogen bonds and salt bridges responsible for stabilizing the linker between CBD1 and CBD2. Distances are in Å.

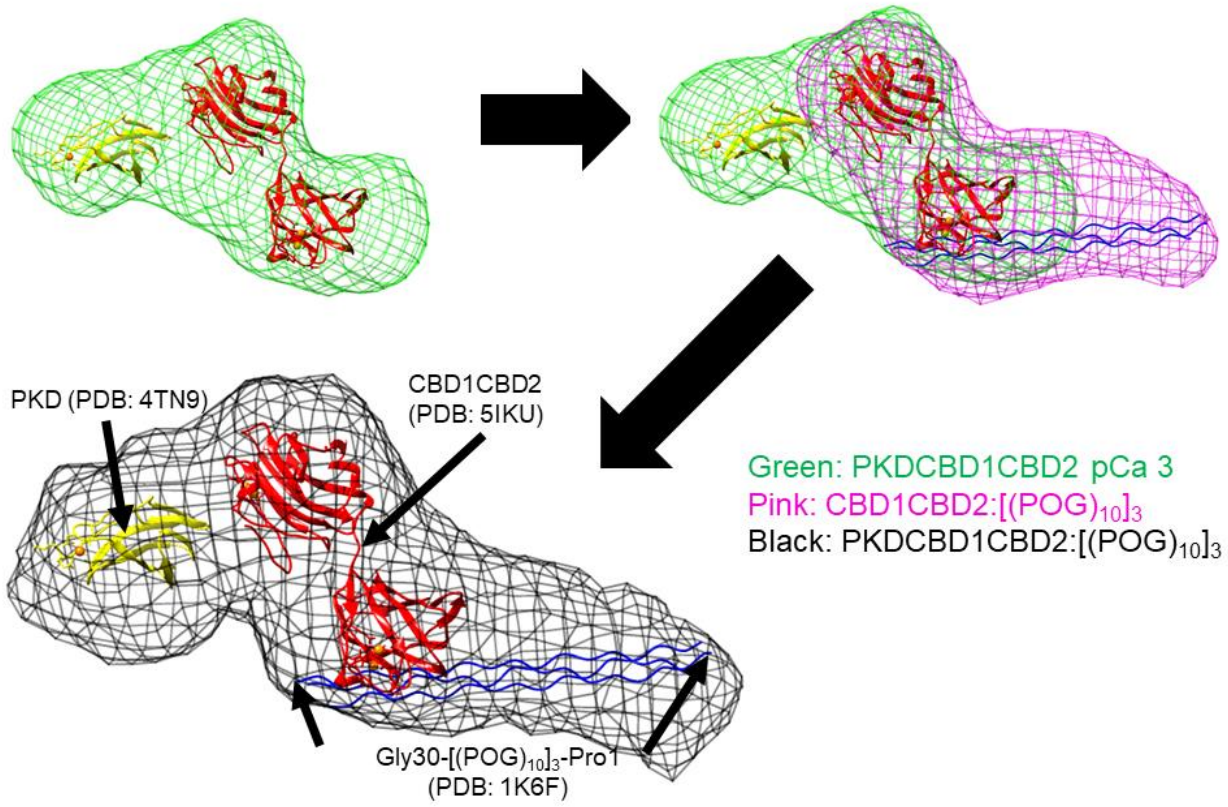


Figure 9: Puzzle-like approach to using multiple SAXS envelopes to derive the position of domains. For PKDCBD1CBD2 at pCa3 approximate position of domains was found by superimposing the envelope with envelope for CBD1CBD2:$[(POG)_{10}]_3$. Using PKDCBD1CBD2:$[(POG)_{10}]_3$ envelope the positions for CBD1, CBD2, PKD and $[(POG)_{10}]_3$ were confirmed. Based on binding data for CBD1CBD2, CBD2 was positioned on the C-terminus of mini-collagen [**3**]. CBD1 was placed in region above CBD2 and PKD2 was positioned in the remaining electron density.

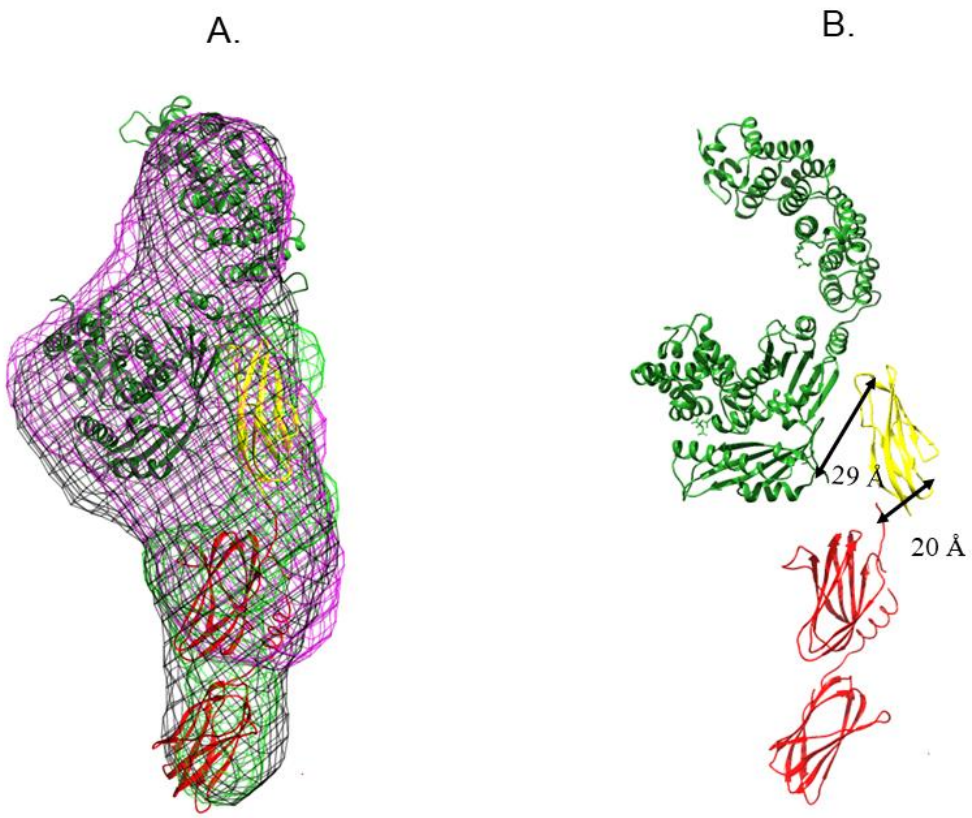


Figure 10: **A.** SAXS derived envelope for full-length Col G at pCa 6 (in black), Col G - ΔCBD2 at pCa 6 (in pink) and PKDCBD1CBD2 at pCa 6 (in green). PKDCBD1CBD2 envelope was first superimposed with full-length Col G in order to find where to dock CBD1CBD2 pCa 6 structure and PKD. Catalytic domain was then placed into remaining electron density. In order to check the position of the domains Col G -ΔCBD2 was superimposed with envelope for full-length structure. The non-overlapping region is where CBD2 was previously placed. **B.** To check if domains can reasonably link to form full-length enzyme the distance between terminal residues for each domain was measured using PyMol. For distance from CBD1 to PKD the 20 Å gap can easily be bridged when taking into account the 20 residues deleted from N-terminal linker of CBD1. 29 Å gap between PKD and catalytic domain cannot easily be made up as only 5 residues are missing between the two. The positions of the PKD and catalytic domain may need to be refined further.

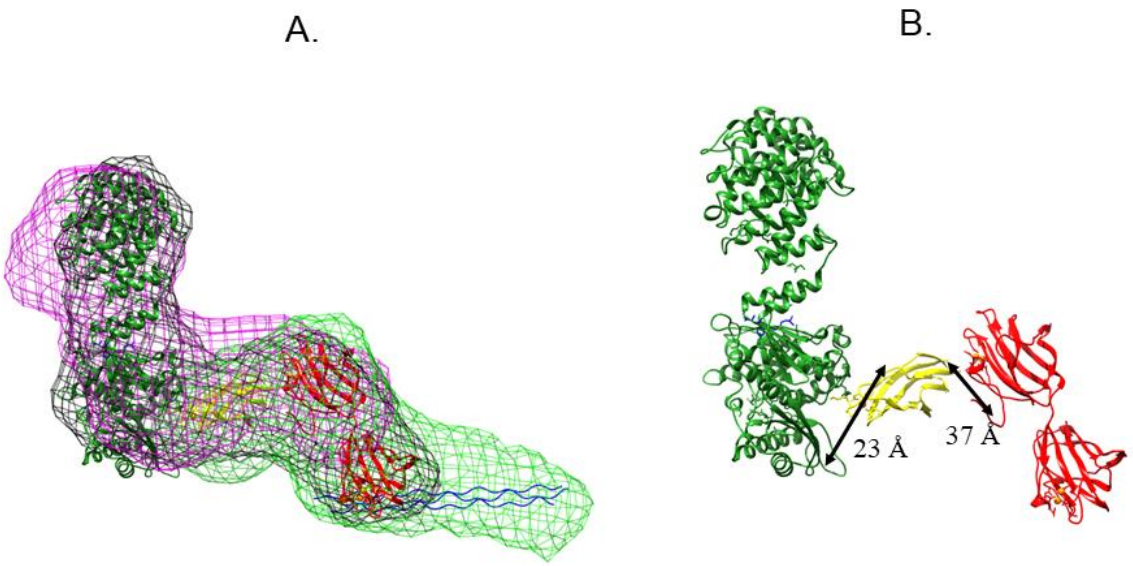


Figure 11: **A.** SAXS derived envelope for full-length Col G at pCa 3 (in black), Col G -ΔCBD2 at pCa 4 (in pink) and PKDCBD1CBD2:$[(POG)_{10}]_3$ at pCa 3 (in green). In order to determine the location of CBD2 in envelope for Col G at pCa 3 PKDCBD1CBD2:$[(POG)_{10}]_3$ envelope was superimposed with pCa 3 evelope. Since CBD2 in CBD1CBD2 is the domain that binds to mini-collagen CBD2 was positioned in the overlapping region between PKDCBD1CBD2:$[(POG)_{10}]_3$ and pCa 3 envelope [**3**] [**4**]. CBD1 and PKD were placed in the remaining overlapping electron density. Finally the catalytic domain was placed in the remaining electron density. As with pCa 6 Col G envelope Col G -ΔCBD2 was used to help check the position of CBD2. **B.** To check if domains can reasonably link to form full-length enzyme the distance between terminal residues for each domain was measured using PyMol. For distance from CBD1 to PKD the 37 Å gap cannot be easily bridged. The position of N-terminal linker in CBD1CBD2 may not be correct for PKDCBD1CBD2. 23 Å gap between PKD and catalytic domain cannot easily be made up as only 2 residues are missing between the two. The positions of the PKD and catalytic domain may need to be refined further.

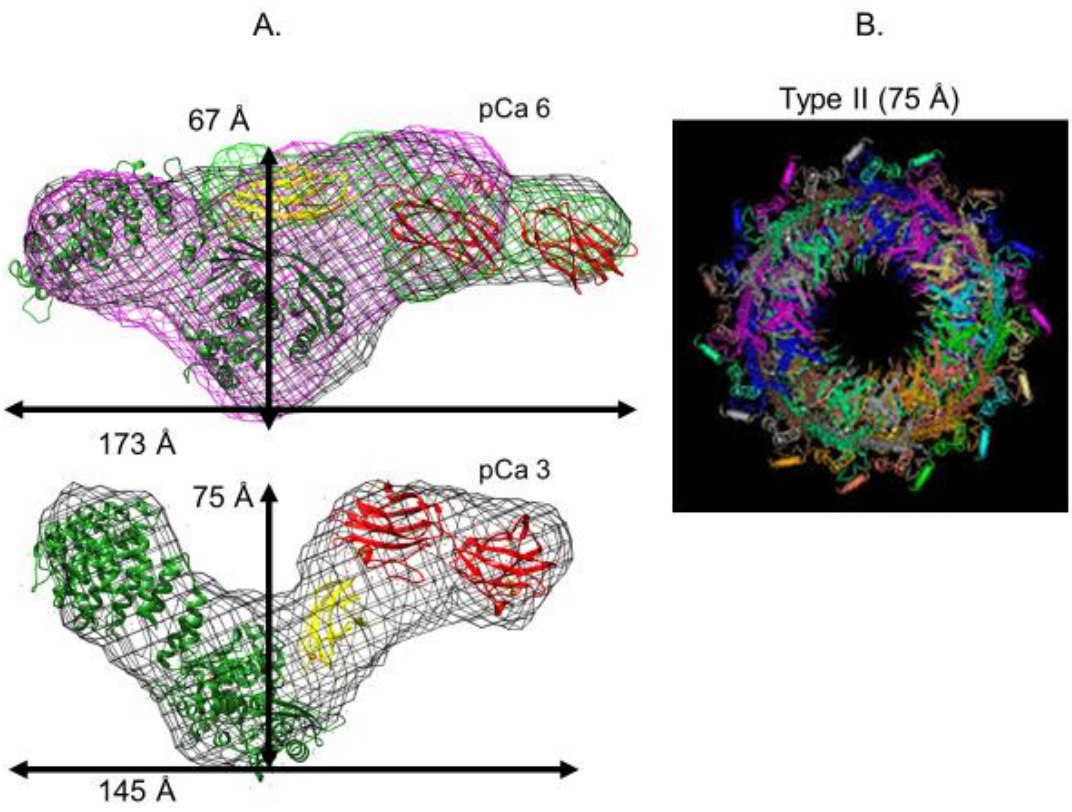


Figure 12: **A.** Comparing the widest portions of SAXS derived envelopes for Col G at pCa 3 and pCa 6. **B.** The interior diameter for the type II injectosome from *V. cholerae* was found to be 75 Å. At pCa 3 Col G would not be able to be secreted efficiently.

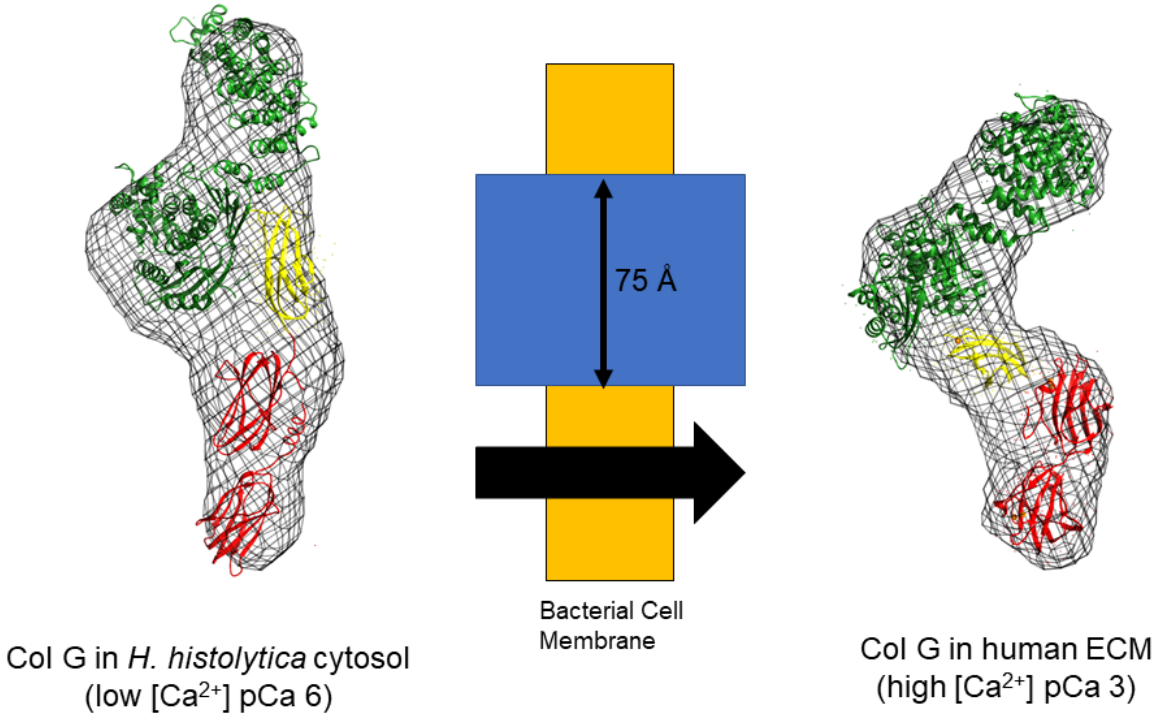


Figure 13: Secretion of Col G into human ECM may cause a change in the orientation of PKD domain which in turn may position catalytic domain for collagen degradation. Repositioning is most likely based on change in N-terminal linker for CBD1 and CBD2 from α-helix to β-sheet. Bacteria may use the difference in Ca^{2+} between cytosol and human ECM in order to quickly secrete collagenase. Once collagenase has entered host Ca^{2+} induced domain rearrangement quickly activates the enzyme.

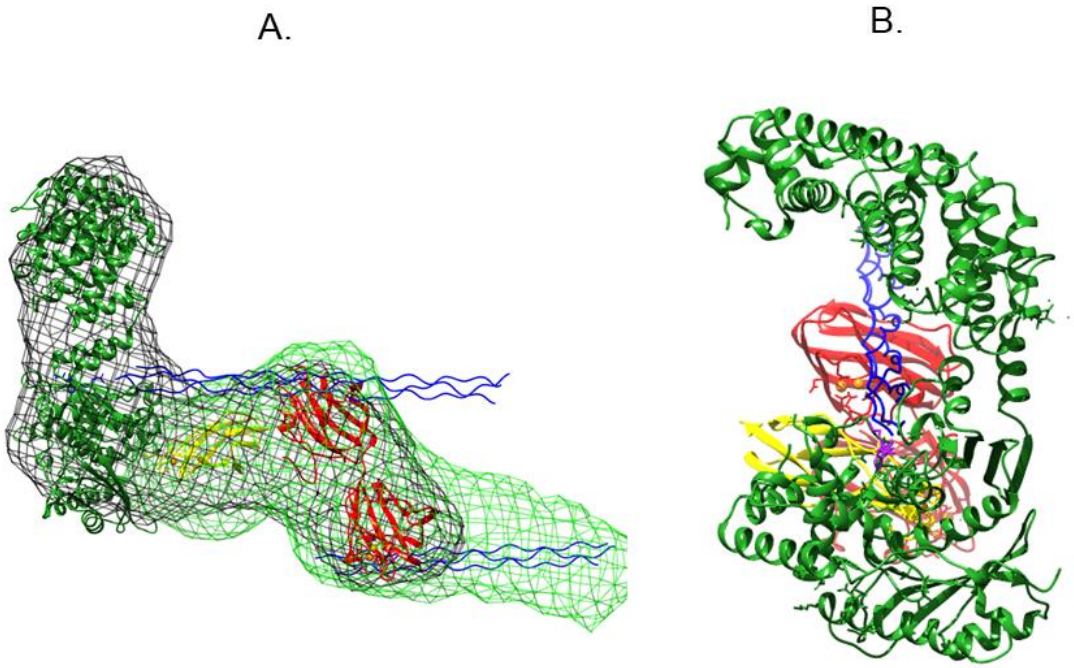


Figure 14: **A.** Superimposed structures for full-length Col G at pCa 3 (in black) and PKDCBDCBD:$[(POG)_{10}]_3$. Additional $[(POG)_{10}]_3$ was built into the collagen binding cleft of CBD1 using Chimera. **B.** Same structure but rotated 90°. Newly built mini-collagen aligns almost perfectly with the active site for catalytic domain. Catalytic domain crystal structure (PDB: 2Y6I) is in a complex with peptide like inhibitor (shown in purple).

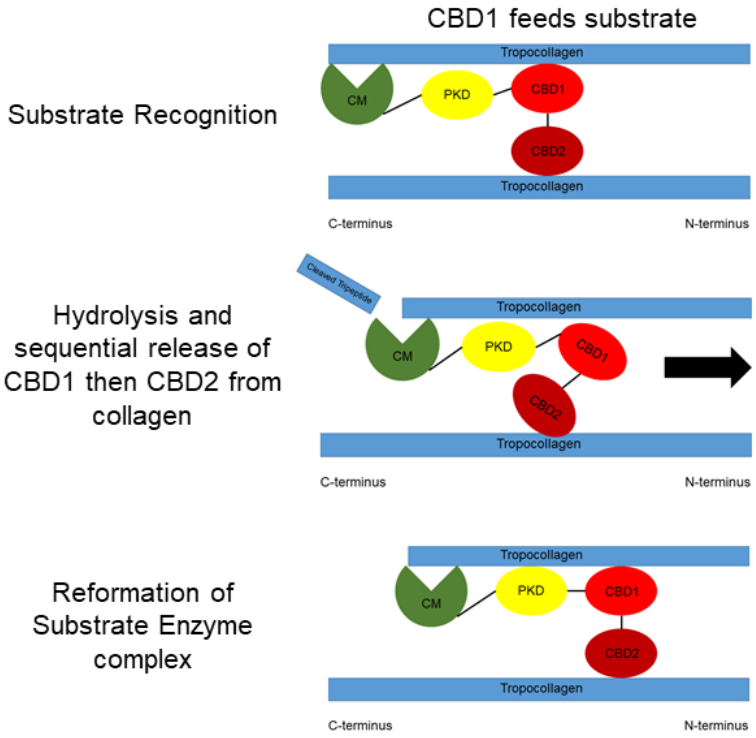


Figure 15: Schematic representation of processive collagenolysis from Col G collagenase. To initiate hydrolysis Col G may intercalate between tropocollagens in a damaged micro-fibril

Works Cited

1. R. Bauer, J. Wilson, S. T. L. Philominathan, D. Davis, O. Matsushita and J. Sakon, *Structural comparison of colh and colg collagen-binding domains from clostridium histolyticum*, Journal of Bacteriology **195** (2013), no. 2, 318-327.

2. T. Watanabe-Nakayama, M. Itami, N. Kodera, T. Ando and H. Konno, *High-speed atomic force microscopy reveals strongly polarized movement of clostridial collagenase along collagen fibrils*, Sci Rep **6** (2016), 28975.

3. S. T. L. Philominathan, T. Koide, O. Matsushita and J. Sakon, *Bacterial collagen-binding domain targets undertwisted regions of collagen*, Protein Science **21** (2012), no. 10, 1554-1565.

4. S. T. L. Philominathan, T. Koide, K. Hamada, H. Yasui, S. Seifert, O. Matsushita and J. Sakon, *Unidirectional binding of clostridial collagenase to triple helical substrates*, Journal of Biological Chemistry **284** (2009), no. 16, 10868-10876.

5. J. J. Wilson, O. Matsushita, A. Okabe and J. Sakon, *A bacterial collagen-binding domain with novel calcium-binding motif controls domain orientation*, EMBO (European Molecular Biology Organization) Journal **22** (2003), no. 8, 1743-1752.

6. K. Kar, S. Ibrar, V. Nanda, T. M. Getz, S. P. Kunapuli and B. Brodsky, *Aromatic interactions promote self-association of collagen triple-helical peptides to higher-order structures*, Biochemistry **48** (2009), no. 33, 7959-7968.

7. M. Nishio, *The ch/π hydrogen bond in chemistry. Conformation, supramolecules, optical resolution and interactions involving carbohydrates*, Physical Chemistry Chemical Physics **13** (2011), no. 31, 13873-13900.

8. P. Caviness, R. Bauer, K. Tanaka, K. Janowska, J. R. Roeser, D. Harter, J. Sanders, C. Ruth, O. Matsushita and J. Sakon, *Ca(2+) -induced orientation of tandem collagen binding domains from clostridial collagenase colg permits two opposing functions of collagen fibril formation and retardation*, Febs j **285** (2018), no. 17, 3254-3269.

9. H. Sekiguchi, K. Uchida, O. Matsushita, G. Inoue, N. Nishi, R. Masuda, N. Hamamoto, T. Koide, S. Shoji and M. Takaso, *Basic fibroblast growth factor fused with tandem collagen-binding domains from clostridium histolyticum collagenase colg increases bone formation*, Biomed Res Int **2018** (2018), 8393194.

10. O. Matsushita, C. M. Jung, J. Minami, S. Katayama, N. Nishi and A. Okabe, *A study of the collagen-binding domain of a 116-kda clostridium histolyticum collagenase*, The Journal of biological chemistry **273** (1998), no. 6, 3643-3648.

11. K. Yoshihara, O. Matsushita, J. Minami and A. Okabe, *Cloning and nucleotide-sequence analysis of the colh gene from clostridium-histolyticum encoding a collagenase and a gelatinase*, Journal of Bacteriology **176** (1994), no. 21, 6489-6496.

12. M. L. Green, A. G. Breite, C. A. Beechler, F. E. Dwulet and R. C. McCarthy, *Effectiveness of different molecular forms of c. Histolyticum class i collagenase to recover islets*, Islets **9** (2017), no. 6, 177-181.

13. R. C. McCarthy, A. G. Breite, M. L. Green and F. E. Dwulet, *Tissue dissociation enzymes for isolating human islets for transplantation: Factors to consider in setting enzyme acceptance criteria*, Transplantation **91** (2011), no. 2, 137-145.

14. A. G. Breite, R. C. McCarthy and F. E. Dwulet, *Characterization and functional assessment of clostridium histolyticum class i (c1) collagenases and the synergistic degradation of native collagen in enzyme mixtures containing class ii (c2) collagenase*, Transplantation Proceedings **43** (2011), no. 9, 3171-3175.

15. U. Eckhard, E. Schonauer and H. Brandstetter, *Structural basis for activity regulation and substrate preference of clostridial collagenases g, h, and t*, J Biol Chem **288** (2013), no. 28, 20184-20194.

16. S. T. Philominathan, O. Matsushita, R. Gensure and J. Sakon, *Ca2+-induced linker transformation leads to a compact and rigid collagen-binding domain of clostridium histolyticum collagenase*, Febs j **276** (2009), no. 13, 3589-3601.

17. N. Ohbayashi, T. Matsumoto, H. Shima, M. Goto, K. Watanabe, A. Yamano, Y. Katoh, K. Igarashi, Y. Yamagata and K. Murayama, *Solution structure of clostridial collagenase h and its calcium-dependent global conformation change*, Biophysical journal **104** (2013), no. 7, 1538-1545.

18. C. R. Sides, R. Liyanage, J. O. Lay, Jr., S. T. L. Philominathan, O. Matsushita and J. Sakon, *Probing the 3-d structure, dynamics, and stability of bacterial collagenase collagen binding domain (apo- versus holo-) by limited proteolysis maldi-tof ms*, Journal of the American Society for Mass Spectrometry **23** (2012), no. 3, 505-519.

19. R. Bauer, K. Janowska, K. Taylor, B. Jordan, S. Gann, T. Janowski, E. C. Latimer, O. Matsushita and J. Sakon, *Structures of three polycystic kidney disease-like domains from clostridium histolyticum collagenases colg and colh*, Acta Crystallographica Section D-Structural Biology **71** (2015), 565-577.

20. O. Matsushita, T. Koide, R. Kobayashi, K. Nagata and A. Okabe, *Substrate recognition by the collagen-binding domain of clostridium histolyticum class i collagenase*, Journal of Biological Chemistry **276** (2001), no. 12, 8761-8770.

21. R. Craig and R. C. Beavis, *A method for reducing the time required to match protein sequences with tandem mass spectra*, Rapid Commun Mass Spectrom **17** (2003), no. 20, 2310-2316.

22. H. Yasui, C. M. Yamazaki, H. Nose, C. Awada, T. Takao and T. Koide, *Potential of collagen-like triple helical peptides as drug carriers: Their in vivo distribution, metabolism, and excretion profiles in rodents*, Peptide Science **100** (2013), no. 6, 705-713.

23. N. J. Greenfield, *Using circular dichroism spectra to estimate protein secondary structure*, Nature Protocols **1** (2006), no. 6, 2876-2890.

24. I. Wittig, H. P. Braun and H. Schägger, *Blue native page*, Nat Protoc **1** (2006), no. 1, 418-428.

25. R. P. Rambo and J. A. Tainer, *Characterizing flexible and intrinsically unstructured biological macromolecules by sas using the porod-debye law*, Biopolymers **95** (2011), no. 8, 559-571.

26. C. D. Putnam, M. Hammel, G. L. Hura and J. A. Tainer, *X-ray solution scattering (saxs) combined with crystallography and computation: Defining accurate macromolecular structures, conformations and assemblies in solution*, Q Rev Biophys **40** (2007), no. 3, 191-285.

27. D. Molodenskiy, E. Shirshin, T. Tikhonova, A. Gruzinov, G. Peters and F. Spinozzi, *Thermally induced conformational changes and protein–protein interactions of bovine serum albumin in aqueous solution under different ph and ionic strengths as revealed by saxs measurements*, Physical Chemistry Chemical Physics **19** (2017), no. 26, 17143-17155.

28. D. Franke, M. V. Petoukhov, P. V. Konarev, A. Panjkovich, A. Tuukkanen, H. D. T. Mertens, A. G. Kikhney, N. R. Hajizadeh, J. M. Franklin, C. M. Jeffries and D. I. Svergun, *Atsas 2.8: A comprehensive data analysis suite for small-angle scattering from macromolecular solutions*, J Appl Crystallogr **50** (2017), no. Pt 4, 1212-1225.

29. Y. Zong, Y. Xu, X. Liang, D. R. Keene, A. Höök, S. Gurusiddappa, M. Höök and S. V. L. Narayana, *A 'collagen hug' model for staphylococcus aureus cna binding to collagen*, The EMBO Journal **24** (2005), no. 24, 4224-4236.

30. M. V. Petoukhov and D. I. Svergun, *Global rigid body modeling of macromolecular complexes against small-angle scattering data*, Biophysical Journal **89** (2005), no. 2, 1237-1250.

31. I. B. Holland, H. E. Jones, A. K. Campbell and A. Jacq, *An assessment of the role of intracellular free ca2+ in e. Coli*, Biochimie **81** (1999), no. 8-9, 901-907.

32. P. Maurer and E. Hohenester, *Structural and functional aspects of calcium binding in extracellular matrix proteins*, Matrix Biol **15** (1997), no. 8-9, 569-580; discussion 581.

33. C. Ruth, "Ca2+ - induced structural change of multi-domain collagen binding segments of collagenases colg and colh from hathewaya histolytica," University of Arkansas2018, p. 38.

34. Z. Yan, M. Yin, D. Xu, Y. Zhu and X. Li, *Structural insights into the secretin translocation channel in the type ii secretion system*, Nat Struct Mol Biol **24** (2017), no. 2, 177-183.

35. T. G. Sana, R. Lomas, M. R. Gimenez, A. Laubier, C. Soscia, C. Chauvet, A. Conesa, R. Voulhoux, B. Ize and S. Bleves, *Differential modulation of quorum sensing signaling through qsla in pseudomonas aeruginosa strains pao1 and pa14*, J Bacteriol **201** (2019), no. 21.

36. T. Kobayasi and T. Karlsmark, *Type v and vi collagen for cohesion of dermal fibrillar structures*, J Submicrosc Cytol Pathol **38** (2006), no. 2-3, 103-108.

37. H. Shima, A. Inagaki, T. Imura, Y. Yamagata, K. Watanabe, K. Igarashi, M. Goto and K. Murayama, *Collagen v is a potential substrate for clostridial collagenase g in pancreatic islet isolation*, Journal of diabetes research **2016** (2016), 4396756-4396756.

38. K. Uchida, O. Matsushita, N. Nishi, G. Inoue, K. Horikawa and M. Takaso, *Enhancement of periosteal bone formation by basic fibroblast-derived growth factor containing polycystic kidney disease and collagen-binding domains from clostridium histolyticum collagenase*, J Tissue Eng Regen Med **11** (2017), no. 4, 1165-1172.

39. G. Inoue, K. Uchida, O. Matsushita, H. Fujimaki, W. Saito, M. Miyagi, H. Sekiguchi, N. Nishi, S. Ohtori, M. Yogoro and M. Takaso, *Effect of freeze-dried allograft bone with human basic fibroblast growth factor containing a collagen-binding domain from clostridium histolyticum collagenase on bone formation after lumbar posterolateral fusion surgery in rats*, Spine (Phila Pa 1976) (2017).

40. W. Saito, K. Uchida, O. Matsushita, G. Inoue, H. Sekiguchi, J. Aikawa, H. Fujimaki and M. Takaso, *Acceleration of callus formation during fracture healing using basic fibroblast growth factor-kidney disease domain-collagen-binding domain fusion protein combined with allogenic demineralized bone powder*, Journal of Orthopaedic Surgery and Research **10** (2015), 6.

41. H. Sekiguchi, K. Uchida, G. Inoue, O. Matsushita, W. Saito, J. Aikawa, K. Tanaka, H. Fujimaki, M. Miyagi and M. Takaso, *Acceleration of bone formation during fracture healing by poly(pro-hyp-gly)10 and basic fibroblast growth factor containing polycystic kidney disease and collagen-binding domains from clostridium histolyticum collagenase*, J Biomed Mater Res A **104** (2016), no. 6, 1372-1378.

42. K. Kuwaba, M. Kobayashi, Y. Nomura, S. Irie and Y. Koyama, *Elongated dermatan sulphate in post-inflammatory healing skin distributes among collagen fibrils separated by enlarged interfibrillar gaps*, The Biochemical journal **358** (2001), no. Pt 1, 157-163.

43. J. P. R. O. Orgel, T. C. Irving, A. Miller and T. J. Wess, *Microfibrillar structure of type i collagen in situ*, Proceedings of the National Academy of Sciences of the United States of America **103** (2006), no. 24, 9001-9005.

44. L. Ye and W. G. Jiang, *Bone morphogenetic proteins in tumour associated angiogenesis and implication in cancer therapies*, Cancer Lett **380** (2016), no. 2, 586-597.

45. X. Chen, *Fusion protein linkers: Property, design and functionality*, **65** (2013), no. 10, 1357-1369.

46. Y.-R. Yun, J. E. Won, E. Jeon, S. Lee, W. Kang, H. Jo, J.-H. Jang, U. S. Shin and H.-W. Kim, *Fibroblast growth factors: Biology, function, and application for tissue regeneration*, Journal of tissue engineering **2010** (2010), 218142-218142.

47. S. Nakamura, M. Nambu, T. Ishizuka, H. Hattori, Y. Kanatani, B. Takase, S. Kishimoto, Y. Amano, H. Aoki, T. Kiyosawa, M. Ishihara and T. Maehara, *Effect of controlled release of fibroblast growth factor-2 from chitosan/fucoidan micro complex-hydrogel on in vitro and in vivo vascularization*, J Biomed Mater Res A **85** (2008), no. 3, 619-627.

48. K. Omata, T. Matsuno, K. Asano, Y. Hashimoto, Y. Tabata and T. Satoh, *Enhanced bone regeneration by gelatin-beta-tricalcium phosphate composites enabling controlled release of bfgf*, J Tissue Eng Regen Med **8** (2014), no. 8, 604-611.

49. M. R. Rickels and R. P. Robertson, *Pancreatic islet transplantation in humans: Recent progress and future directions*, Endocrine reviews **40** (2019), no. 2, 631-668.

50. I. Fujita, R. Utoh, M. Yamamoto, T. Okano and M. Yamato, *The liver surface as a favorable site for islet cell sheet transplantation in type 1 diabetes model mice*, Regenerative Therapy **8** (2018), 65-72.

51. O. Matsushita, T. Koide, R. Kobayashi, K. Nagata and A. Okabe, *Substrate recognition by the collagen-binding domain of clostridium histolyticum class i collagenase*, J Biol Chem **276** (2001), no. 12, 8761-8770.

52. A. N. Balamurugan, A. G. Breite, T. Anazawa, G. Loganathan, J. J. Wilhelm, K. K. Papas, F. E. Dwulet, R. C. McCarthy and B. J. Hering, *Successful human islet isolation and transplantation indicating the importance of class 1 collagenase and collagen degradation activity assay*, Transplantation **89** (2010), no. 8, 954-961.

53. M. R. Wilkins, E. Gasteiger, A. Bairoch, J. C. Sanchez, K. L. Williams, R. D. Appel and D. F. Hochstrasser, *Protein identification and analysis tools in the expasy server*, Methods Mol Biol **112** (1999), 531-552.

Chapter 3: Identifying the role of PKD domains 1 and 2 in Col H collagenolysis

Abstract

Col H and Col G differ in the number of Collagen binding domains (CBDs), Polycystic kidney disease (PKD) like domains. While structurally the CBDs for Col H and Col G are similar the PKD domains for Col H (PKD1 and PKD2) display different surface features than the PKD domain for Col G (PKD). Most importantly are the presence of surface aromatic residues (PKD1: Tyr689, Phe712; PKD2: Tyr780, His782, Tyr796, Tyr801). Difference in domain orientation and surface features for collagenase PKD domains further hint that Col G and Col H target different regions in collagen. In this chapter we seek to understand the significance of these residues and the role they play in Col H collagenolysis through mutagenesis and Small Angle X-ray Scattering (SAXS)

3.1 Introduction

Hathewaya (*Clostridium*) *histolytica* is a gram-positive bacteria that is known to cause gas gangrene. The β-exotoxin of this bacteria consist of two enzymes, Col G and Col H, which seek out and degrade collagen fibers located in extracellular space. Destruction of collagen exposes new nutrient sources for the invading bacteria and helps spread the infection [**2**]. Both Col G and Col H have a similar multi-domain structure consisting of collagen binding domains (CBDs; CDB1 and CBD2 for Col G and CBD for Col H), polycystic kidney disease like-domains (PKD for Col G and PKD1 and PKD2 for Col H) and a catalytic domain, a zinc peptidase from the M9B family that is responsible for collagen hydrolysis (see figure 1) [**3**].

Currently, the specific role of PKD domains in collagen degradation is unknown. Col G contains only a single PKD domain while Col H contains two PKD domains, PKD1 and PKD2.

Polycystic kidney disease like domains (PKD) adopt a variable immunoglobulin fold (V-set Ig) that was first found in polycystic kidney disease protein. Col G and Col H PKD domain crystal structures resemble the variable antibody domain and consist of a β-sandwich of seven strands stabilized through a hydrogen bonding network (see figure 2A) [**4**]. Structurally PKD domains are thought to assist in protein-protein or protein-carbohydrate interactions [**5**]. The difference in the number of PKD domains as well as the different structural features for Col G and Col H PKD domains hints that the two play different roles in degrading collagen fibrils.

Similar to CBDs PKD domains for Col G and Col H all have conserved Asp and Asn residues responsible for binding calcium ions (see figure 2A). In the presence of calcium the mid-section of S2 becomes more flexible but for S2A and S2B calcium causes the domains to become more rigid [**5**]. Unlike CBD domains, PKD domains do not undergo significant structural changes in the presence of Ca^{2+} [**5**]. Col H PKD domains contain surface aromatic residue that are found along a prominent ridge that are not present in Col G PKD domain, suggesting that at least for Col H the PKD domains help to bind the enzyme to collagen (PKD1: Tyr689, Phe712 PKD2: Tyr780, His782, Tyr796, Tyr801) (see figure 2B) [**5**]. The different surface features of Col H and Col G PKD domains show that the two different types of PKD domains from Col G and Col H have different functions in collagen degradations.

The specific role PKD domains play for *H. histolytica* collagenases is not known. PKD domain may enhancing binding capabilities, act as a spacer or swell collagen [**3**] [**6**] [**7**] [**8**] [**9**]. PKD from Col G does not contain surface aromatic residues which suggests that it does not play a role in collagen binding. Binding studies for Col H PKD2CBD with insoluble type I collagen indicate that binding occurs in two phases [**3**]. The first dissociation constant is 2.11 μM and the second dissociation constant is 339 nM. The first phase of binding indicates that just the presence

of PKD2 causes CBD domain to bind to insoluble collagen tighter. While PKD2CBD has two phases of binding PKD2 by itself has negligible binding affinity for insoluble collagen [**3**]. Surface plasmon resonance derived dissociation constants for PKD2CBD and CBD are roughly identical [**10**]. PKD2's binding ability is only present for higher ordered collagen substrates. Currently it is unknown if Col H PKD1 plays a role in collagen degradation. Full length non-active Col H was found to have the highest affinity for insoluble collagen fibril which suggests that the catalytic domain helps contribute towards collagen binding [**3**]. A collagen binding mechanism for full-length Col H has yet to be postulated. Currently it is unknown what role PKD1 and PKD2 play in Col H collagenolysis. Binding and structural study results for Col H are described below and the role of PKD1 and PKD2 in Col H collagenolysis is discussed.

Col G PKD domain has no detectable binding to insoluble collagen so its role may be purely structural [**11**]. B-factor analysis of the crystal structure of Col G PKD domain indicates that the protein has high flexibility so perhaps the role of PKD is to help position the catalytic domain in Col G so that it aligns with the collagen binding clefts of CBD1 or CBD2. PKD1 in Col H may play the same role. In addition to assisting in collagen binding and processivity, PKD domains may also assist in collagen "swelling". The ability to widen the distance from one tropocollagen molecule to another in a fibril or micro-fibril is known as collagen swelling. Some serine proteases have been determined to swell collagen fibers using their PKD domains to assist in degradation [**8**]. Type I collagens bear a net positive charge and interactions with serine protease PKD domain is expected to increase this charge. This may cause an increase in electrostatic repulsive forces between tropocollagens within a fibril and may result in exposure of tropocollagens. Similar to the surface exposed residues in the Col H PKD domains surface exposed aromatic residues from PPC domains are necessary for binding and swelling collagen.

When these aromatic residues were mutated to alanine swelling was significantly reduced [**9**]. For PPC domains binding is necessary for swelling to occur and this may be the same for Col H PKD domains. However, it is currently unknown if clostridium collagenases are capable of swelling collagen.

3.2 Methods

3.2a. Site Directed Mutagenesis of Col H PKD domains

Surface exposed aromatics (TYR, HIS and PHE) along a proposed PKD collagen binding ridge in PKD1 and PKD2 were mutated to Ser using PCR site directed mutagenesis [**5**]. Wild type genes for PKD1PKD2CBD and PKD2CBD were used as gene templates for PCR. 5' and 3' pGEX primers carrying the desired mutations were purchased from Eurofins Genomics and used to insert mutations into wild type genes similar to experiment from Wilson et al. 2003 [**12**]. Primers are listed in Table 1 and mutated nucleotides are underlined.

After 30 PCR cycles, the presence of newly synthesized gene product was confirmed using gel electrophoresis. PCR products were transformed into *E. Coli* Dh5α cells by heat shock for plasmid replication. The transformation product was then plated onto LB media containing ampicillin. After 24 hours a single colony was inoculated into 5 ml of LB-Amp and incubated at 37° C overnight with vigorous shaking. Plasmids were isolated from cells using ethanol precipitation [**13**]. After transformation, the presence of the desired mutation in the final plasmid product was confirmed using Sanger sequencing [**14**]. DNA concentration was found using NanoDrop ND-1000 by measuring the ratio of absorbance at 260 nm to absorbance at 280 nm. Finally, each mutant plasmid was then transformed into BL21 competent cells for expression using previously described methods [**15**] [**3**]. The following expression vectors were constructed: PKD2CBD Y780S, PKD2CBD H782S, PKD2CBD Y796S, PKD2CBD Tyr801Ser, PKD2CBD

Tyr780Ser His782Ser Tyr796Ser Tyr801Ser (4Ser PKD2CBD), PKD1PKD2CBD Tyr689Ser Phe712Ser (2Ser PKD1PKD2CBD) and PKD1PKD2CBD Tyr689Ser Phe712Ser Tyr780Ser His782Ser Tyr796Ser Tyr801Ser (6Ser PKD1PKD2CBD). Expression vectors containing multiple mutations (4Ser PKD2CBD, 2Ser PKD1PKD2CBD and 6Ser PKD1PKD2CBD) were created by performing site directed mutagenesis experiments with primers containing single mutations until desired plasmid is formed. For 6Ser PKD1PKD2CBD, PCR annealing temperature was lowered to 50°C so that reverse primer for Tyr780Ser His782Ser could anneal to template.

Table 1

Primer	Sequence (5’ to 3’)
PKD2–Y780S-H782S–For	ATATCTATGAGTGTACCTAAATCCGGAGC
PKD2–Y780S-H782S–Rev	TACACTCATAGATATAACTGGAAGTTTAT
PKD2–Y796S-Y801S–For	CTGGAAAAGGAACATCTGACCCAGATGGATCTAT
PKD2–Y796S-Y801S–Rev	GATGTTCCTTTTCCAGAGAAAACAACTTTTTGAT
PKD2–Y780S–For	GTTATATCGATGCATGTACCTAAATCC
PKD2–Y780S–Rev	ATGCATCGATATAACTGGAAGTTTATT
PKD2–H782S–For	TATATGAGTGTACCTAAATCCGGAGCC
PKD2–H782S–Rev	AGGTACACTCATATATATAACTGGAAG
PKD2–Y796S–For	GTTTTCTCGGGAAAAGGAACATATGAC
PKD2–Y796S–Rev	TTTTCCCGAGAAAACAACTTTTTGATT
PKD2–Y801S–For	GGAACATCGGACCCAGATGGATCTATC
PKD2–Y801S–Rev	TGGGTCCGATGTTCCTTTTCCATAGAA

Table 1 Cont.

Primers	Sequence (5’ to 3’)
PKD1–Y689S–For	TTACCTTCGGGCAAGATCAATGGAACT
PKD1–Y689S–Rev	CTTGCCCGAAGGTAATGAATTTTTGGA
PKD1–F712S–For	GGCTCTTCGGATCCAGATGGTAAAATA
PKD1–F712S–Rev	TGGATCCGAAGAGCCTTCACTAGAGAA

3.2b. Production and Purification of Col G and Col H segments

The expression and purification of wild type/mutant Col H PKD1PKD2CBD and PKD2CBD was performed at Okayama University Graduate School of Medicine Dentistry & Pharmaceutical Sciences in the Matsushita lab using previously described methods [**3**] [**15**]. Briefly, pGEX-4T-2 plasmids containing the sequences to be expressed were transformed into *E. coli* BL21 competent cells by heat shock. Cells were plated onto LB Agar plates supplemented with ampicillin and chloramphenicol and allowed to grow at 37° C. After 24 hours 10 ml of 2-YTG broth supplemented with ampicillin was inoculated with a single colony; broth was incubated at 37° C for 12 hours with agitation. 6 ml of new culture was transferred to 120 ml of 2-YTG broth containing ampicillin and the culture was incubated once again at 37° C for 12 hours with agitation. Finally, after 12 hours 25 ml of this culture was transferred to into 2 liters of 2-YTG containing ampicillin. Cells were incubated until OD_{600} of 0.7. IPTG was added to the 2-YTG broth and expression was carried out for 4-8 hours. After addition of 20 mL of 100 mM PMSF, cells were harvested using centrifugation (7,000 RPM, 4° C, 15 min). Cell pellet was suspended in 40 ml of 1X PBS supplemented with 1 mM PMSF and French press (10,000 PSI)

was used to lyse the cells into centrifugation tubes (20 ml). 1 ml of Triton X-100 was added to each tube and tubes were rotated for 30 min. (4° C). Centrifugation was performed twice to clear out cell debris (15,000 RPM, 4° C, 30 min).

The fusion proteins were purified using batch purification with glutathione sepharose beads. Junk proteins were eluted with the PBS wash. Fusion proteins were eluted with 50 mM Tris pH 8.0 containing 10 mM glutathione. Incubation with thrombin cleaves the GST tag and a second-round chromatography with the glutathione sepharose column was used to separate the protein from the tag. Q-sepharose ion exchange chromatography was used to further purify the proteins. Flow through fractions were pooled and dialyzed into 50 mM Tris, 1 mM $CaCl_2$ pH 7.5. Following dialysis, a BCA assay was used to quantify concentration for each sample.

3.2c Circular Dichroism and Tryptophan Emission Spectra of Col H segments

Circular Dichroism and Tryptophan fluorescence emission were used to check if introduction of Ser mutations along the proposed collagen binding ridge in PKD1 and PKD2 would alter the structure of PKD1PKD2CBD and PKD2CBD. For both Circular Dichroism and Fluorescence experiments proteins were dialyzed into 10 mM Sodium phosphate pH 7.5 1 mM $CaCl_2$. Circular Dichroism and Tryptophan emission spectra of PKD1PKD2CBD and PKD2CBD proteins were found using a Jasco CD spectrometer at room temperature.

3.2d. Collagen Fibril Binding Study for Col H segments

Binding information for Col H segments was found using a previously described method slightly modified [**3**]. Binding analysis was performed using type I insoluble collagen from a Bovine Achilles tendon (Sigma). Briefly, 5 mg of insoluble type I collagen will be added to a .22 µm micro centrifugal membrane. Collagen was swollen in 200 µl of 50 mM Tris-HCl 150 mM

NaCl 5 mM $CaCl_2$ pH 7.5. Collagen-buffer mixture was then incubated for 30 min. at room temperature. After incubation, the mixture was centrifuged at 15,000 x G for 30 min. The collagen is then resuspended in 100 µl of the wild type or mutant proteins at various concentrations (.3 mg/ml - .04 mg/ml). 100 µl of Bovine Serum Albumin at an equal concentration to the protein was added. The mixture was incubated for 30 min and then filtered with a fresh .22 µm micro centrifugal membrane at 15,000 x G for 15 min. After centrifugation the filtrate was collected for densitometry analysis

3.2e Densitometry analysis of Col H binding segments

Filtrate for each concentration from a sample was ran on an 15% SDS PAGE gel for 2 hrs. at 20 mA. Gels were stained with comassie blue and destained with 10% Acetic acid/50% Methanol for 24 hrs. After destaining, Gels were scanned using Epson scanner and were analyzed using ImageJ. Greyscale images were used and background was subtracted (light background, sliding parabola and disable smoothing). Rectangles were drawn for each lane to include both BSA and sample bands. ImageJ densitometry analysis was used to determine the amount of protein in each filtrate by constructing a calibration curve with the BSA bands [**16**]. To potentially compensate for improper staining or destaining scaling factors were derived based on the multiplication values needed for BSA intensity values to equal theoretical values from calibration curve. Derived scaling factors were used for samples whose BSA intensities diverged greatly from the calibration curve.

The amount of sample bound and the amount of sample in the filtrate was calculated for each well by using calibration curve from BSA band intensity. Saturation curves were constructed to give the number of binding phases for each sample. For saturation curves, finding the binding parameters such as dissociation constants (K_D) and number of binding sites (B_{Max})

for proteins with two phases of binding is difficult as the concentration needed to find these parameters may not be achievable. To find binding parameters Scatchard plots were constructed by plotting concentration of bound sample over sample vs. bound sample. Slope and x-intercept were noted to find association constant, K and amount of protein bound to collagen fibril, B_{Max}. Individual results were obtained using linear regression and averaged in triplicate.

3.2f. Identifying lower molecular weight contaminants in full length Col H from VitaCyte

Full length Col H was purchased from VitaCyte for structural analysis. 10% SDS PAGE gel was run for Col H to determine if contaminants were present in the samples. Gels were stained with comassie blue and destained with 10% Acetic acid/50% Methanol for 24 hrs. Analysis of SDS PAGE gel showed the presence of contaminants for Col H. ImageJ densitometry analysis of gel revealed that Col H is sufficiently pure (>93% purity for both Col H) (see figure 3).

Bands were excised, washed twice with 25 mM ammonium bicarbonate, destained with ammonium bicarbonate/50% acetonitrile, and dried with pure acetonitrile. Then, the proteins were reduced using 10 mM dithiothreitol in 50 mM ammonium bicarbonate for 1hr. at 55°C water bath. Proteins were alkylated by adding 10 mM iodoacetamide (in 50 mM ammonium bicarbonate). Reaction was carried out for 1 hr. in the dark. Next, gel pieces were washed with ammonium bicarbonate, and dried with pure acetonitrile. Enough trypsin gold (Promega, Mass spectrometry grade, 10 ng/µl in 25 mM ammonium bicarbonate) was added to cover the gel pieces and the mixture was incubated at 4°C for 30 min. After digestion, 100 µl of 25mM ammonium bicarbonate was added and the mixture was left overnight at room temperature. Solutions were transferred to separate Eppendorf tubes; left over tryptic peptides being extracted by vortexing with 100 µl of 50% acetonitrile/5% formic acid for 5 min. Solutions were

completely evaporated and reconstituted in 50 µl of 0.1% formic acid and mass spectrometry analysis was performed.

3.2g. Mass Spectrometry analysis of Col G and Col H

Dr. Rohanna Liyanage in the mass spectrometry department at the University of Arkansas performed all trypsin digestions and mass spectrometry analysis. The tryptic digests obtained from all samples were analyzed by ESI-LC-MS/MS (Electrospray ionization liquid chromatography-tandem mass spectrometry) at the State-Wide Mass Spectrometry Facility at the University of Arkansas (Fayetteville, AR). Analysis was performed using data-dependent acquisition mode (DDA) with collision-induced dissociation (UltraScan mode, one min, 32,500 m/z per second). Agilent 1200 series micro flow HPLC system in line with Bruker Amazon-SL quadrupole ion trop ESI mass spectrometer (QIT-ESI-MS) was used for analyzing tryptic digests. Reverse-phase high performance liquid chromatography (RP-HPLC) with a Zorbax SB C18 column (150 mm x 0.3 mm, 3.5 µm particle size, 300 Å pore size, Agilent Technologies) and a solvent flow rate of 4 µL/min. was used for separating the tryptic digests. Gradient of 5%-38% consisting of 0.1% formic acid and acetonitrile was used during elution. Auto MS mode was used to choose peaks for MS/MS analysis. Analysis was performed in positive ion mode using Bruker captive electrospray source with a dry nitrogen gas temperature of 200 °C with a flow rate of 3 L/min with enhanced scanning mode (8,100 m/z per second).

Peaks were picked in the LC-MS/MS chromatogram using default settings in Bruker Data Analysis 4.2 software. Bruker Proteinscape bioinformatics suite coupled with MASCOT 2.1 (http://www.matrixscience.com) was used to search and match LC-MS/MS data against Col H FASTA sequences in the uniprot format (Col H: Q46085). Scaffold (version Scaffold_4.8.4, Proteome Software Inc., Portland, OR) was used to validate MS/MS based peptide and protein

identifications [**17**]. Scaffold software was used to give high confident peptide identifications using identification probability and false discovery rates. Sequence coverage for each band revealed that small amounts of Col H lose C-terminal domains during protein expression (see figure 3).

3.2h. Col H SAXS studies

Small Angle X-ray Scattering (SAXS) was used to find a low-resolution envelope for both full length Col H enzyme as well as Col H non-catalytic segments bound to different mini-collagens. All mini-collagens were synthesized by Dr. Takaki Koide using Fmoc-based solid-phase method on Rink amide resin [**18**].

For Col H, SAXS data was collected for full length Col H, both wild type and Ser mutant versions of Col H segments apo and complexed to $[(POG)_{10}]_3$ and $[GPRG(POG)_{13}]_3$. To analyze complexes of mini-collagen with Col H Ser mutants, a 5 mg/ml $[GPRG(POG)_{13}]_3$ stock was formed by dissolving 1 mg of the powder in 200 µl of 50 mM Tris pH 7.5, 1 mM $CaCl_2$. Solution was heated at 90° C for three minutes, then cooled to room temperature and placed in the refrigerator and the triple helix formed overnight. $[(POG)_{10}]_3$ was formed using the same method. Circular dichroism was used to confirm that the triple helix formed (see figure 4) [**1**]. Samples for PKD2CBD, 4Ser PKD2CBD, PKD1PKD2CBD, 2Ser PKD1PKD2CBD and 6Ser PKD1PKD2CBD complexed to $[GPRG(POG)_{13}]_3$ were prepared by allowing equal molar amounts of proteins and $[GPRG(POG)_{13}]_3$ to incubate together at 4° C overnight. After incubation, each complex was dialyzed into 50 mM Hepes pH 7.5, 100 mM NaCl, 5 mM $CaCl_2$ as this buffer was previously found to be suitable for collagenase non-catalytic domains complexed to mini-collagen [**19**]. SAXS data for PKD2CBD in a complex with $[(POG)_{10}]_3$ was collected as a control by using the same technique. For free 4Ser PKD2CBD, 2Ser

PKD1PKD2CBD and 6Ser PKD1PKD2CBD 50 mM Tris pH 7.5, 1 mM $CaCl_2$ was used as the buffer since the proteins are stable in this buffer. For SAXS full length Col H was dialyzed into the pCa 3 buffer used for Col G. Full-length Col H was previously found to be stable in high Ca^{2+} buffer [**20**]. As such this buffer was used for Col H during SAXS

For each sample decreasing concentrations were pipetted into a skirted 96 well SAXS plate from VWR (30 µl – 20 µl; 5 wells for each sample, 3 containing protein and 2 containing matching buffer; concentration should be no lower than 1 mg/ml). Buffer from either the last dialysis exchange or the last centrifugation step were used for dilutions and for the matching buffer wells. High-Throughput SAXS data was collected at the SIBYLS beamline 12.3.1 (Advanced Light Source; Berkeley, CA) using the SIBYLS mail-in program [**21**]. Plate was centrifuged at 3700 rpm for 10 minutes at 4°C to remove air bubbles. SAXS data was collected at 10°C using a Pilatus 2M detector (Detector distance: 1.5 m, X-ray wavelength: 1.127 Å). Each sample was exposed to the beam for .3 seconds for 33 to 50 exposures.

Scatter version 3.0 from Bioisis containing ATSAS 2.7.2 software was used to analyze scattering data for each sample. Data from the concentration gradient that were not affected by either aggregation, radiation damage or detector saturation were scaled to the highest concentration in the series to identify concentration dependent behavior [**22**]. Radius of gyration (R_G) and maximum diameter (D_{Max}) of each sample was found using Scatter and GNOM. To find R_G, Guinier peak analysis function in Scatter was used. The q-range used in constructing the Guinier plot was found by adding and removing points so that the best fit line in the low q-range of data was linear [**23**] [**24**]. The q-range was further refined by adding or removing points so that the normalized Guinier plot follows that of an ideal, globular particle [**24**]. To find D_{Max}, pair distribution plots were constructed in GNOM [**25**]. The number of points omitted at the

beginning is equal to the number omitted before the start of the q-range. Plots were constructed so that they smoothly approached P(r) = 0 (D_{Max}). Porod-Debye (s vs. $I*s^4$) plots and dimensionless Kratky plots (sR_g^2 vs. (sR_g^2) x I(q)/I(0)) were constructed to analyze protein fold and flexibility [**26**].

Ab initio shape reconstruction of each sample was carried out using DAMMIF; calculations were repeated twenty-three times and the resulting envelopes were averaged and filtered using DAMAVER and DAMFILT [**27**] [**28**]. Solved crystal structures for each Col H domain (Col H CBD: 3JQW, Col H PKD2: 4JGU, Col H PKD1: 4U7K, Col H/Col G catalytic domain: 2Y6I, 2Y3U, 4ARE and 4ARF) were docked into SAXS derived envelopes based binding, NMR and previous SAXS data using Chimera [**29**] [**12**] [**30**] [**31**]. For mini-collagen, $[(POG)_{10}]_3$ or [GPRG$(POG)_{13}$], were built from $[(PPG)_{10}]_3$ template (PDB: 1K6F) and additional residues were added to template using PyMol. To help position domains into SAXS derived envelopes FASTA sequences from Col H (Uniprot format; Col H: Q46085) was used to orient N and C-termini of each domain.

3.3 Results and Discussion

3.3a. Introduction of Ser mutations along a proposed PKD collagen binding ridge

In Col H PKD domains 1 and 2 surface exposed aromatic residues (PKD1: Tyr689, Phe712 PKD2: Tyr780, His782, Tyr796, Tyr801) are present along a potential collagen binding ridge [**5**]. In Col G CBD2 mutation of Tyr residues in collagen binding cleft was found to greatly affect binding [**12**]. Structure based sequence alignment for PKD domains revealed that PKD domains in collagenases with only one CBD contain surface aromatics while PKD domains in collagenases with two CBDs do not [**5**]. Conservation of PKD surface aromatic residues in collagenases with only one CBD suggest that the domains assist in binding to collagen. In order

to investigate if Col H surface exposed aromatic residues are involved in binding to substrate PKD1PKD2CBD and PKD2CBD with surface exposed aromatic residues mutated to Ser were expressed [**5**]. The following mutants were expressed and purified PKD2CBD Tyr780Ser, PKD2CBD His782Ser, PKD2CBD Tyr796Ser, PKD2CBD Tyr801Ser, 4Ser PKD2CBD, Tight binding PKD2CBD, 2Ser PKD1PKD2CBD and 6Ser PKD1PKD2CBD. Circular dichroism spectra for all samples indicates that proteins are folded with predominantly beta-sheet structure as predicted (see figure 5) [**5**].

3.3b. WT PKD2CBD, 4Ser PKD2CBD and single point mutant binding results

Previously performed insoluble bovine type I collagen binding experiment for wild type PKD2CBD was repeated as a control [**3**]. While the binding data was not identical to literature values Saturation plots showed two phases of binding for the protein. Binding information could not be derived from saturation plots so a scatchard plot was created. Scatchard plot confirmed that biphasic binding with a tighter binding phase (K_2 = 2.75 x 10^6 +/- 1.07 x 10^6 M, B_{Max2} = 0.087 +/- 0.021 nmole/mg of collagen) and a less tight binding phase (K_1 = 1.72 x 10^5 +/- 7.33 x 10^4 M, B_{Max1} = 0.523 +/- 0.187 nmole/mg of collagen) (see figure 6) [**3**]. The reported binding information for WT PKD2CBD is as follows K_2 = 2.95 x 10^6 M, B_{Max2} = 0.201 nmole/mg of collagen and K_1 = 4.74 x 10^5 M, B_{Max1} = 0.628 nmole/mg of collagen [**3**]. Difference in B_{Max} values between experimental and literature values may be due to variations in substrate.

To demonstrate that surface aromatic residues on PKD2 play a role in binding to collagen the binding experiment was repeated for 4Ser PKD2CBD. Saturation and Scatchard plots for 4Ser PKD2CBD show only one phase of binding (K = 5.34 x 10^5 +/- 9.43 x 10^4 M, B_{Max} = 0.187 +/- 0.023 nmole/mg of collagen) with the tighter binding phase being lost (see figure 6). Knocking out all surface aromatic residues on PKD2 seems appears to keep the domain from

interacting strongly with collagen fibril. In order to investigate if there is a key residue responsible for binding the experiment was repeated for the following PKD2CBD single point mutants, i.e. Y780S, H782S, Y796S and Y801S and saturation and scatchard plots were created for each. Among the single point mutants, only H782S and Y796S exhibited a single binding phase with roughly identical binding information (for H782S PKDCBD $K = 1.07 \times 10^6$ +/- 1.92×10^5 M, $B_{Max} = 0.210$ +/- 0.026 nmole/mg of collagen and for Y796S PKDCBD $K = 1.49 \times 10^6$ +/- 2.11×10^5 M, $B_{Max} = 0.132$ +/- 0.011 nmole/mg of collagen) (see figure 7). Results for Y780S and Y801S were biphasic similar to the wild type (Y780S: $K_2 = 8.83 \times 10^6$ +/- 2.56×10^6 M, $B_{Max2} = 0.069$ +/- 0.01 nmole/mg of collagen fibril and $K_1 = 7.62 \times 10^4$ +/- 1.09×10^5 M, $B_{Max1} = 1.34$ +/- 0.059 nmole/mg of collagen fibril; Y801S: $K_2 = 5.90 \times 10^6$ +/- 3.65×10^6 M, $B_{Max2} = 0.056$ +/- 0.016 nmole/mg of collagen; $K_1 = 1.64 \times 10^5$ +/- 1.85×10^5 M, $B_{Max1} = 0.433$ +/- 0.231 nmole/mg of collagen) (see figure 8).

3.3c. WT PKD1PKD2CBD, 2Ser PKD1PKD2CBD and 6Ser PKD1PKD2CBD binding results

In order to determine if PKD1 plays a role in binding Col H to collagen fibril the binding experiment was performed for wild type PKD1PKD2CBD and saturation and scatchard plots were made. Binding results were nearly identical to those of the wild type PKD2CBD (biphasic; $K_2 = 3.18 \times 10^6$ +/- 7.90×10^5 M, $B_{Max2} = 0.057$ +/- 0.008 nmole/mg of collagen; $K_1 = 1.85 \times 10^5$ +/- 9.96×10^4 M, $B_{Max1} = 0.235$ +/- 0.126 nmole/mg of collagen) (see figure 9). Decrease in B_{Max} values from wild type PKD2CBD may be due to steric hindrance from addition of PKD1. PKD1 does not interact strongly enough with collagen to form a third phase of binding. However PKD1 may contribute to binding by acting together with PKD2 on collagen fibril. In order to examine this, binding experiment was performed for 2Ser PKD1PKD2CBD. For 2Ser PKD1PKD2CBD binding is biphasic ($K_2 = 1.38 \times 10^7$ +/- 1.22×10^7 M, $B_{Max2} = 0.031$ +/- 0.007 nmole/mg of

collagen; K_1 = 4.26 x 10^5 +/- 2.87 x 10^5 M, B_{Max1} = 0.102 +/- 0.014 nmole/mg of collagen) and the second phase seems to be much tighter for 2Ser mutant than for wild type PKD1PKD2CBD (see figure 9). Tighter binding could be from potential hydrogen bonding between Ser hydroxyl group and amine group from collagen α-chains [**12**]. While plausible, this seems as other PKD2CBD Ser mutants do exhibit significantly increased binding to insoluble substrate. To determine if CBD alone can anchor full-length Col H to collagen fibril, binding experiment for 6Ser PKD1PKD2CBD was performed. For PKD1PKD2CBD with all PKD surface aromatics knocked out binding was single phased (K_D = 5.94 x 10^5 +/- 1.32 x 10^5 M, B_{Max} = 0.079 +/- 0.011 nmole/mg of collagen) (see figure 9). Results indicate that CBD by itself may be strong enough to anchor Col H to collagen fibril.

3.3d. Small angle X-ray Scattering of Col H domains

SAXS data was collected in order to give a three-dimensional shape of Col H domains interacting with collagenous peptide. Since SAXS measures solution scattering insoluble collagen fibrils cannot be used to model protein/collagen interaction [**32**]. In order to model the interaction between proteins and collagen, complexes with water soluble collagen-like peptides (mini-collagen) were formed. Mini-collagen are short peptides with collagen like sequence, generally Pro-Hyp-Gly (POG), repeated throughout the peptide that assemble into triple helix when dissolved into buffer [**33**]. $[GPRG(POG)_{13}]_3$ was used as a the mini-collagen for SAXS experiments with $(POG)_{10}$ being used as a control. Collagenase segments have been shown to bind readily to mini-collagen and is thought to mimic how enzyme binds collagen α-chains [**30**] [**31**] [**19**]. After dissolving mini-collagen in buffer, solution was allowed to equilibrate overnight and triple helix formation was confirmed using circular dichroism (see figure 4) [**1**].

Small Angle X-ray Scattering (SAXS) intensity was collected for the following samples; PKDCBD, PKD2CBD:$[(POG)_{10}]_3$, PKD2CBD:$[GPRG(POG)_{13}]_3$, 4 Ser PKD2CBD, 4 Ser PKD2CBD:$[GPRG(POG)_{13}]_3$, PKD1PKD2CBD, PKD1PKD2CBD:$[GPRG(POG)_{13}]_3$, 2 Ser PKD1PKD2CBD, 2 Ser PKD1PKD2CBD:$[GPRG(POG)_{13}]_3$, 6 Ser PKD1PKD2CBD, 6 Ser PKD1PKD2CBD:$[GPRG(POG)_{13}]_3$ and full length Col H. For PKD2CBD, 4 Ser PKD2CBD, PKD1PKD2CBD, 2 Ser PKD1PKD2CBD and 6 Ser PKD1PKD2CBD scattering data was collected to understand how the Col H domains come together in three-dimensional space. SAXS data for PKD2CBD was previously collected by Dr. Ryan Bauer and was only reprocessed for this experiment. Envelopes for PKD2CBD and PKD1PKD2CBD were used to determine where the enzyme is in relation to the mini-collagen. SAXS profiles, guinier plots, pair distribution plots and Kratky and porod-debye plots for Col H samples are shown in figures 9 – 14. Dimensionless kratky plots indicate that all samples are folded (see figures 10 – 15) [**26**].

Briefly, radius of gyration (R_G) is a way to give a rough estimate on the shape of the molecule and can be used to identify aggregation in a protein [**24**]. D_{Max} value found from pair distribution function gives the maximum diameter present in a molecule [**24**]. D_{Max} values are used in the development of SAXS derived envelopes. NSD values are used to determine the agreement between derived envelopes used to calculate an average structure [**24**]. Q-range is the range of scattering values from initial scattering profile used in deriving SAXS envelope. The R_G, D_{Max}, NSD and Q-range for all samples are summarized in Tables 2. Identical to Col G, agreement between Col H scattering profiles and theoretical profiles from crystal structure coordinates for each envelope was also found by using Crysol from Atsas online server. Results are listed in appendix.

Table 2

Sample	R_G (Å) Guinier Approx.	R_G (Å) Defined by P(r)	D_{Max} (Å)	Mean NSD value	Q-range ($Å^{-1}$)
PKD2CBD	26	26	87	0.509	0.0188 – 0.3319
PKD2CBD:$[(POG)_{10}]_3$	33	35	145	0.592	0.0129 – 0.3220
PKD2CBD:$[GPRG(POG)_{13}]_3$	33	39	180	0.572	0.0111 – 0.3932
4 Ser PKD2CBD	25	26	90	0.512	0.0195 – 0.3528
4 Ser PKD2CBD:$[GPRG(POG)_{13}]_3$	32.3	37	180	0.545	0.0111 – 0.3924
PKD1PKD2CBD	33	35	110	0.686	0.0111 – 0.3912
PKD1PKD2CBD:$[GPRG(POG)_{13}]_3$	40	48	220	0.551	0.0109 – 0.2890
2 Ser PKD1PKD2CBD	36	36	115	0.778	0.0164 – 0.2946
2 Ser PKD1PKD2CBD:$[GPRG(POG)_{13}]_3$	44	50	210	0.605	0.0129 – 0.3220
6 Ser PKD1PKD2CBD	33	34	110	0.749	0.0184 – 0.3570
6 Ser PKD1PKD2CBD:$[GPRG(POG)_{13}]_3$	40	46	220	0.576	0.0139 – 0.3751
Col H	40	41	140	0.798	0.0124 – 0.3215

3.3e. SAXS derived envelopes for Col H segments free and bound to mini-collagen

To construct each PKD2CBD:mini-collagen complex the envelopes for each complex and for reprocessed PKD2CBD were superimposed with each other. The crystal structures for

$[(POG)_{10}]_3$ (PDB: 1K6F) and $[GPRG(POG)_{10}]_3$ (built from $(POG)_{10}$ code) were positioned in the most rod-like regions of the envelopes while CBD was positioned on the C-terminus of the mini-collagens as determined by previous works for Col G's CBD2 [**30**] [**31**]. Binding surface of CBD derived from mutagenesis and collagen binding studies was positioned to interact with the C-terminus of mini-collagen [**12**] [**30**] [**34**]. PKD2 was positioned in the remaining electron density. For WT PKD2CBD with $[(POG)_{10}]_3$ and $[GPRG(POG)_{10}]_3$ it was discovered that PKD2 hangs off the end of mini-collagen and does not make any contact (see figure 16A). Construction of 4Ser:$[GPRG(POG)_{13}]_3$ complex was performed in the same manner as listed with envelope for 4Ser PKD2CBD being used in place of PKD2CBD (see figure 16B). The results for 4Ser PKD2CBD were very similar to that of the wild type. SAXS envelopes for PKD2CBD and 4 Ser PKD2CBD may explain surface plasmon resonance results from Sekiguchi et al. 2018. Plasmon resonance results for binding to mini-collagen revealed that measured dissociation constants for Col H CBD and PKD2CBD are similar (CBD: 75.2 +/- 0.41 x 10^{-5} M, PKD2CBD: 44.5 +/- 0.55 x 10^{-5} M) [**10**]. When CBD binds to C-terminus of mini-collagen linker to PKD2 must not be long enough to allow the domain to interact with substrate. While it is known that Col G CBD2 binds to C-terminus of mini-collagen with the front facing the N-terminus the directionality of binding for Col H is unknown [**30**].

In order to construct the complexes for samples involving PKD1PKD2CBD the position of PKD1 in free PKD1PKD2CBD was first determined. The positions of CBD and PKD2 in PKD2CBD can be inferred based on the envelopes for the wild type complexes. Envelope for wild type PKD2CBD was superimposed with golf club shaped envelope for PKD1PKD2CBD. CBD was positioned in the "head" of the club with PKD2 following. PKD1 was then positioned in the remaining electron density envelope (see figure 17A). PKD1PKD2CBD complexes were

constructed by superimposing the reprocessed envelope for PKD1PKD2CBD with the envelope for PKD1PKD2CBD:[GPRG$(POG)_{13}]_3$. [GPRG$(POG)_{13}]_3$, CBD and PKD2 were positioned based on the results for PKD2CBD:mini-collagen envelopes and PKD1 was placed in the remaining electron density (see figure 17A). Both PKD1 and PKD2 do not make contact the same mini-collagen as CBD. For both 2Ser and 6Ser PKD1PKD2CBD:[GPRG$(POG)_{13}]_3$ complexes were constructed using the same method with 2Ser and 6Ser PKD1PKD2CBD envelopes used in place of WT PKD1PKD2CBD (see figure 17B). The results for 2Ser and 6Ser PKD1PKD2CBD were identical to that of the wild type.

3.3f. SAXS derived structure of full-length Col H

In order to investigate how the full length enzyme is able to hydrolyze collagen the SAXS derived envelope for full-length Col H at pCa 3 was superimposed with the previously solved complexes for PKD2CBD:[GPRG$(POG)_{13}]_3$ and PKD1PKD2CBD:[(GPRG$(POG)_{13}]_3$. Envelope for full-length Col H has a globular shape with a protrusion (see figure 16A). Based on results from Ohbayashi et al. 2013 CBD was positioned in the "head" region and PKD2 was positioned on the protrusion based on how the Col H envelope overlapped with PKD2CBD and PKD2CBD:[GPRG$(POG)_{13}]_3$ envelopes (see figure 18A) [**20**]. Next PKD1 was positioned based on how the Col H envelope overlaps with envelopes for PKD1PKD2CBD and PKD1PKD2CBD:[GPRG$(POG)_{13}]_3$. Catalytic domain was positioned in the remaining electron density. Crystal structure for full catalytic domain of Col H has not been solved, however crystal structure for its peptidase domain has been solved (PDB: 4ARF) [**35**]. Therefore, the crystal structure of catalytic domain from Col G (PDB: 2Y3U) was used to find the position of the activator domain. Peptidase domain from Col H with peptide like inhibitor was superimposed

with Col G collagenase module (R.M.S.D. = 0.878 Å). Col G and Col H catalytic domains are homologous and share sequence identity of 48% and sequence similarity of 66% [**7**] [**36**] [**37**].

N-terminal residue in CBD crystal structure is not present in sequence of Col H CBD. As such it was hidden and new N-terminal residue (Val862) of CBD crystal structure and C-terminal residue (Asp860) of PKD2 crystal structure are oriented as close as possible distance between the two was found to be 14 Å (see figure 18B). The single residue missing between CBD and PKD2 would not be enough to make up the 14 Å distance. Coloring both CBD and PKD2 Cα by B-factor showed that N-terminus of CBD has a high B-factor relative to the rest of the domain which may make it dynamic enough to make up the distance without interfering with the Ca^{2+} binding residues. For average B-factor of CBD N-terminus Cα B-factors for residues 861 to 867 were chosen for finding the average B-factor as Glu868 side chain is involved in chelating Ca^{2+} ions. For the PKD2 and PKD1 overlapping residues in PKD2 were hidden so that no residues are missing between N-terminus of PKD2 and C-terminus of PKD1. The distance between the new N-terminal residue of PKD2 (Leu771) and the C-terminal residue of PKD1 (Asp770) was found to be 9 Å. Coloring the Cα atoms for the N-terminal linker of PKD2 and the C-terminal linker of PKD1 by relative B-factor reveal that N-terminal linker of PKD2 and C-terminal linker of PKD1 have high B-factors relative to their respective structure which may allow the two domains to connect (see figure 18C). PKD1 was positioned within superimposed envelope for PKD1PKD2CBD so that the N-terminal residue of PKD1 (Asn685) was as close as possible to C-terminal residue of Col H peptidase domain (Gly681). The distance between the two residues was found to be 33 Å. While two residues are missing between C-terminus of peptidase domain and N-terminus of PKD1 that would not bridge the 33 Å gap. The C-terminal region of the peptidase domain is known as the helper domain and is used to help the catalytic domain fold

correctly [**7**]. In the full-length enzyme, the helper domain may be folded differently than in the peptidase domain alone which may help link PKD1 to peptidase domain. Unpublished high-speed atomic force microscopy (AFM) results from Osaka University for full-length Col H reveals that the enzyme is highly dynamic which may help to link the domains (see Appendix video 2). Since the method used for positioning domain can only give a relative position CBD, PKD2, PKD1 and catalytic domain may not be positioned optimally in SAXS envelope for full-length Col H. In the future, SASREF from ATSAS online server may need to be used to find the best position for each domain in the envelope.

3.3g. PKD2 Surface Aromatic residues could allow PKD2CBD to wedge between tropocollagen molecules

Mechanism of how PKD's aromatic side chains can assist in binding tighter with collagen fibril is currently unknown. The binding clefts for CBDs from both Col G and Col H contain conserved Tyr and Phe residues that constitute 'hot spots' for CBD binding to collagen [**12**] [**34**]. Mutation of these residues in CBD from Col G either obliterated or diminished binding to mini-collagen [**12**]. HSQC-titration of mini-collagen to CBD2 from Col G also demonstrated their interaction to mini-collagen [**30**]. The orientation of aromatic residues in the binding pocket of Col H closely resemble those of Col G. Therefore, both Col G and Col H are likely to use these same residues to bind to collagen fibril and mini-collagen. The aromatic residues may interact with the Pro residues of collagen through aromatic CαH-π interactions as demonstrated for model peptides Pro and aromatic residues, i.e. Tyr, Phe and Trp [**38**] [**39**].

While addition of PKD2 by itself cannot bind to collagen, it was shown to enhance collagen fibril binding of CBD but not so much to mini-collagen [**3**] [**10**]. Catalytic domain by itself (gelatinase) does have the ability to degrade acid solubilized collagen and denatured

collagen it cannot degrade the ability degrade insoluble collagen fibril [**40**] [**3**]. PKD2 has the ability to enhance binding of CBD to collagen fibril however, PKD2 and PKD1PKD2 by themselves have no affinity for collagen fibril [**3**]. To explain PKD2's ability to enhance collagen binding for CBD we speculate that PKD2CBD has the ability to wedge between adjacent tropopocollagens within a micro-fibril [**41**]. Previously described unpublished high-speed AFM results show that free Col H is a highly dynamic enzyme which may allow to wiggle between adjacent tropocollagens (see video 2 in Appendix). Wedges work by converting a force applied to its blunt end to into perpendicular forces. Blunt end force may be the dynamic movement of PKD2CBD which causes PKD2CBD to bind tighter than CBD alone.

Once PKD2CBD is wedged, aromatic residues on PKD2 may interact to collagen via CαH-π interactions [**39**]. H782 and Y796 play a key-roles in allowing PKD2 to bind collagen fibril. In the crystal structure of PKD2 H782 and Y796 are located near the C-terminal linker of PKD2 [**5**]. Once CBD binds to a tropocollagen, H782 and Y796 could provide the initial surface for adjacent tropocollagen to bind. While they appear to not be as important for binding, Y780 and Y801 may also help PKD2 bind to adjacent tropocollagen. These interactions cause PKD2CBD to bind even tighter and explains the higher binding affinity phase seen in WT PKD2CBD [**3**].

3.3h. PKD1 is not involved in allowing Col H to bind to collagen

While binding results indicate that PKD1 appears to play no role in allowing Col H to interact with collagen fibril it could still play a role in collagenolysis. PKD1 could still act as either a spacer or collagen "sweller". PKD1 domain may provide the proper distance between CBD and catalytic domain. PKD1 have a similar role to that of PKD domain from Col G. Since it has no apparent binding apparent to insoluble collagen or mini-collagen PKD1's role may be

to align the active site of the catalytic domain with the binding clefts of a CBD [**6**] [**7**]. Other role for PKD domains demonstrated are that PKD domains from some serine proteases have been shown to "swell" collagen fibrils by increasing the distance between tropocollagens within a micro-fibril [**8**] [**42**]. Interactions between serine protease and tropocollagens increase the overall positive charge on collagen and electrostatic repulsive forces may cause individual collagen fibers to swell [**8**]. It is currently unknown if PKD, PKD1 or PKD2 can "swell" collagen fibrils. Future experiments are being planned to use AFM to measure the distance between microfibrils after being treated with either PKD1PKD2CBD or PKD2CBD [**8**] [**42**].

3.3i. SAXS derived envelope of Col H reveals PKD2 could feed a tropocollagen into catalytic domain

For a brief recap, in this chapter it was discovered that in SAXS experiments to visualized PKD1PKD2CBD and PKD2CBD complexed to mini-collagen PKD domains do not interact with mini-collagen. However, fibril binding assay wild type and mutant PKD1PKD2CBD and PKD2CBD proteins revealed that aromatic residues present on PKD2 enhance binding of CBD. Below we attempt to use these new pieces of information to tease out a mechanism for Col H collagen hydrolysis.

Superimposed PKD2CBD:$[GPRG(POG)_{13}]_3$ and PKD1PKD2CBD$[GPRG(POG)_{13}]_3$ complexes with the SAXS derived envelope for full-length Col H reveals that PKD2 could be used to feed a tropocollagen molecule into the catalytic domain. Due to binding results for WT PKD2CBD and 4Ser PKD2CBD, PKD2 was positioned so that aromatic residues involved are pointed away from the tropocollagen CBD is bound to. In PKD2CBD:$[GPRG(POG)_{13}]_3$ envelope PKD2 is positioned diagonally from CBD at the C-terminus of the mini-collagen. However if PKD2 was positioned so that it is side-by-side with CBD the distance from the

binding cleft of CBD to the surface aromatic residues of PKD2 would be approximately 15 Å. Consequentially, the distance from the center of one tropocollagen molecule to an adjacent one in a micro-fibril is approximately 15 Å [**43**]. A second tropocollagen molecule was positioned 15 Å from the crystal structure for $[GPRG(POG)_{13}]_3$ so that PKD2 aromatic residues would make contact. This second mini-collagen aligns almost perfectly with the active site cleft of the catalytic domain (see figure 19). PKD1PKD2CBD:$[GPRG(POG)_{13}]_3$ envelope suggests that PKD1 is necessary for positioning the catalytic domain. If it has any involvement in binding its contribution may not be large enough to be picked up by the methods used in this study. While CBD is key towards anchoring the enzyme, PKD2 surface aromatic residues allow Col H to wedge between adjacent tropocollagen molecules in a micro-fibril. PKD2 does not bind to tropocollagen by itself however, its potential loose interactions via surface aromatics may allow it be fed into the catalytic domain easier. It is possible that mini-collagen bound to CBD may be the one being fed into catalytic domain. Based on mechanism teased out form multiple SAXS envelopes tropocollagen bound to CBD does not align with active site cleft of peptidase domain. Positioning of domains in envelopes could be incorrect. Potential SAXS experiment with mutationally inactivated E416D Col H and mini-collagen would help to support or discredit current PKD2 focused hypothesis [**3**].

Col G prefers to bind to slightly untwisted regions in collagen and based on sequence similarity Col H may do the same [**31**] [**34**]. High speed AFM video shows that Col G prefers to attack disorder regions in a collagen micro-fibril mimic and the same could be true for Col H [**44**]. Col H would have an easier time wedging between tropocollagens at less tightly twisted regions in collagen fibril. While Col G is a processive enzyme there is no evidence to suggest that Col H has processive activity. However, processivity cannot be disproven until AFM

experiment with Col H is performed [**44**]. Limited proteolysis of acid solubilized type I, type II and type III collagens by Col H shows that the enzyme prefers to initially attack the hypersensitive sequence of tropocollagen located about 1/3 of way from N-terminus [**45**] [**46**]. After initial hydrolysis event Col H processes C-terminal fragment at specific sites [**45**] [**47**]. Col H's role in collagen hydrolysis may be to act as an endopeptidase hydrolyzing hypersensitive sites in collagen fibrils and micro-fibrils. Hydrolysis at these sites may open up new regions for Col G to begin processive hydrolysis.

3.3j. Col H segments are successful in increasing the systemic effects of Parathyroid Hormone

Like Col G segments, Col H segments can be used for localizing the effects of bFGF. For localization of bFGF using PKD2CBD is necessary for getting the best results, however Col G segment CBD1CBD2 is still the best for using bFGF [**48**] [**49**] [**50**] [**10**]. For biomolecules that have a more systemic effect, such as Parathyroid hormone (PTH), Col H non-catalytic segments are required for increasing their efficacy. Since they do not bind as tightly as Col G segments, they are more likely to be distribute throughout blood accessible collagen rich tissues such as skin and bone [**10**]. PTH fused to the N-terminus of CBD has been used to promote growth of bone and hair in mice [**51**] [**52**] [**53**] [**54**]. PTH-CBD was found to distribute to both skin and bone after intraperitoneal or subcutaneous injection [**52**]. PTH-CBD is an anabolic agent and works by directly targeting bone and skin [**55**]. Addition of PKD2, PKD2-CBD, was found to be less effective than using CBD alone and approximately equivalent to using PTH alone [**54**]. Preferable pharmacodynamics of CBD is potentially the reason for that. Linker between PTH and PKD2 may be too long and is cleaved in vivo before it can interact with its receptor [**54**].

Scatchard plot and saturation plot data has revealed how tight non-catalytic segments of Col H bind to collagen. Based on SAXS derived envelopes for PKD2CBD and PKD1PKD2CBD

bound to mini-collagen it was discovered that the N-terminal linker of PKD2 and PKD1 domain are not involved in binding to collagen. For increasing systemic affects using Col H segments, finding the optimal binding affinity for Col H segments is necessary. For WT PKD2CBD, PKD2 aromatic residues are key to binding the domain to collagen. However, even when these residues are knocked out PKD2CBD still has increased affinity for collagen [**3**]. Fusing PTH to N-terminus PKD1 in 6 Ser PKD1PKD2CBD or to the N-terminus of PKD2 in 4Ser PKD2CBD may be more efficacious than using PTH fused to CBD. In order to make these potential fusion proteins affective linker region between PTH and N-terminus of PKD2 may need to be rigid to prevent loss of PTH into the serum [**54**] [**55**].

3.4 Conclusion

For Col H, site directed mutagenesis of PKD1 and PKD2 coupled with a collagen binding revealed that PKD2 surface aromatics play a key role in binding the enzyme to substrate while PKD1 plays no apparent role in collagen binding [**3**]. SAXS derived envelopes of Col H non-catalytic segments PKD1PKD2CBD and PKD2CBD complexed with mini-collagen revealed that PKD2 works similar to CBD1 in Col G and allows Col H to intercalate between adjacent tropocollagens within a micro-fibril [**19**]. Superimposing full-length Col H envelope with the previous non-catalytic envelopes revealed that PKD2 is used to feed substrate into the catalytic domain. In order to support proposed mechanism of hydrolysis of Col H more data is needed. FRET imaging and High-speed AFM could potentially be used to visualize Col H on reconstituted collagen fibrils. Observing Col H on collagen fibril may reveal if micro-fibril wedging hypothesis is plausible. Understanding how Col H binds substrate would reveal best ways to link PTH for delivery.

Chapter 3: Figures

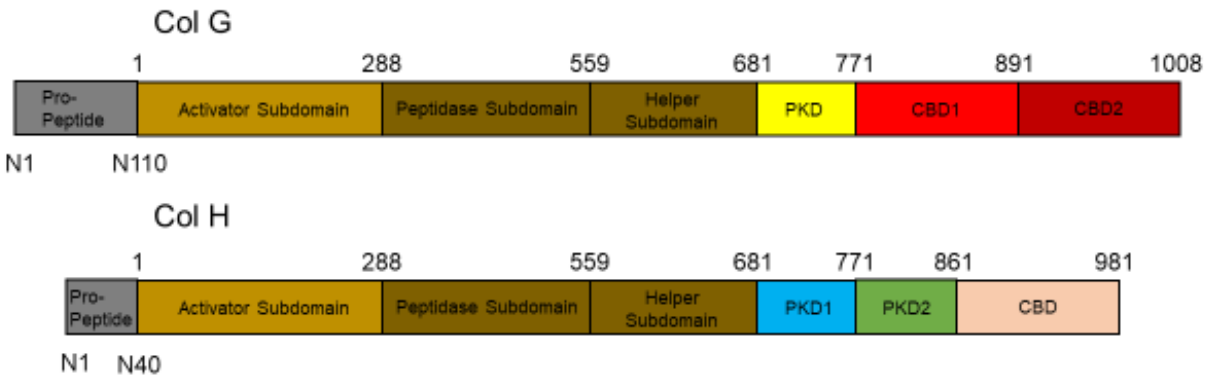


Figure 1: Domain map of Col G and Col H from *H. histolytica*. In order to activate Col G and Col H Pro-peptide (gray, Col G: N1 – N110 Col H: N1-N40) must be cleaved. Catalytic domain of Col G and Col H are composed of Activator sub-domains (light brown, residues 1 – 288) and Peptidase subdomain (dark brown, residues 289 – 559), which itself contains a helper subdomain to help the domain fold correctly (dart brown, residues 560 – 681). PKD domains (single PKD for Col G in yellow, residues 682 – 771; PKD1 and PKD2 for Col H in blue and green, residues 681 – 771 for PKD1 and residues 772 – 861 for PKD2) connect catalytic domains of both enzymes to CBDs. CBDs (CBD1 and CBD2 for Col G in red and dark red, residues 772 – 891 for CBD1 and residues 892 – 1008 for CBD2; single CBD for Col H in pink, residues 862 – 981) are responsible for binding both enzymes to collagen

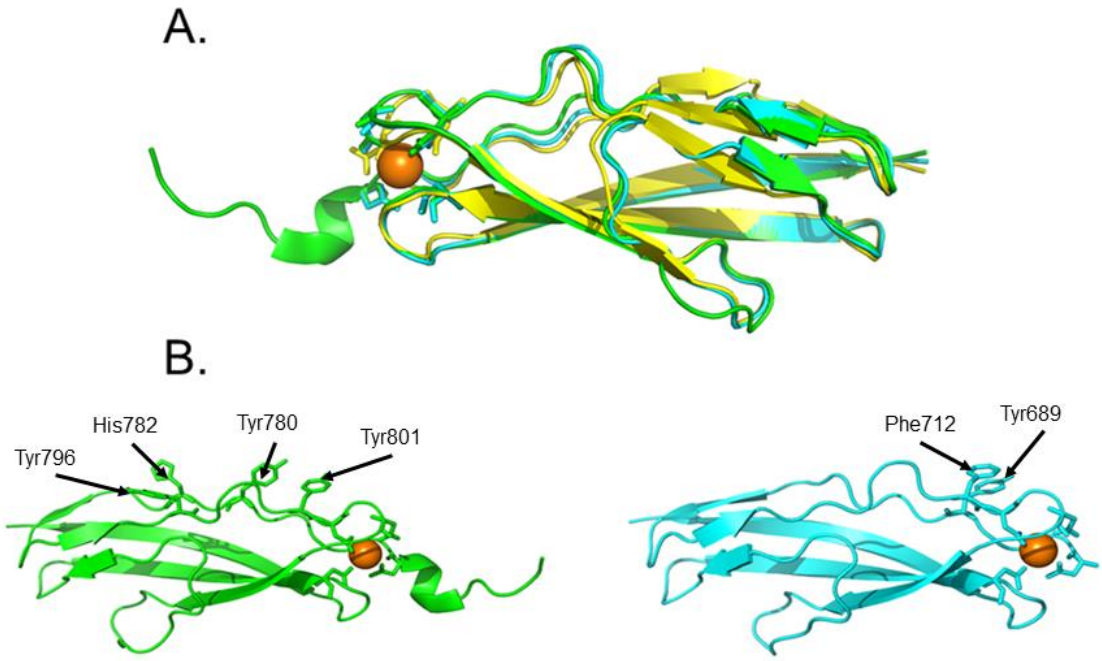


Figure 2: **A.** Structural alignment for Col G PKD (yellow), Col H PKD2 (green) and Col H PKD1 (cyan). R.M.S.D. PKD to PKD2 = 0.619 Å, R.M.S.D. PKD to PKD1 = 0.639 Å and R.M.S.D. PKD1 to PKD2 = 0.490 Å. Residues responsible for Ca^{2+} chelation are shown as sticks. **B.** Surface aromatic residues for PKD2 (Tyr780, His782, Tyr796 and Tyr801) and PKD1 (Tyr689 and Phe712) lie along a proposed collagen binding domain.

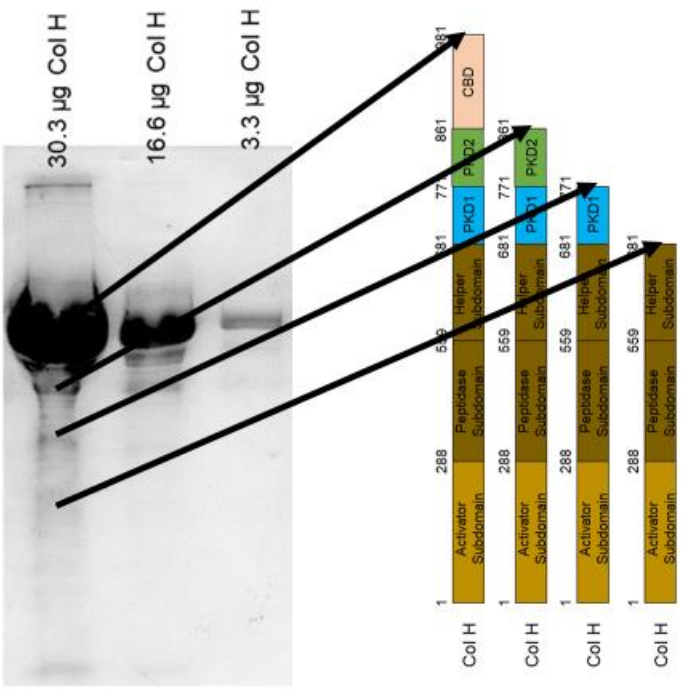


Figure 3: Greyscale image of Col H SDS PAGE gel with different amounts of protein. Not pictured, ImageJ densitometry results used to calculate purity. Rectangle was drawn around all bands for 30.3 µg well. Total intensity was calculated and intensity for each band was divided by total intensity to give purity. Purity was found to be 96%. Bands from were excised for trypsin digest and mass spectrometry analysis. FASTA sequences for full-length Col H (uniprot: Q46085) was used with mass spectrometry data to find the residue coverage for Col H. Results show that small amounts of Col H lose C-terminal domains during expression.

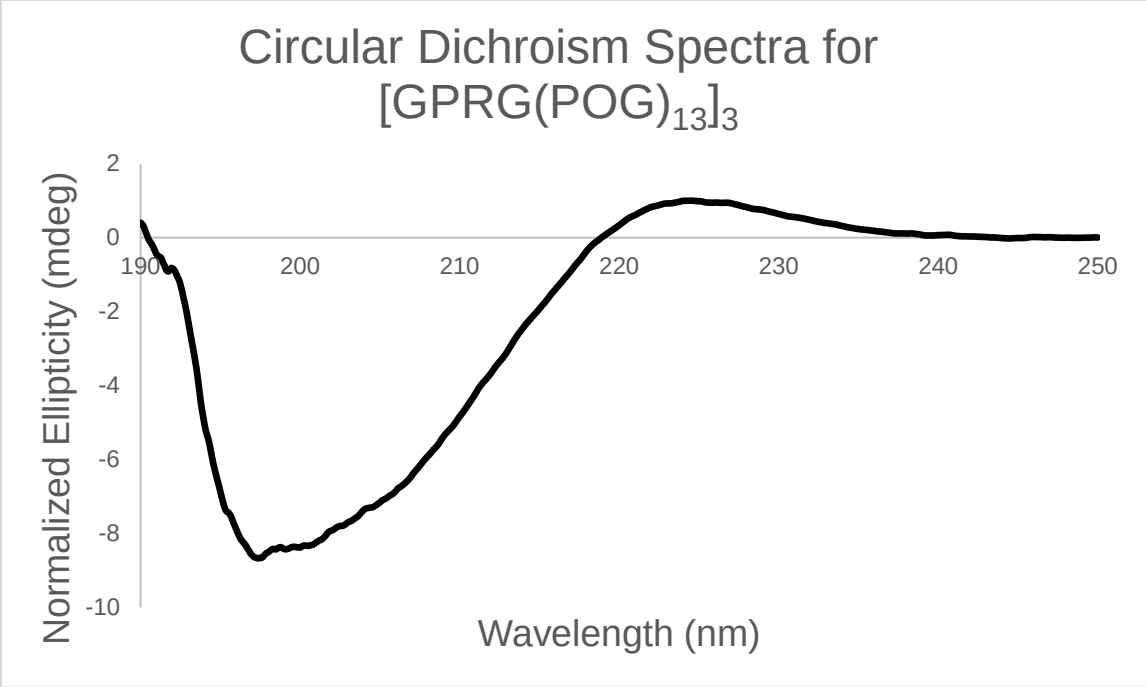


Figure 4: Circular dichroism spectrum for $[GPRG(POG)_{13}]_3$. Maximum at 225 nm and minimum at 198 nm is consistent with triple helical collagen [**1**].

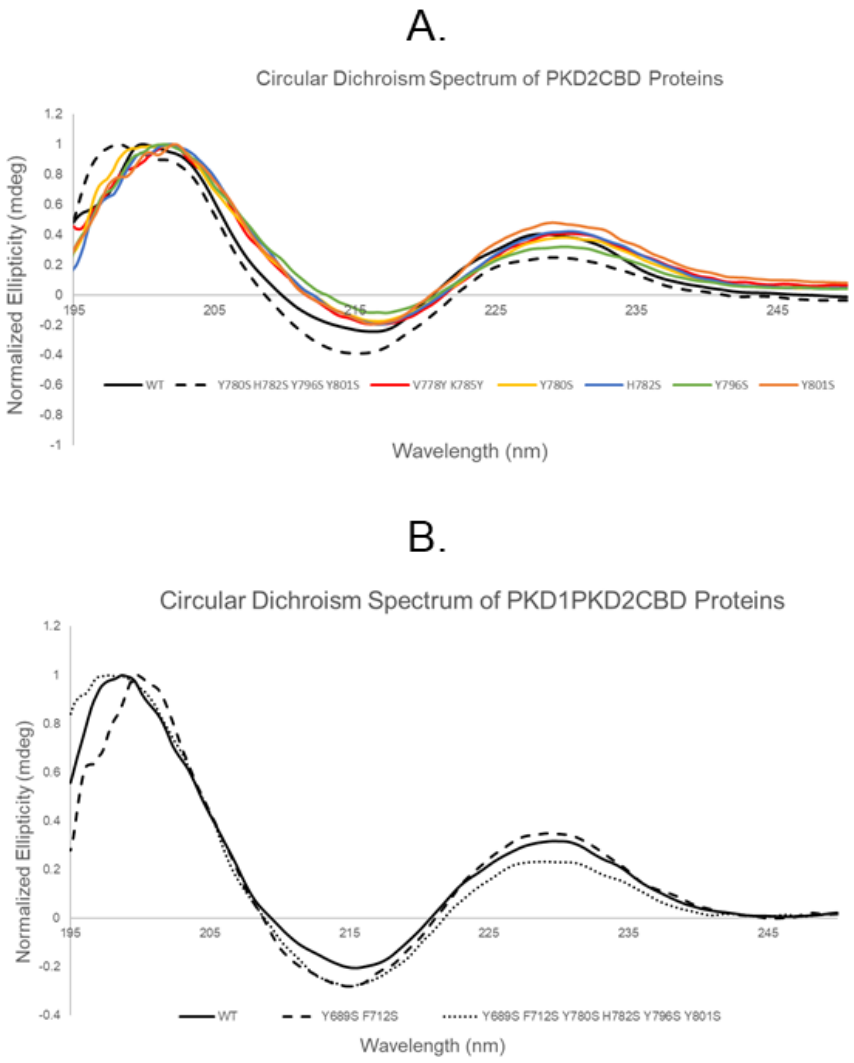


Figure 5: **A.** Circular dichroism spectra for PKD2CBD, 4 Ser PKD2CBD, V778Y K785Y PKD2CBD, Y780S PKD2CBD, H782S PKD2CBD, Y796 PKD2CBD, Y801S PKD2CBD. V778Y K785Y PKD2CBD is a potential tight binding mutant of PKD2CBD. V778Y K785Y PKD2CBD did not express well, only enough protein was produced for Circular dichroism. **B.** Circular dichroism spectra for PKD1PKD2CBD, 2 Ser PKD1PKD2CBD and 6 Ser PKD1PKD2CBD. Raw data was plotted on Excel and smoothed with exponential smoothing function.

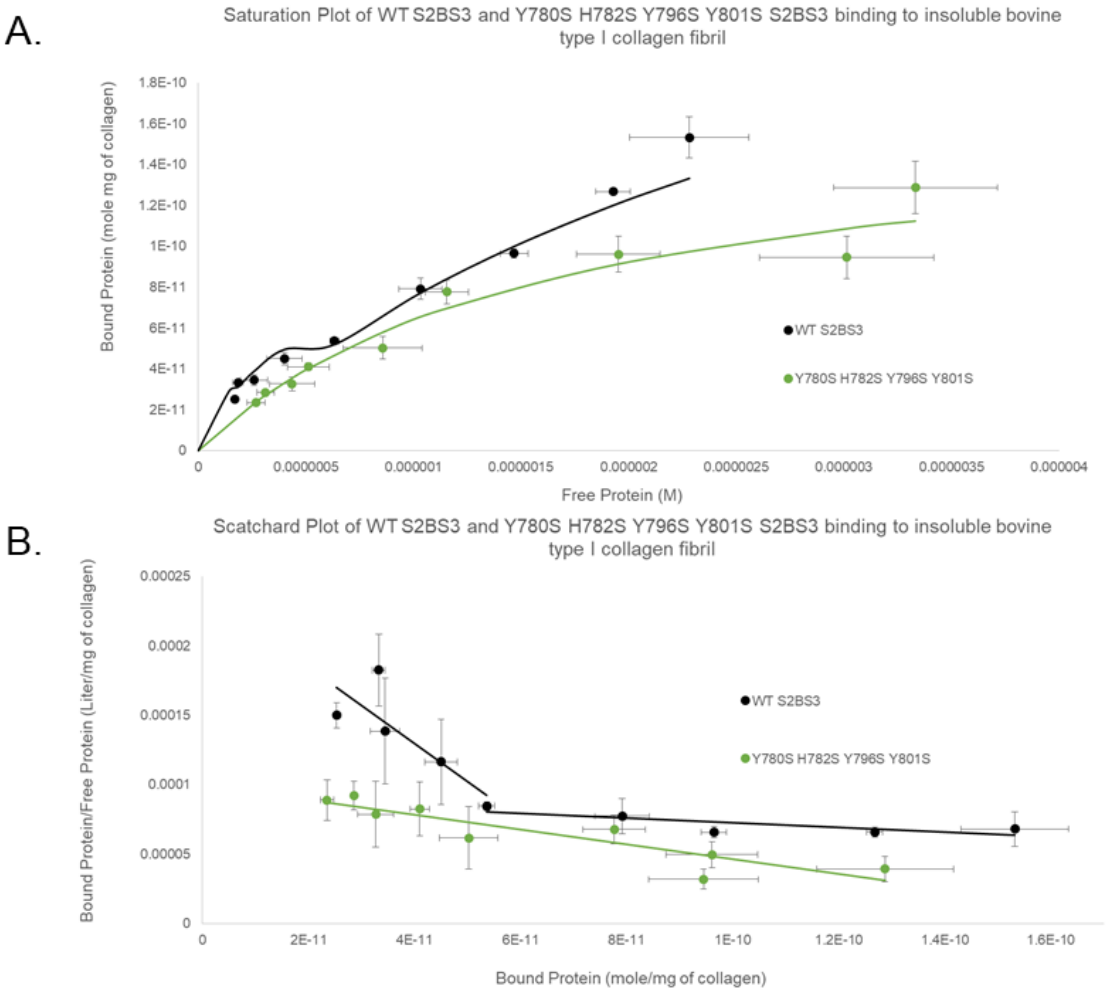


Figure 6: **A.** Saturation plot for PKD2CBD (black) and 4 Ser PKD2CBD (green) with bovine type I insoluble collagen fibril. Wild type PKD2CBD appears to still have two phases of binding while 4 Ser PKD2CBD lost a phase **B.** Scatchard plots for PKD2CBD and 4 Ser PKD2CBD also shows that 4 Ser PKD2CBD has lost a phase of binding. While uncertainty for scatchard plots are generally high plots should still be able to give association constant within an order of 10.

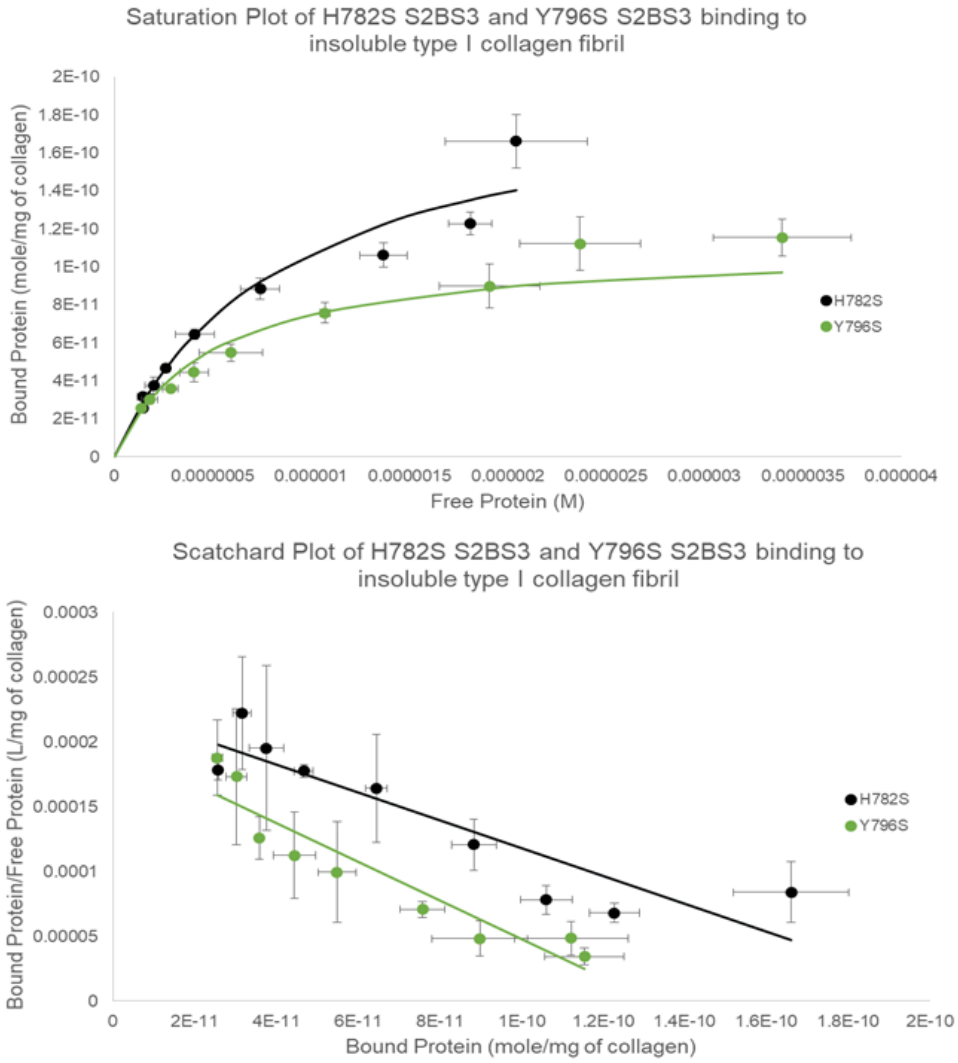


Figure 7: **A.** Saturation plot for H782S PKD2CBD (black) and Y796S PKD2CBD (green) with bovine type I insoluble collagen fibril. H782S PKD2CBD and Y796S PKD2CBD both show single phase in saturation plots. **B.** Scatchard plots for H782S PKD2CBD and Y796S PKD2CBD both confirm single phase of binding and were used to estimate association constant. In PKD2 crystal structure H782 and Y796 are closest to C-terminal linker to CBD. These two residues might provide the initial surface for binding to collagen fibril.

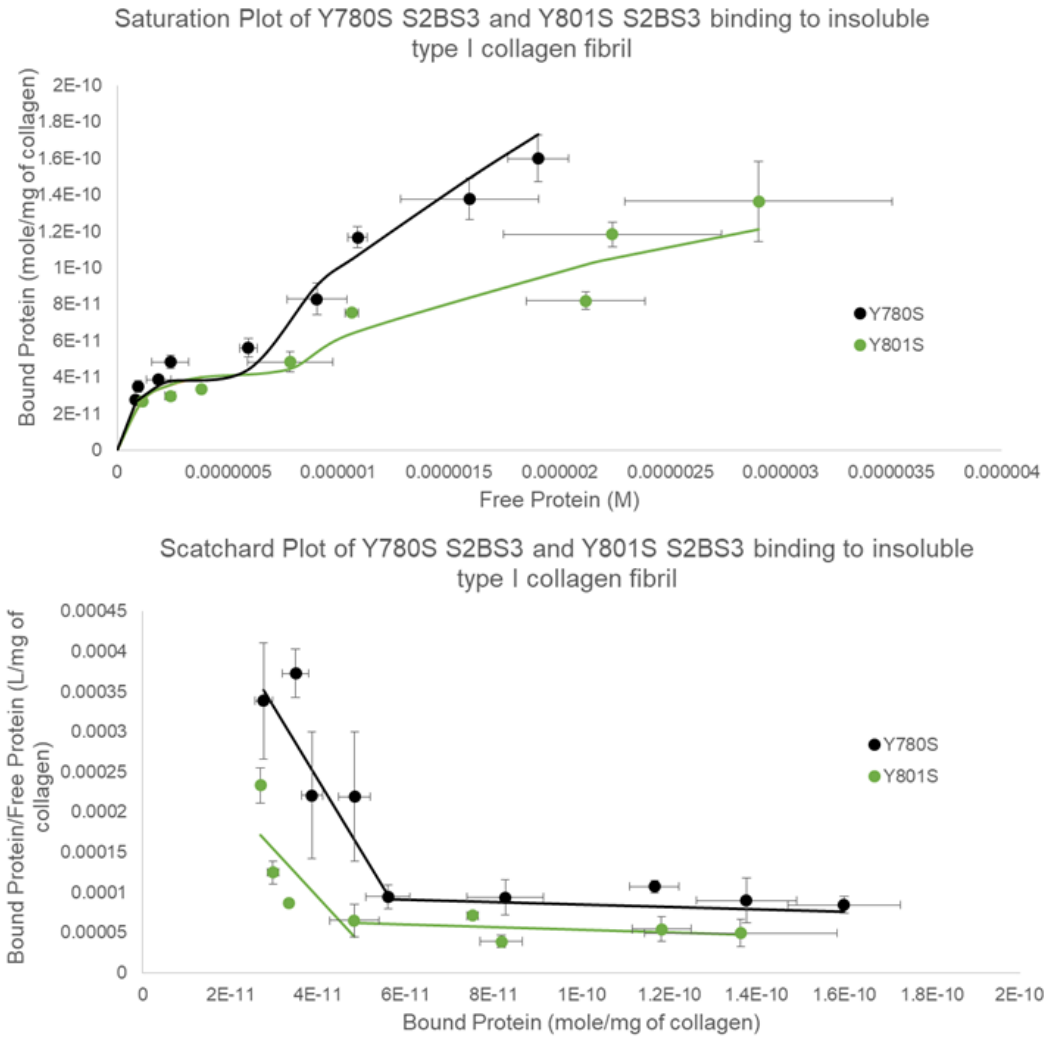


Figure 8: **A.** Saturation plot for Y780S PKD2CBD (black) and Y801S PKD2CBD (green) with bovine type I insoluble collagen fibril. Y780S PKD2CBD and Y801S PKD2CBD both show two phases in saturation plots. **B.** Scatchard plots for Y780S PKD2CBD and Y801S PKD2CBD both confirm single phase of binding and were used to estimate association constant. Results seem to give support to H782 and Y796 being responsible for binding in PKD2.

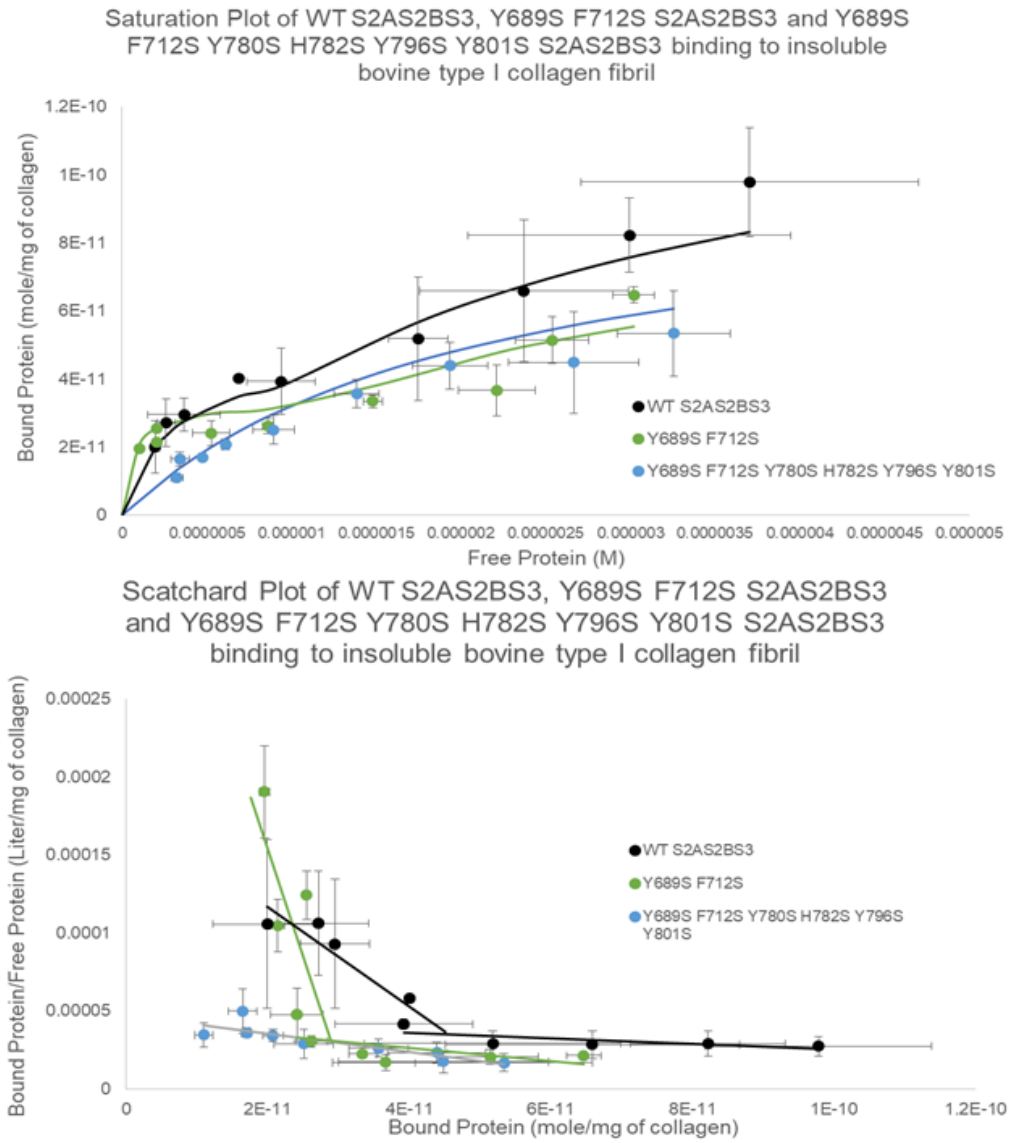


Figure 9: **A.** Saturation plot for PKD1PKD2CBD (black) 2 Ser PKD1PKD2CBD (green) and 6 Ser PKD1PKD2CBD (blue) with bovine type I insoluble collagen fibril. PKD1PKD2CBD and 2 Ser PKD1PKD2CBD both show biphasic binding in the saturation plot while 6 Ser PKD1PKD2CBD shows only a single phase. **B.** Scatchard plots for PKD1PKD2CBD, 2 Ser PKD1PKD2CBD and 6 Ser PKD1PKD2CBD confirm binding

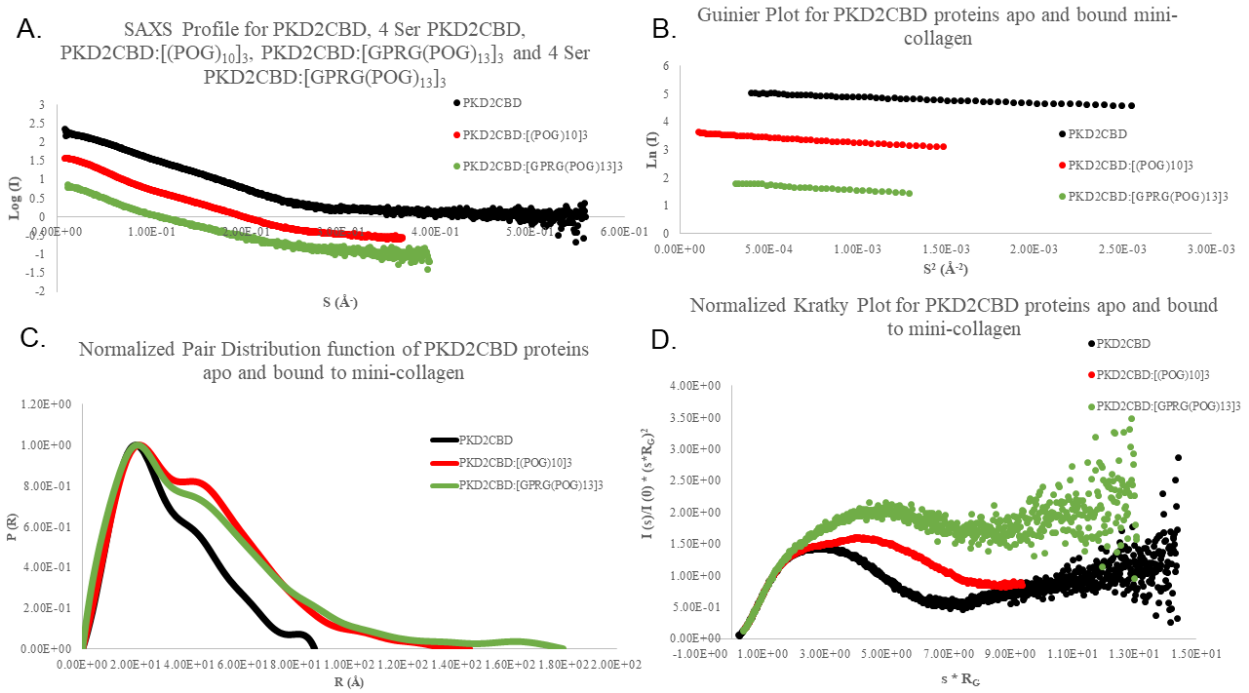


Figure 10: **A.** Experimental Scattering profile for PKD2CBD (black), PKD2CBD:$[(POG)_{10}]_3$ (red), PKD2CBD:$[GPRG(POG)_{13}]_3$ (green). **B.** Guinier plot for PKD2CBD, PKD2CBD:$[(POG)_{10}]_3$, PKD2CBD:$[GPRG(POG)_{13}]_3$. R_G values are calculated from slope of Guinier region using Guinier approximation. **C.** Normalized pair distribution function for PKD2CBD, PKD2CBD:$[(POG)_{10}]_3$, PKD2CBD:$[GPRG(POG)_{13}]_3$. Extended D_{Max} values are from addition of mini-collagen **D.** Dimensionless Kratky plot for PKD2CBD, PKD2CBD:$[(POG)_{10}]_3$, PKD2CBD:$[GPRG(POG)_{13}]_3$. Plateau like region for PKD2CBD:$[GPRG(POG)_{13}]_3$ may be due to extended mini-collagen used for complex making the structure appear more rod-like.

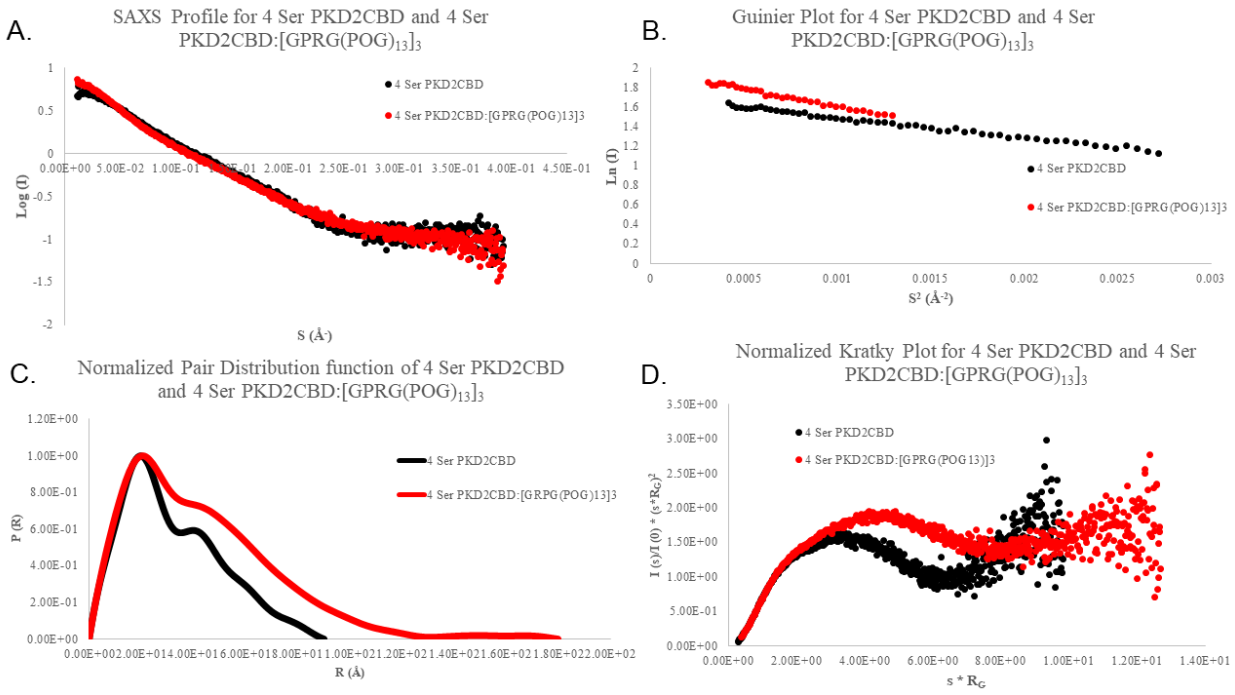


Figure 11: **A.** Experimental Scattering profile for 4 Ser PKD2CBD (black) and 4 Ser PKD2CBD:[GPRG(POG)$_{13}$]$_3$ (red). **B.** Guinier plot for 4 Ser PKD2CBD and 4 Ser PKD2CBD:[GPRG(POG)$_{13}$]$_3$. R_G values are calculated from slope of Guinier region using Guinier approximation. **C.** Normalized pair distribution function for 4 Ser PKD2CBD and 4 Ser PKD2CBD:[GPRG(POG)$_{13}$]$_3$. Extended D_{Max} values are from addition of mini-collagen **D.** Dimensionless Kratky plot for 4 Ser PKD2CBD and 4 Ser PKD2CBD:[GPRG(POG)$_{13}$]$_3$. Plateau like region for 4 Ser PKD2CBD:[GPRG(POG)$_{13}$]$_3$ may be due to extended mini-collagen used for complex making the structure appear more rod-like.

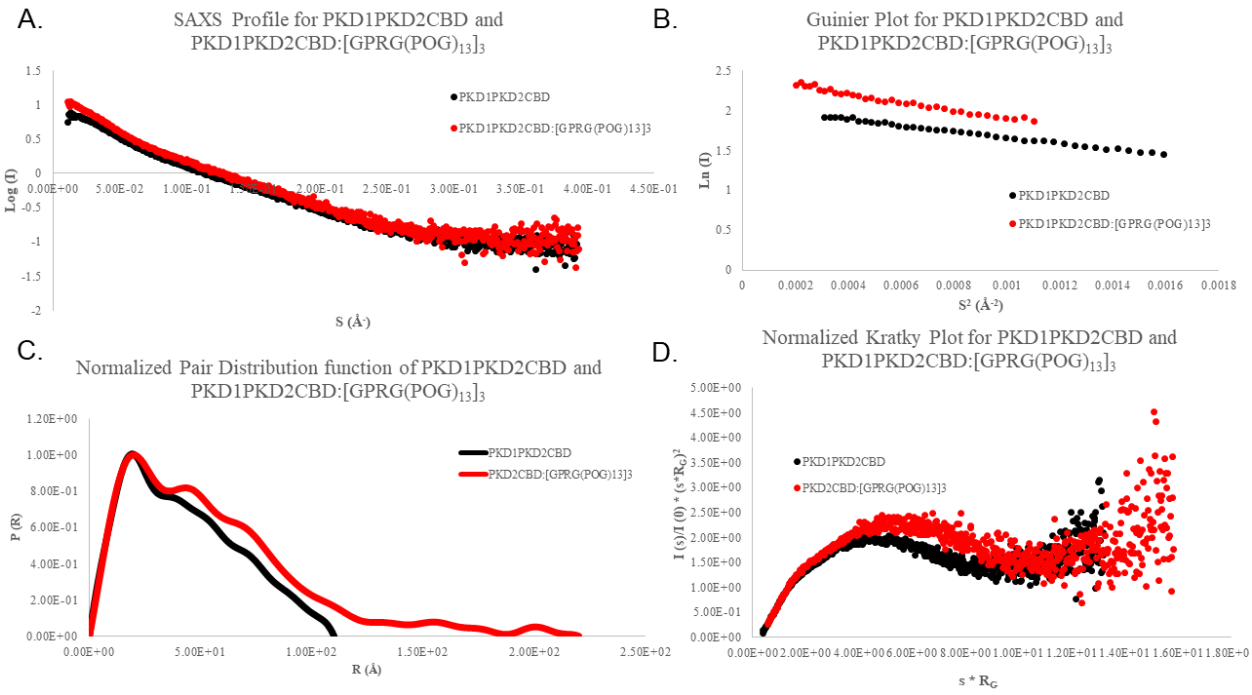


Figure 12: **A.** Experimental Scattering profile for PKD1PKD2CBD (black) and PKD1PKD2CBD:[GPRG(POG)$_{13}$]$_3$ (red). **B.** Guinier plot for PKD1PKD2CBD and PKD1PKD2CBD:[GPRG(POG)$_{13}$]$_3$. R_G values are calculated from slope of Guinier region using Guinier approximation. **C.** Normalized pair distribution function for PKD1PKD2CBD and PKD1PKD2CBD:[GPRG(POG)$_{13}$]$_3$. Extended D_{Max} values are from addition of mini-collagen **D.** Dimensionless Kratky plot for PKD1PKD2CBD and

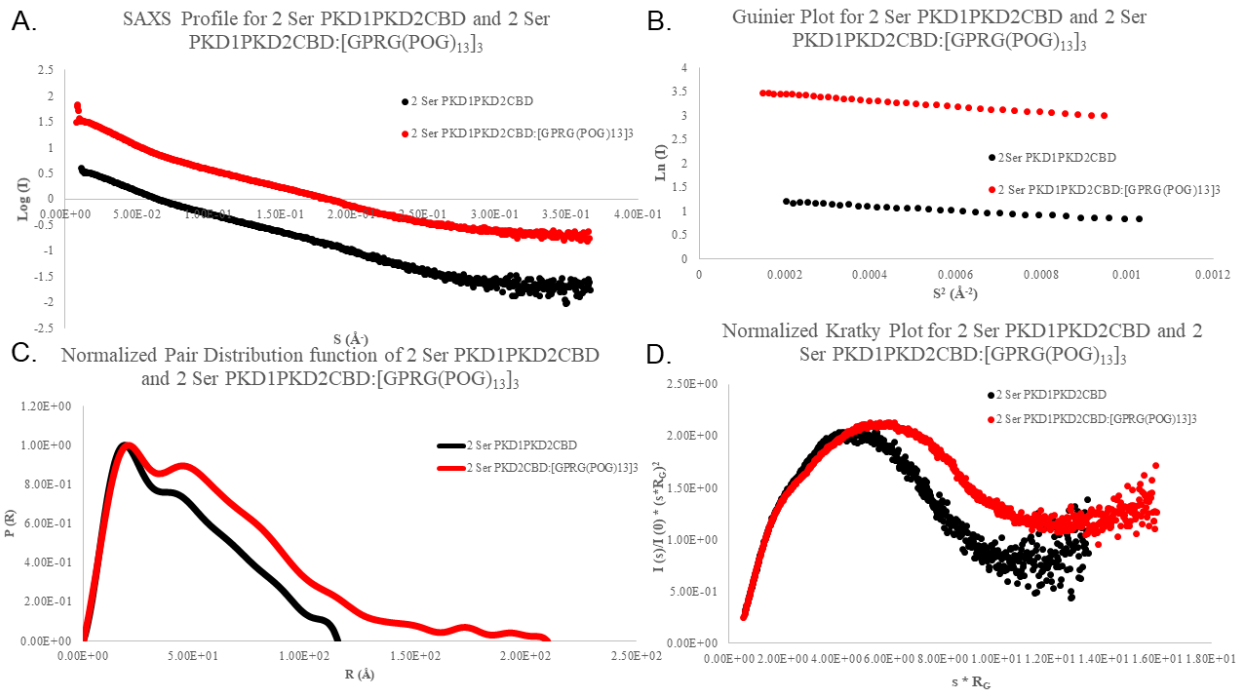


Figure 13: **A.** Experimental Scattering profile for 2 Ser PKD1PKD2CBD (black) and 2 Ser PKD1PKD2CBD:[GPRG(POG)$_{13}$]$_3$ (red). **B.** Guinier plot for 2 Ser PKD1PKD2CBD and 2 Ser PKD1PKD2CBD:[GPRG(POG)$_{13}$]$_3$. R_G values are calculated from slope of Guinier region using Guinier approximation. **C.** Normalized pair distribution function for 2 Ser PKD1PKD2CBD and 2 Ser PKD1PKD2CBD:[GPRG(POG)$_{13}$]$_3$. Extended D_{Max} values are from addition of mini-collagen **D.** Dimensionless Kratky plot for 2 Ser PKD1PKD2CBD and 2 Ser PKD1PKD2CBD:[GPRG(POG)$_{13}$]$_3$.

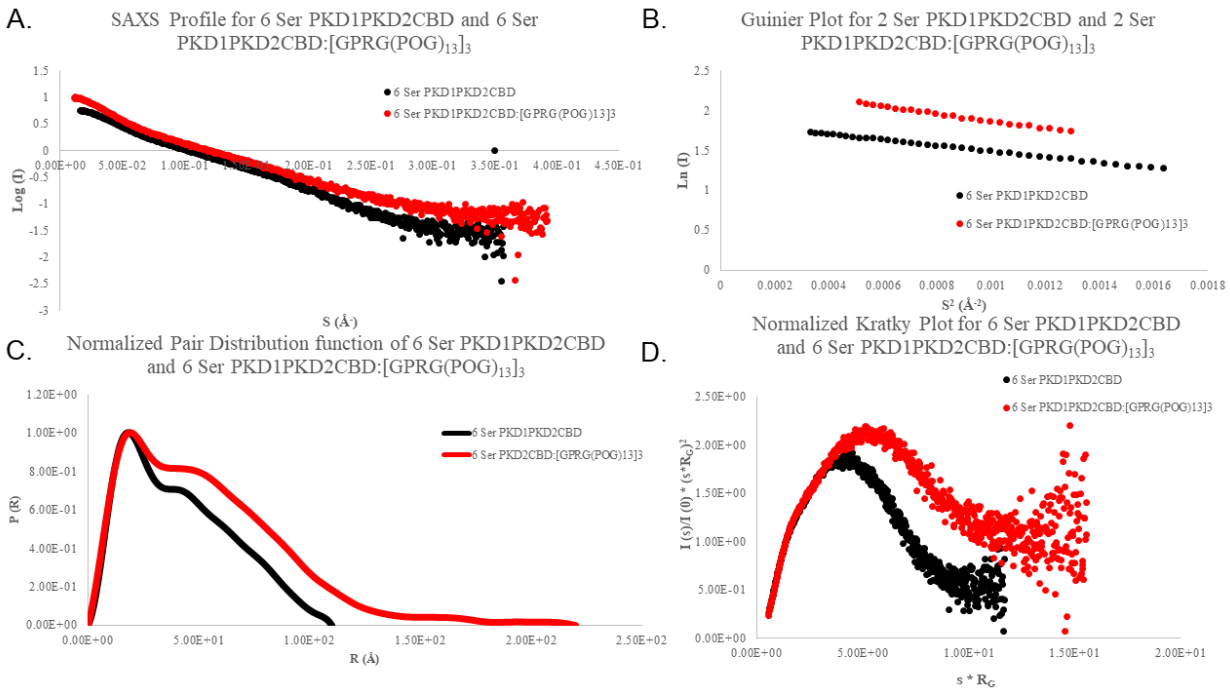


Figure 14: **A.** Experimental Scattering profile for 6 Ser PKD1PKD2CBD (black) and 6 Ser PKD1PKD2CBD:$[GPRG(POG)_{13}]_3$ (red). **B.** Guinier plot for 6 Ser PKD1PKD2CBD and 6 Ser PKD1PKD2CBD:$[GPRG(POG)_{13}]_3$. R_G values are calculated from slope of Guinier region using Guinier approximation. **C.** Normalized pair distribution function for 6 Ser PKD1PKD2CBD and 6 Ser PKD1PKD2CBD:$[GPRG(POG)_{13}]_3$. Extended D_{Max} values are from addition of mini-collagen **D.** Dimensionless Kratky plot for 6 Ser PKD1PKD2CBD and 6 Ser PKD1PKD2CBD:$[GPRG(POG)_{13}]_3$.

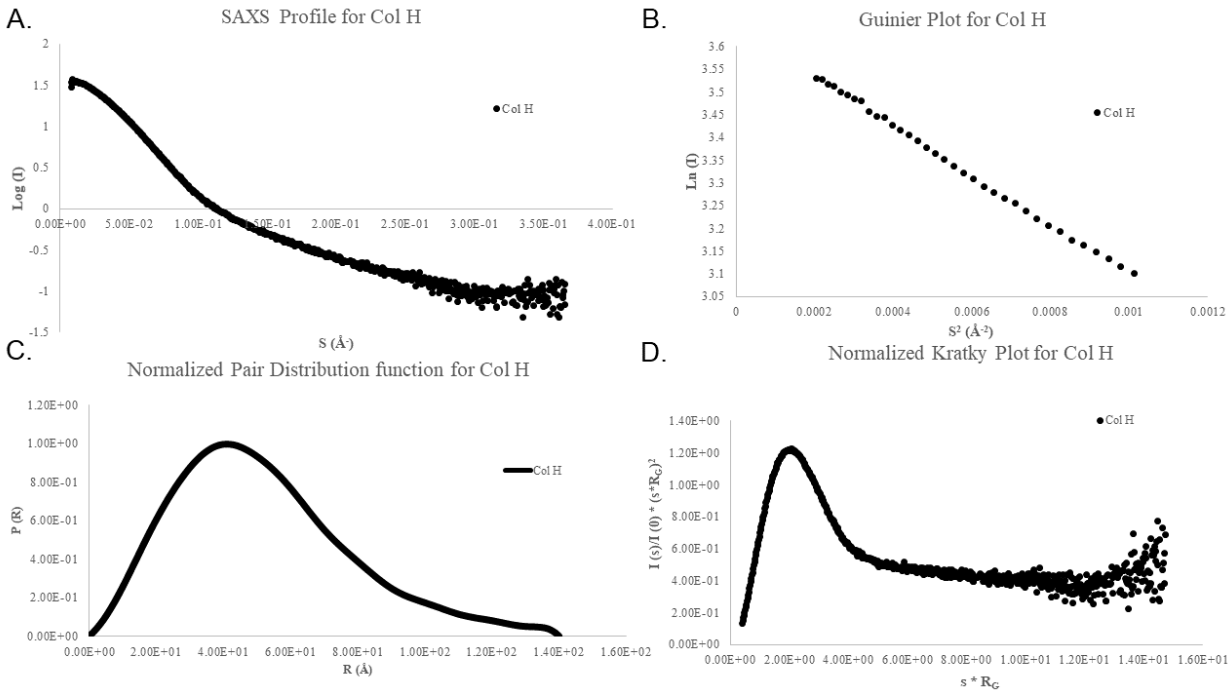


Figure 15: **A.** Experimental Scattering profile for Col H **B.** Guinier plot for Col H. R_G values was calculated from slope of Guinier region using Guinier approximation. **C.** Normalized pair distribution function for Col H. **D.** Dimensionless Kratky plot for Col H.

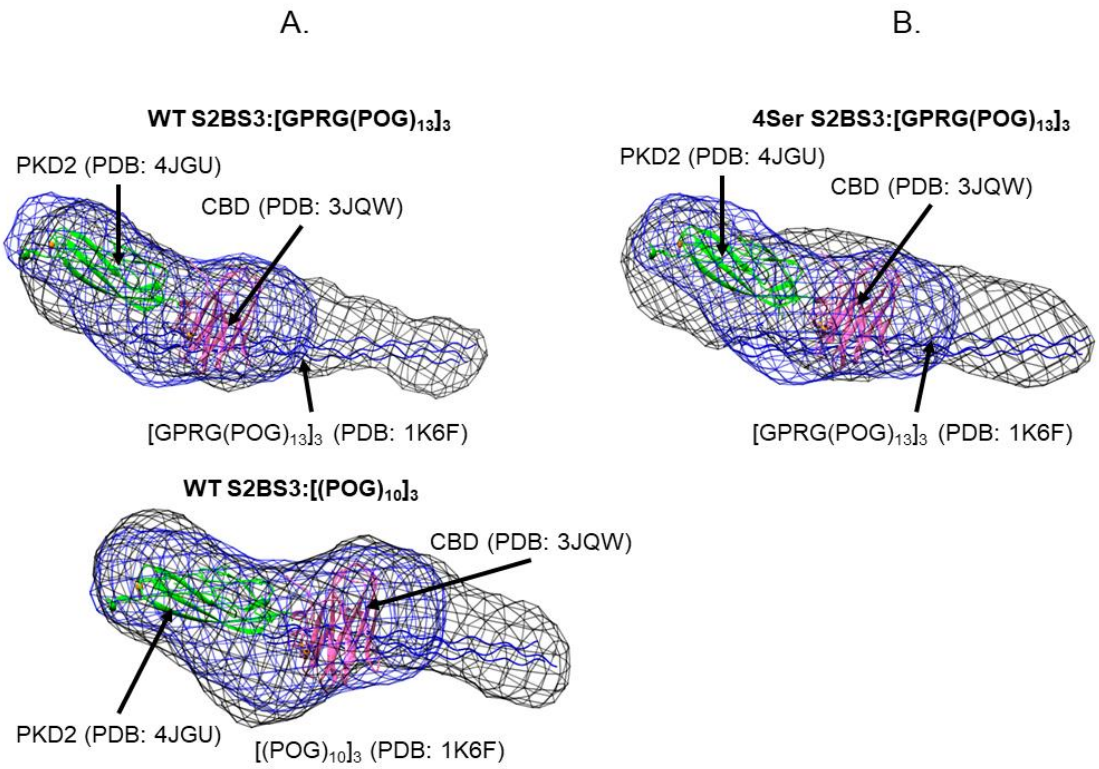


Figure 16: **A.** SAXS derived envelope for PKD2CBD bound to $[GPRG(POG)_{13}]_3$ and $[(POG)_{10}]_3$. Position of domains was determined by superimposing envelope for PKD2CBD with docked crystal structures with envelopes for PKD2CBD bound to $[GPRG(POG)_{13}]_3$ and $[(POG)_{10}]_3$ contrary to fibril binding results PKD2 does not appear to bind to mini-collagen **B.** SAXS derived envelope for 4 Ser PKD2CBD bound to $[GPRG(POG)_{13}]_3$ envelope appears identical to that for PKD2CBD bound to mini-collagen

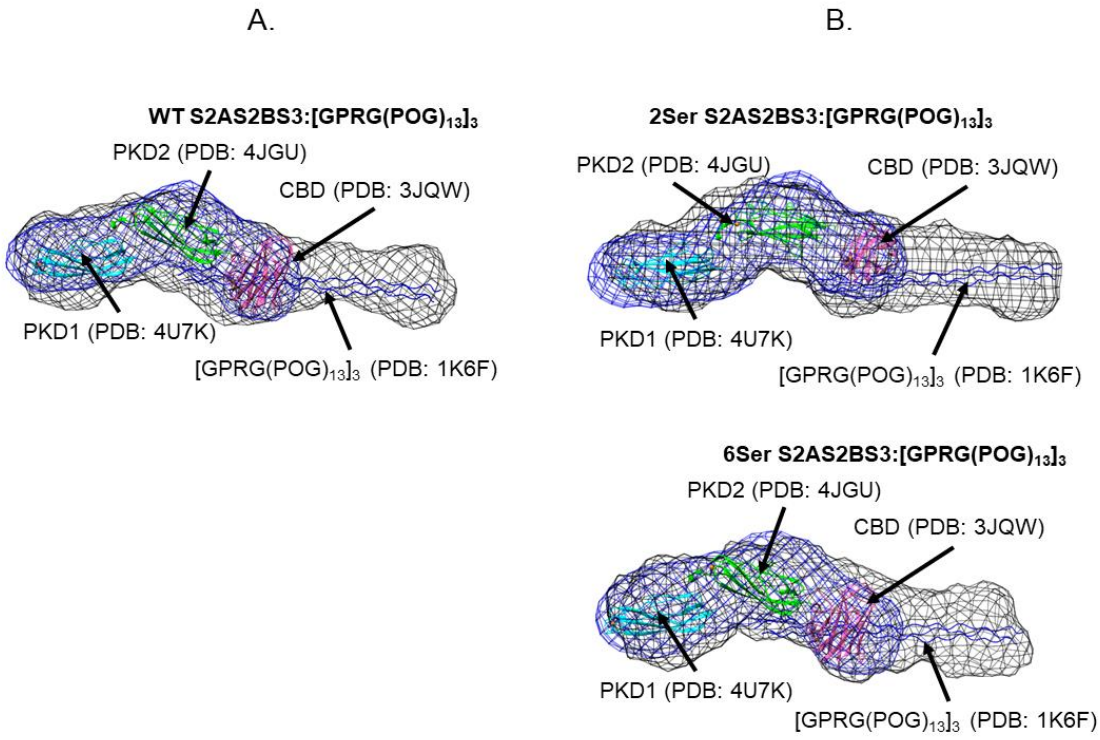


Figure 17: **A.** SAXS derived envelope for PKD1PKD2CBD bound to $[GPRG(POG)_{13}]_3$. Position of domains was determined by superimposing envelope for PKD1PKD2CBD with docked crystal structures with envelopes for PKD1PKD2CBD bound to $[GPRG(POG)_{13}]_3$. Similar to results for PKD2CBD bound to mini-collagen PKD1 and PKD2 do not appear to bind to mini-collagen **B.** SAXS derived envelope for 2 Ser PKD1PKD2CBD and 6 Ser PKD1PKD2CBD bound to $[GPRG(POG)_{13}]_3$. The envelope appears identical to that for

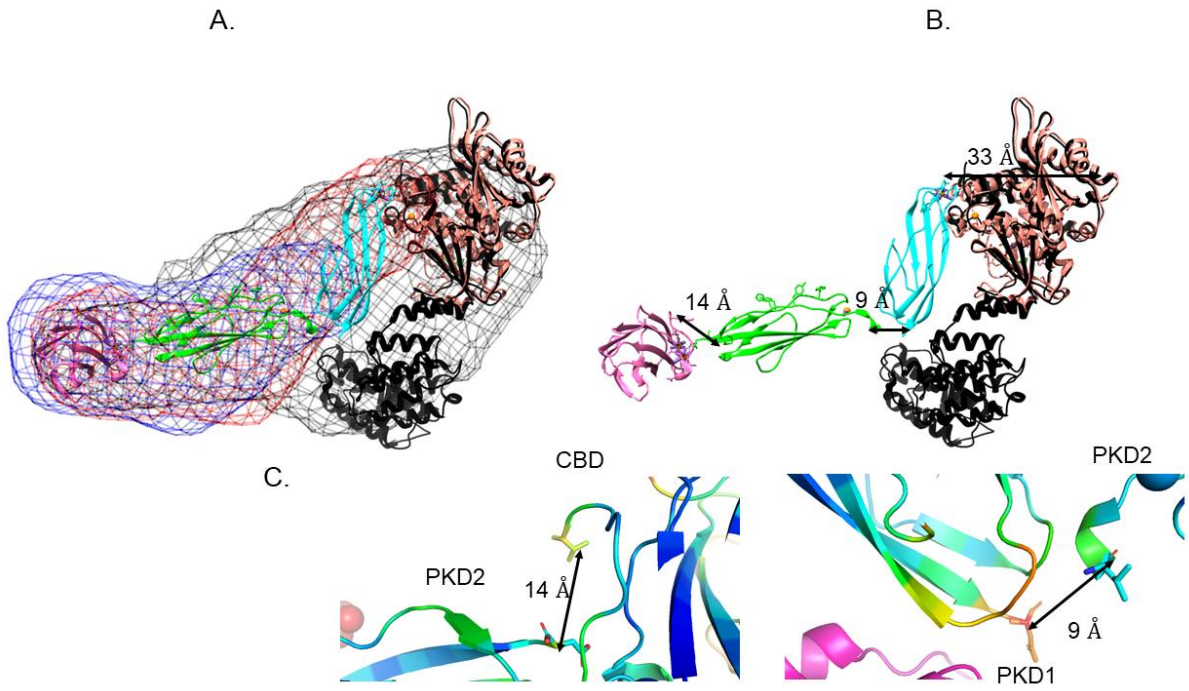


Figure 18: **A.** SAXS derived envelope for full-length Col H. Domains were positioned by superimposing envelope for PKD2CBD and PKD1PKD2CBD. Catalytic domain was positioned in remaining electron density. **B.** Figure A with envelopes removed. **C.** Zoomed in region between CBD/PKD2 and PKD2/PKD1 with Cα B-factors colored (Red: relative high B-factor, Blue: relative low B-factors). B-factors may allow the domains to link together to form the full-length enzyme.

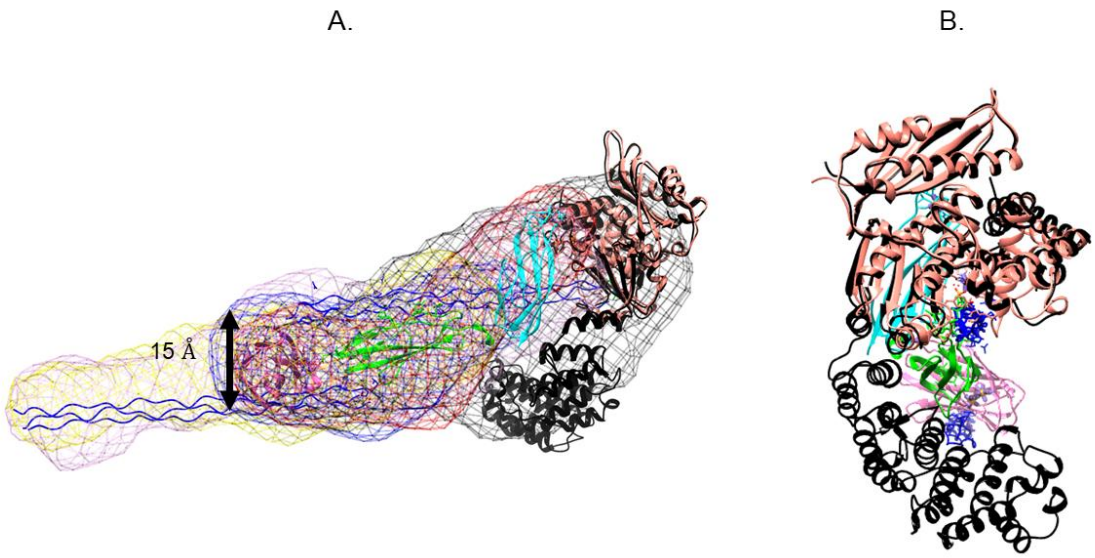


Figure 19: **A.** Teasing out collagen binding mechanism for Col H. Envelope for Col H was superimposed with envelopes for PKD1PKD2CBD (pink) and PKD2CBD (yellow) bound to $[GPRG(POG)_{13}]_3$ as well as with PKD1PKD2CBD (red) and PKD2CBD (blue) alone. Based on PKD2 wedging hypothesis a second $[GPRG(POG)_{13}]_3$ was built near PKD2 so that it makes contact with PKD2 residues. Distance from first mini-collagen to the second was found to be 15 Å the same as the distance between tropocollagens within a micro-fibril. Col H may use PKD2CBD to wedge between tropocollagens in a micro-fibril for hydrolysis. **B.** Figure A rotated 90°. Second $[GPRG(POG)_{13}]_3$ aligns almost perfectly with active site of Col H peptidase domain (in orange).

Works Cited

1. N. J. Greenfield, *Using circular dichroism spectra to estimate protein secondary structure*, Nature Protocols **1** (2006), no. 6, 2876-2890.

2. A. S. Duarte, A. Correia and A. C. Esteves, *Bacterial collagenases - a review*, Crit Rev Microbiol **42** (2016), no. 1, 106-126.

3. O. Matsushita, C. M. Jung, J. Minami, S. Katayama, N. Nishi and A. Okabe, *A study of the collagen-binding domain of a 116-kda clostridium histolyticum collagenase*, J Biol Chem **273** (1998), no. 6, 3643-3648.

4. *Polycystic kidney disease: The complete structure of the pkd1 gene and its protein. The international polycystic kidney disease consortium*, Cell **81** (1995), no. 2, 289-298.

5. R. Bauer, K. Janowska, K. Taylor, B. Jordan, S. Gann, T. Janowski, E. C. Latimer, O. Matsushita and J. Sakon, *Structures of three polycystic kidney disease-like domains from clostridium histolyticum collagenases colg and colh*, Acta Crystallographica Section D-Structural Biology **71** (2015), 565-577.

6. U. Eckhard and H. Brandstetter, *Polycystic kidney disease-like domains of clostridial collagenases and their role in collagen recruitment*, Biological Chemistry **392** (2011), no. 11, 1039-1045.

7. U. Eckhard, E. Schonauer, D. Nuss and H. Brandstetter, *Structure of collagenase g reveals a chew-and-digest mechanism of bacterial collagenolysis*, Nat Struct Mol Biol **18** (2011), no. 10, 1109-1114.

8. Y. K. Wang, G. Y. Zhao, Y. Li, X. L. Chen, B. B. Xie, H. N. Su, Y. H. Lv, H. L. He, H. Liu, J. Hu, B. C. Zhou and Y. Z. Zhang, *Mechanistic insight into the function of the c-terminal pkd domain of the collagenolytic serine protease deseasin mcp-01 from deep sea pseudoalteromonas sp sm9913 binding of the pkd domain to collagen results in collagen swelling but does not unwind the collagen triple helix*, Journal of Biological Chemistry **285** (2010), no. 19, 14285-14291.

9. J. Huang, R. Wu, D. Liu, B. Liao, M. Lei, M. Wang, R. Huan, M. Zhou, C. Ma and H. He, *Mechanistic insight into the binding and swelling functions of prepeptidase c-terminal (ppc) domains from various bacterial proteases*, Applied and environmental microbiology **85** (2019), no. 14, e00611-00619.

10. H. Sekiguchi, K. Uchida, O. Matsushita, G. Inoue, N. Nishi, R. Masuda, N. Hamamoto, T. Koide, S. Shoji and M. Takaso, *Basic fibroblast growth factor fused with tandem collagen-binding domains from clostridium histolyticum collagenase colg increases bone formation*, Biomed Res Int **2018** (2018), 8393194.

11. O. Matsushita, T. Koide, R. Kobayashi, K. Nagata and A. Okabe, *Substrate recognition by the collagen-binding domain of clostridium histolyticum class i collagenase*, J Biol Chem **276** (2001), no. 12, 8761-8770.

12. J. J. Wilson, O. Matsushita, A. Okabe and J. Sakon, *A bacterial collagen-binding domain with novel calcium-binding motif controls domain orientation*, EMBO (European Molecular Biology Organization) Journal **22** (2003), no. 8, 1743-1752.

13. S. Au - Zhang and M. D. Au - Cahalan, *Purifying plasmid dna from bacterial colonies using the qiagen miniprep kit*, JoVE (2007), no. 6, e247.

14. F. Sanger and A. R. Coulson, *A rapid method for determining sequences in dna by primed synthesis with dna polymerase*, J Mol Biol **94** (1975), no. 3, 441-448.

15. O. Matsushita, T. Koide, R. Kobayashi, K. Nagata and A. Okabe, *Substrate recognition by the collagen-binding domain of clostridium histolyticum class i collagenase*, Journal of Biological Chemistry **276** (2001), no. 12, 8761-8770.

16. C. A. Schneider, W. S. Rasband and K. W. Eliceiri, *Nih image to imagej: 25 years of image analysis*, Nat Methods **9** (2012), no. 7, 671-675.

17. R. Craig and R. C. Beavis, *A method for reducing the time required to match protein sequences with tandem mass spectra*, Rapid Commun Mass Spectrom **17** (2003), no. 20, 2310-2316.

18. H. Yasui, C. M. Yamazaki, H. Nose, C. Awada, T. Takao and T. Koide, *Potential of collagen-like triple helical peptides as drug carriers: Their in vivo distribution, metabolism, and excretion profiles in rodents*, Peptide Science **100** (2013), no. 6, 705-713.

19. P. Caviness, R. Bauer, K. Tanaka, K. Janowska, J. R. Roeser, D. Harter, J. Sanders, C. Ruth, O. Matsushita and J. Sakon, *Ca(2+) -induced orientation of tandem collagen binding domains from clostridial collagenase colg permits two opposing functions of collagen fibril formation and retardation*, Febs j **285** (2018), no. 17, 3254-3269.

20. N. Ohbayashi, T. Matsumoto, H. Shima, M. Goto, K. Watanabe, A. Yamano, Y. Katoh, K. Igarashi, Y. Yamagata and K. Murayama, *Solution structure of clostridial collagenase h and its calcium-dependent global conformation change*, Biophysical journal **104** (2013), no. 7, 1538-1545.

21. K. N. Dyer, M. Hammel, R. P. Rambo, S. E. Tsutakawa, I. Rodic, S. Classen, J. A. Tainer and G. L. Hura, *High-throughput saxs for the characterization of biomolecules in solution: A practical approach*, Methods in molecular biology (Clifton, N.J.) **1091** (2014), 245-258.

22. S. Skou, R. E. Gillilan and N. Ando, *Synchrotron-based small-angle x-ray scattering of proteins in solution*, Nature protocols **9** (2014), no. 7, 1727-1739.

23. L. A. Feigin and D. I. Svergun, *Structural analysis by small angle x-ray and neutron scattering*, Plenum Press1987.

24. C. D. Putnam, M. Hammel, G. L. Hura and J. A. Tainer, *X-ray solution scattering (saxs) combined with crystallography and computation: Defining accurate macromolecular structures, conformations and assemblies in solution*, Q Rev Biophys **40** (2007), no. 3, 191-285.

25. D. I. Svergun, *Determination of the regularization parameter in indirect-transform methods using perceptual criteria*, Journal of Applied Crystallography **25** (1992), no. 4, 495-503.

26. R. P. Rambo and J. A. Tainer, *Characterizing flexible and intrinsically unstructured biological macromolecules by sas using the porod-debye law*, Biopolymers **95** (2011), no. 8, 559-571.

27. V. Volkov and D. Svergun, *Uniqueness of ab initio shape determination in small-angle scattering*, Journal of Applied Crystallography - J APPL CRYST **36** (2003), 860-864.

28. D. Franke and D. I. Svergun, *Dammif, a program for rapid ab-initio shape determination in small-angle scattering*, J Appl Crystallogr **42** (2009), no. Pt 2, 342-346.

29. E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng and T. E. Ferrin, *Ucsf chimera--a visualization system for exploratory research and analysis*, J Comput Chem **25** (2004), no. 13, 1605-1612.

30. S. T. L. Philominathan, T. Koide, K. Hamada, H. Yasui, S. Seifert, O. Matsushita and J. Sakon, *Unidirectional binding of clostridial collagenase to triple helical substrates*, Journal of Biological Chemistry **284** (2009), no. 16, 10868-10876.

31. S. T. L. Philominathan, T. Koide, O. Matsushita and J. Sakon, *Bacterial collagen-binding domain targets undertwisted regions of collagen*, Protein Science **21** (2012), no. 10, 1554-1565.

32. D. I. Svergun, *Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing*, Biophysical journal **76** (1999), no. 6, 2879-2886.

33. R. Masuda, M. Kudo, Y. Dazai, T. Mima and T. Koide, *Collagen-like antimicrobial peptides*, Biopolymers **106** (2016), no. 4, 453-459.

34. R. Bauer, J. Wilson, S. T. L. Philominathan, D. Davis, O. Matsushita and J. Sakon, *Structural comparison of colh and colg collagen-binding domains from clostridium histolyticum*, Journal of Bacteriology **195** (2013), no. 2, 318-327.

35. E. Schönauer, A. M. Kany, J. Haupenthal, K. Hüsecken, I. J. Hoppe, K. Voos, S. Yahiaoui, B. Elsässer, C. Ducho, H. Brandstetter and R. W. Hartmann, *Discovery of a potent inhibitor class with high selectivity toward clostridial collagenases*, Journal of the American Chemical Society **139** (2017), no. 36, 12696-12703.

36. U. Eckhard, D. Nuss, P. Ducka, E. Schonauer and H. Brandstetter, *Crystallization and preliminary x-ray characterization of the catalytic domain of collagenase g from clostridium histolyticum*, Acta Crystallogr Sect F Struct Biol Cryst Commun **64** (2008), no. Pt 5, 419-421.

37. U. Eckhard, E. Schonauer and H. Brandstetter, *Structural basis for activity regulation and substrate preference of clostridial collagenases g, h, and t*, J Biol Chem **288** (2013), no. 28, 20184-20194.

38. K. Kar, S. Ibrar, V. Nanda, T. M. Getz, S. P. Kunapuli and B. Brodsky, *Aromatic interactions promote self-association of collagen triple-helical peptides to higher-order structures*, Biochemistry **48** (2009), no. 33, 7959-7968.

39. M. Nishio, *The ch/π hydrogen bond in chemistry. Conformation, supramolecules, optical resolution and interactions involving carbohydrates*, Physical Chemistry Chemical Physics **13** (2011), no. 31, 13873-13900.

40. K. Yoshihara, O. Matsushita, J. Minami and A. Okabe, *Cloning and nucleotide sequence analysis of the colh gene from clostridium histolyticum encoding a collagenase and a gelatinase*, J Bacteriol **176** (1994), no. 21, 6489-6496.

41. J. P. R. O. Orgel, T. C. Irving, A. Miller and T. J. Wess, *Microfibrillar structure of type i collagen in situ*, Proceedings of the National Academy of Sciences of the United States of America **103** (2006), no. 24, 9001-9005.

42. X.-L. Chen, B.-B. Xie, F. Bian, G.-Y. Zhao, H.-L. Zhao, H.-L. He, B.-C. Zhou and Y.-Z. Zhang, *Ecological function of myroilysin, a novel bacterial m12 metalloprotease with elastinolytic activity and a synergistic role in collagen hydrolysis, in biodegradation of deep-sea high-molecular-weight organic nitrogen*, Applied and environmental microbiology **75** (2009), no. 7, 1838-1844.

43. J. Parvizi and G. K. Kim, "Chapter 53 - collagen," *High yield orthopaedics*, J. Parvizi and G. K. Kim (Editors), W.B. Saunders, Philadelphia, 2010, pp. 107-109.

44. T. Watanabe-Nakayama, M. Itami, N. Kodera, T. Ando and H. Konno, *High-speed atomic force microscopy reveals strongly polarized movement of clostridial collagenase along collagen fibrils*, Sci Rep **6** (2016), 28975.

45. M. F. French, K. A. Mookhtiar and H. E. Van Wart, *Limited proteolysis of type i collagen at hyperreactive sites by class i and ii clostridium histolyticum collagenases: Complementary digestion patterns*, Biochemistry **26** (1987), no. 3, 681-687.

46. K. A. Mookhtiar and H. E. Van Wart, *Clostridium histolyticum collagenases: A new look at some old enzymes*, Matrix Suppl **1** (1992), 116-126.

47. M. F. French, A. Bhown and H. E. Van Wart, *Identification of clostridium histolyticum collagenase hyperreactive sites in type i, ii, and iii collagens: Lack of correlation with local triple helical stability*, J Protein Chem **11** (1992), no. 1, 83-97.

48. W. Saito, K. Uchida, O. Matsushita, G. Inoue, H. Sekiguchi, J. Aikawa, H. Fujimaki and M. Takaso, *Acceleration of callus formation during fracture healing using basic fibroblast growth factor-kidney disease domain-collagen-binding domain fusion protein combined with allogenic demineralized bone powder*, Journal of Orthopaedic Surgery and Research **10** (2015), 6.

49. K. Uchida, O. Matsushita, N. Nishi, G. Inoue, K. Horikawa and M. Takaso, *Enhancement of periosteal bone formation by basic fibroblast-derived growth factor containing polycystic kidney disease and collagen-binding domains from clostridium histolyticum collagenase*, J Tissue Eng Regen Med **11** (2017), no. 4, 1165-1172.

50. K. Uchida, G. Inoue, O. Matsushita, K. Horikawa, H. Sekiguchi, W. Saito, S. Takano, H. Fujimaki, M. Miyagi and M. Takaso, *Basic fibroblast growth factor-anchored multilayered mesenchymal cell sheets accelerate periosteal bone formation*, Biomed Res Int **2017** (2017), 4371460.

51. R. Katikaneni, T. Ponnapakkam, O. Matsushita, J. Sakon and R. Gensure, *Treatment and prevention of chemotherapy-induced alopecia with pth-cbd, a collagen-targeted parathyroid hormone analog, in a non-depilated mouse model*, Anti-Cancer Drugs **25** (2014), no. 1, 30-38.

52. T. Ponnapakkam, R. Katikaneni, H. Suda, S. Miyata, O. Matsushita, J. Sakon and R. C. Gensure, *A single injection of the anabolic bone agent, parathyroid hormone-collagen binding domain (pth-cbd), results in sustained increases in bone mineral density for up to 12 months in normal female mice*, Calcified Tissue International **91** (2012), no. 3, 196-203.

53. R. Katikaneni, T. Ponnapakkam, H. Suda, S. Miyata, J. Sakon, O. Matsushita and R. C. Gensure, *Treatment for chemotherapy-induced alopecia in mice using parathyroid hormone agonists and antagonists linked to a collagen binding domain*, Int J Cancer **131** (2012), no. 5, E813-821.

54. T. Ponnapakkam, R. Katikaneni, E. Miller, A. Ponnapakkam, S. Hirofumi, S. Miyata, L. J. Suva, J. Sakon, O. Matsushita and R. C. Gensure, *Monthly administration of a novel pth-collagen binding domain fusion protein is anabolic in mice*, Calcif Tissue Int **88** (2011), no. 6, 511-520.

55. R. Stratford, Jr., C. Vu, J. Sakon, R. Katikaneni, R. Gensure and T. Ponnapakkam, *Pharmacokinetics in rats of a long-acting human parathyroid hormone-collagen binding domain peptide construct*, J Pharm Sci **103** (2014), no. 2, 768-775.

Chapter 4: Working hypothesis of Col G/Col H collagenolytic synergy for purposes of Cell isolation

Abstract

While it is known that both Col G and Col H are needed to most effectively degrade collagen fiber, the optimal ratio between the two is not agreed upon. This chapter focuses on the degradation of type I collagen fiber from Col G and Col H. In order to isolate different cells type I collagen fibers, which are abundant in all tissues, would need to be degraded [**5**]. This chapter seeks to understand the steps that *H. histolytica* takes to degrade each level of hierarchical collagen and expose micro-fibrils. Based on time course hydrolysis of acid solubilized bovine type I collagen a method of synergistic hydrolysis of collagen micro-fibril by Col G and Col H was speculated. Optimal ratio between Col G and Col H can improve cell isolation yields for transplant.

4.1 Introduction

4.1a. Clinical applications for Col G/Col H mixtures

The ability of full-length Col G and Col H to seek out and synergistically degrade collagen fibers makes Col G and Col H molecules of interest for clinical research purposes. Santyl ointment is a mixture of Col G and Col H that is used as a wound debridement and they are also used in the treatment diseases characterized by a buildup collagen fibrils (Dupuytren's contracture and Peyronie's disease) [**6**]. Pancreatic islet transplantation was mentioned earlier in the chapter. It is an experimental treatment for type I diabetes and pancreatic cancer where pancreatic islets are isolated from the patient pancreas or from a donor using a mixture of Col G, Col H and neutral protease (normally thermolysin) (see figure 1). Newly isolated islets are

cultured and then injected into the patients liver where pancreatic function is restored [**7**]. Minimizing the amount of neutral protease used in isolation is important as excess limits islet survival and viability *in vitro* [**8**]. In order to minimize the amount of protease used the optimal ratio of Col G to Col H needed for optimizing the islet yield needs to be found. Currently different ratios of Col G to Col H are being investigated on their cell isolation ability but a consensus has yet to be reached [**9**]. Based on what is known about Col G and Col H collagen hydrolysis, in this chapter the optimal ratio of Col G to Col H needed for hydrolysis of fibrillar collagen is proposed [**4**] [**10**].

In addition to pancreatic islet isolation, collagenases can be used for the isolation of stem cells from adipose fat tissue. Recently a method was proposed by Gentile and Sterodimas for the isolation of stem cells from adipose fat tissue using human MMP for purposes of reversing Covid-19 lung damage [**11**]. Mesenchymal stem cell isolation process may be optimized by using bacterial collagenases instead of human MMP. Stem cells may also be isolated from synovial fluid [**12**]. In addition to treating Covid-19, collagenases could be used to isolate skeletal muscle stem cells for the purposes of treating muscular dystrophy and age associated muscle atrophy [**13**]. Knowing the optimal ratio of Col G to Col H necessary for collagen fiber degradation would improve the many clinical applications for *H. histolytica* collagenases. Below a method for unbundling collagen fibers by *H. histolytica* is proposed based on the known structures and collagen hydrolysis patterns for Col G and Col H.

4.1b. Proposed method for Hathewaya histolytica unbundling collagen fibers

Collagen fiber are the largest ordered structure of tropocollagen molecules possible. Collagen fibers within the interstitial matrix give structural support to cells and keep them tethered to basement membranes from different organs. For Col G/Col H mixtures to be effective

at cell isolation and collagen degradation in treating Dupuytren's contracture and Peyronie's disease collagen fibers must be unbundled [**6**] [**14**] [**5**]. For a brief reminder on type I collagen hierarchy, three alpha propeptide chains (two α1 and one α2) intertwine to form triple helical tropocollagen. Orgel et al. 2006 posits a micro-fibril structures where five to six tropocollagens molecules are packed together in a super-twisted, right-handed quasi-hexagonal pattern called a micro-fibril (see figure 2). As tropocollagens are twisted together individual tropocollagens are longitudinally staggered throughout the micro-fibril by 67 nm to create D-period gaps [**15**].

Tropocollagens within a micro-fibril are both inter and intra molecularly cross-linked at Lys residues located at the N and C-telopeptide regions by Lysyl oxidase (LOX), cross-links provide additional strength to tropocollagen and micro-fibril [**16**]. In order to hydrolyze tropocollagens within a micro-fibril telopeptide region must be cleaved off. Intermolecular cross-linking at the telopeptides is known to play a key role in collagen fibril formation [**17**]. In order to form collagen fibril from micro-fibril LOX catalyzes cross-linking at Lys residues from tropocollagen molecules on the exterior of the micro-fibril [**18**]. Finally, collagen fibrils are linked together by proteoglycan bridges to form collagen fibers (see figure 3) [**19**]. While proteoglycan bridges are hydrolyzed in acid solubilized collagen for MMPs proteoglycan bridges prevent the enzyme from cleaving collagen fibers [**20**]. It may be the same for bacterial collagenases. Below a brief description on the mechanism by which proteoglycan bridges promote collagen fiber formation is described.

Small leucine-rich repeat proteoglycans (SLRPs) are a specific type of proteoglycan that is thought to promote fiber formation. In SLRP knockouts mice, collagen fibrils are highly disorganized, and some connective tissue function is lost [**21**]. SLRPs are two part constructs with one N-terminal variable domain containing sulfonated Tyr or acidic amino acids and one

core protein domain that is rich with Leu repeats [**21**]. The core protein domain of decorins bind to D-period gaps in collagen via covalent interactions most likely to 3HYP repeats found at the C-terminus of type I tropocollagen [**22**] [**2**] [**23**]. Glycosaminoglycans (GAG) chains are present on the N-terminal domain of SLRPs and affect collagen fiber formation by binding to non-fibril collagens, FACIT collagens, that exist in the collagen matrix between fibrils [**21**]. They also interact with GAG from adjacent collagen fibrils via electrostatic interactions [**24**]. This ability to bridge multiple collagen fibrils through FACIT collagens and electrostatic interactions results in the association of several collagen fibrils forming fibers.

It has previously been shown that a crude mixture of Col G and Col H has the ability to unbundle collagen fibers from sea cucumber [**24**]. Collagenase type I from Worthington biochemical was used during this experiment. Type I from Worthington is a mixture of Col G and Col H and contains trace amounts caseinase, clostripain and trypsin. TEM imaging revealed that collagen treated with this mixture began to unbundle and proteoglycans connecting fibrils were lost [**24**]. SDS-PAGE analysis of collagen fiber incubated with this crude mixture for various intervals was used to visualize the hydrolysis of proteoglycans over time. SDS-PAGE gel was stained with alcian blue-periodic acid stiff, which is used to stain sugars, revealed the presence of a diffuse sugar band appearing from 87 kDa to 55 kDa. As the incubation time increased this band became denser and wider, implying that over time more GAG is released from solubilized collagen fiber [**24**]. In addition to SDS-PAGE, DMNB assay also revealed that the amount of GAG present in the supernatant of samples treated with Col G/Col H mixture increases over time [**24**].

Despite these results, it is currently unknown if the collagenases Col G or Col H can hydrolyze the proteoglycan bridge or if the added caseinase, clostripain and trypsin do the

hydrolysis. Clostripain and trypsin have been previously shown to efficiently hydrolyze choindroitin sulfate aggregates from rats [**25**]. Currently it is unknown if caseinase has any activity towards proteoglycans involved in fiber formation. We speculated that clostripain is primarily responsible for degradation of proteoglycan bridge. Clostripain is a cysteine protease with a preference for hydrolysis on the C-terminal side of Arg and Lys [**26**]. Arg rich SLRPs have been found in a variety of collagens [**27**]. The protease was initially isolated from *C. histolyticum* (now *H. histolyticum*) and found in all *Clostridium* species. Core protein domain for SLRPs are resistant to degradation from most enzymes, including clostripain, suggesting that if cleavage is to occur it must be cleaved off at the variable N-terminal domain [**28**]. For *Clostridium* species clostripain's role in virulence is not entirely understood. In *C. perfringens* α-clostripain structural gene ccp was mutationally inactivated. In mice models of myonecrosis, inactivation of the ccp gene was found to have no affect on the progression of the disease indicating that clostripain may not play a major role in the progression of gas gangrene but it may be one of many enzymes that assist in initially unbundling collagen fibers for Col G and Col H to attack [**29**].

4.1c. Col G may be responsible for dismantling Extrafibrillar Space between collagen fibers

Once proteoglycan bridges are dismantled collagen fibers are only partially exposed. However, in the interstitial space, where Col G and Col H primarily work, there exists a complex connective matrix between collagen fibers. This region must be properly degraded before collagen fibrils are fully exposed. This extrafibrillar space consists of polysaccharides, proteoglycans and auxiliary collagens that are necessary for the formation of collagen fibers (see figure 4) [**30**]. The tandem CBDs of Col G bare larger bore size and may make uniquely suited towards cleaving some of auxiliary, non-fibril collagens present in the extrafibrillar matrix that

hold collagen fibrils together to form fibers [**31**]. Below we speculate how Col G dismantles this region

Type V collagen is a key component of the extrafibrillar matrix and associates with type I collagen to form heterotypic fibrils [**2**]. It has three different α-chains α1, α2 and α3 that come together in different combinations with $(\alpha1)_2\alpha2$ being the most common for forming heterotypic fibrils [**2**]. This isoform has the normal triple helical regions of collagen however, at the N-terminus there is a propeptide domain that is important in interacting with the extrafibrillar matrix. In heterotypic collagen fibrils, type V collagen is in the interior and due to the presence of a kink region between the N-terminal propeptide and the triple helical region it is thought that the propeptide is projected through the D-period gap region [**2**] [**32**]. Once the propeptide is projected through the D-period gap it is able to interact with different non-fibrillar (FACIT) collagens in extrafibrillar space (see figure 5) [**32**] [**33**]. Immunoelectron microscopy has been able to reveal the presence of type VI collagen, a FACIT collagen, in the banded regions of collagen fibers from the extracellular matrix of chicken embryo and human fibroblast [**34**]. This interaction with the extrafibrillar space is thought to be important for fiber formation; in mice with gene for type V collagen knocked out fiber formation was significantly inhibited [**35**].

In addition to fiber formation, Type V collagen is important for connecting cells to basement membranes. Type V collagen N-terminal propeptide interacts with type VI collagen which then interacts with type IV collagen, a key component of basement membranes [**2**] [**30**]. This ability to act as a sort of "bridge" between heterotypic collagen fibrils and basement membranes makes type V collagen very important in anchoring cells to basement membranes. SDS-PAGE analysis and mass spectrometry have revealed that type V collagen is a substrate for

Col G but not for Col H [**31**]. Unlike digestion of type I collagen which occurs from the C-terminus; digestion of type V collagen was found to occur near the N-terminus [**36**] [**37**] [**31**].

This ability to hydrolyze type V collagen near the N-terminus may hint towards Col G's role in cell isolation. Col G's dual CBDs may allow the enzyme to wedge between a heterotypic collagen fibril containing type V collagen and the type IV/VI basement membrane collagens it is connected to. While both type IV and type V collagens are substrates for Col G it is currently unknown if type VI collagen is for Col G [**38**] [**31**]. Type VI collagen contains triple helical regions similar to fibril collagens suggesting that it to can be processed by *H. histolytica* collagenases [**39**]. If type VI collagens are resistant to Col G, the enzyme still may be able to bind to heterotypic collagen fibril in ECM and position its catalytic domain to optimally process the N-terminal propeptide from type V collagen. In heterotypic collagen fibrils the only portions of type V collagen that would be available for processing are the N-terminal propeptide and the hinge region. Once this bridge between heterotypic collagen fibrils and tissue basement membranes is cut isolation could be performed more easily. Below a method for collagenolytic synergy between Col G and Col H is proposed.

4.2 Methods

4.2a. Time Dependent Hydrolysis of Bovine Type I Collagen by Col G and Col H

Experiments regarding time dependent hydrolysis of Bovine Type I collagen were performed by Dr. Keisuke Tanaka at the Nippi Research Institute of Biomatrix in Toride, Japan. Analysis of data was done at the University of Arkansas. To prevent formation of gelatin all steps were performed at 4°C unless stated otherwise. Briefly, 12.5 µl of acid solubilized bovine type I collagen (2 mg/ml) was mixed with 50 µl of a 2X collagenase buffer (100 mM Bis-Tris-HCl pH 7.5, 0.4 M NaCl, 10 mM $CaCl_2$), 36.5 µl of MilliQ water and 1 µl of Col G or Col H (0.1

mg/ml). Final concentration of collagen and collagenase was 0.25 mg/ml and 0.001 mg/ml respectively. Mixture was incubated at 30°C and samples were collected at 0, 0.5, 1, 3, 5 and 21 hrs. respectively. The reaction for each sample aliquot was stopped by adding 15 µl of 2X SDS PAGE sample buffer. After all samples were collected, ther were boiled at 100°C for 3 min and 30 µl of each sample were applied to a 4-20% gradient Tris-Glycine gel and gel electrophoresis was performed.

Gels were stained with comassie blue and destained with 10% Acetic acid/50% Methanol for 24 hrs. After destaining, gels were scanned and ImageJ software was used to perform densitometry analysis for Col G and Col H [**40**]. Greyscale images were used and background was subtracted (light background, sliding parabola and disable smoothing). Rectangles were drawn for each well to include γ, β11 variant, β11, β12, α1, α2 and any potential fragments from hydrolysis [**41**]. Densitometry peaks were integrated. To help reveal how Col G and Col H process different regions in collagen, the intensity for each band was plotted over the course of 21 hrs. Relative mobility of each band relative to protein ladder was used to help determine the molecular weight for each band.

4.3 Results and Discussion

4.3a. Hydrolysis of γ and β collagens by Col G and Col H gives an initial increase in α1 and α2 chains

Fragments resulting from hydrolysis of acid solubilized bovine type I collagen from Col G and Col H over 21 hours were separated by SDS-PAGE and analyzed. Since the bands for β11 and β12 could not be separated, intensity was quantified for a single β11/β12 band. For Col G and Col H bands corresponding to γ, β11 variant, β11/β12, α1 and α2 as well as any collagen fragments were assigned based on past study (see figure 6A & 7A) [**4**]. Cross-link of two α1

chains forms β11 collagen and cross-link of α1 chain with α2 chains forms β12 collagen and cross-link of three α chains (commonly α1α1α2) forms γ collagen [**41**]. Background from gels were subtracted using imageJ and rectangles were drawn around each lane to quantify band intensities. For Col G and Col H total intensity for each band was plotted over time (see figure 6B & 7B).

β11 variant intensity appears to be negligible for both Col G and Col H. For both Col G and Col H, the intensities for the γ and β11/β12 bands decrease over the course of the experiments while the intensities for α1 and α2 collagens increase during the first 30 minutes then decrease over the rest of the experiment. For Col G at 3 hrs. bands were almost fully hydrolyzed and for Col H bands were still present after 21 hrs. For Col G, as the hydrolysis experiment proceeded the ratio of α1 band intensity to α2 band intensity changes from 2:1 to closer to 1:1 (2.30:1 at 0 hr., 1.15:1 at 0.5 hr., 1.37:1 at 1 hr.). For Col H, the ratio of α1 to α2 stays closer to 2:1 (2.14:1 at 0 hr., 1.76:1 at 0.5 hr., 1.60:1 at 1 hr., 1.49:1 at 3 hr., 1.83:1 at 5 hr.) (see figure 6B & 7B). For Col G, only one additional fragment was present outside of the normal six isoforms for type I collagen (see figure 6A) [**41**]. This additional fragment appeared after 30 min. and quickly disappeared at 1 hr. Due to its low intensity and short lifespan it is not included in plot for hydrolysis of acid solubilized collagen by Col G. For Col H six fragments emerged at different times after hydrolysis. Molecular weight of Col H fragments were calculated by using relative mobility of protein ladder to create a calibration plot (see figure 7C) [**42**]. What the the appearance of these bands mean for a potential collagen hydrolysis mechanism for Col H is discussed further on in the chapter.

4.3b. Col G collagenolysis starts at C-terminus and ends with hydrolysis of both N & C telopeptides

To understand the increase in the release of α2 chains the formation of tropocollagens and their packing within a micro-fibril must be considered. While acid solubilized type I collagen is not identical to collagen micro-fibril hydrolysis of solubilized substrate may give hint to how *H. histolytica* collagenases degrade insoluble collagen substrate. During collagen fibril formation tropocollagens are packed in parallel and staggered by a repeat distance of 67 nm known as the D-period [**16**]. Inter and Intra molecular covalent crosslinks located at regions near the N and C-terminal known as telopeptides (see figure 3B) [**16**]. At the C-telopeptide Lys residues from both α1 chains are intermolecularly crosslinked with α1 chain Lys in the triple helical region of an adjacent tropocollagen. There are no Lys residues in the C-telopeptide of the α2 chain [**16**]. While at the N-telopeptide Lys from one α1 chain intermolecularly crosslinks with Lys in the triple helical region of an adjacent tropocollagen [**17**]. Lys on the other α1 chain intramolecularly crosslinks with Lys in the N-telopeptide of the α2 chain [**17**].

N-telopeptide is thought to be the main region in collagen involving α2 cross-linking [**17**] [**16**]. Initial increase in intensity for α2 band and decrease in intensity for γ, β11 and β12 bands must be from hydrolysis of N-telopeptide. For the amount of α2 chain to be increased in the SDS-PAGE during the first 30 mins. Col G must hydrolyze the N-telopeptide. Hydrolysis would release α2 chain as the intramolecular crosslink between Lys from α1 and Lys from α2 would be degraded and release of α1 chains would be slowed due to the presence of intermolecular crosslinks at the C-telopeptide region. Previously it has been shown that Col G cannot bind to telopeptide regions but it is still able to have the region efficiently released [**43**] [**44**]. Based on previous results it is believed that Col G is also involved in the hydrolysis of the C-telopeptide

[**44**]. However, the intensity increase in α1 from hydrolysis would be difficult to pick up due to α1 being cross-linked at both the C and N-telopeptides.

SAXS and NMR data show that CBD2 from Col G prefers to bind to loosely wound regions of tropocollagen and will bind so in a unidirectional manner [**45**] [**37**]. Previous high speed AFM video of Col G revealed that the enzyme flocks to disordered regions in collagen micro-fibril mimic [**36**]. MMP binding site in collagen is roughly one quarter distance from the C-terminus of collagen [**46**]. Low hydroxyproline content near the C-terminus suggest that this region is unstable as it has been reported to locally "melt" [**47**]. Similar to its preference for disrupted regions in collagen micro-ribbon Col G may prefer the partially unwound region at the C-terminus since less energy would be needed for unwinding substrate before hydrolysis [**45**]. While processing collagen; polarized movement occurs as Col G moves from the C-terminus to the N-terminus. Saddle like shape of catalytic domain active site, tight binding of CBD2 and lower collagen affinity of CBD1 allows Col G to processively degrade substrate [**1**]. A mechanism for this processivity was previously proposed and below we attempt to reconcile this mechanism with known information on Col G and collagen micro-fibril structure with new information pertaining to the enzymes ability to hydrolyze N-telopeptide regions in acid-solubilized collagen [**1**]. Using all of this information we attempt to tease out how Col G hydrolyzes collagen micro-fibril *in vivo*.

Col G would still attack disordered regions in the micro-fibril (either from damage or at the C-terminus) of the collagen and wedge between two parallel oriented tropocollagens. Once wedged the enzyme would proceed to move towards the N-terminus of one tropocollagen molecule while processively degrading substrate [**36**]. C-terminus of tropocollagen CBD2 is bound to is staggered, with the N-terminus of an adjacent tropocollagen molecule overlapping

(see figure 3B). This overlapping from the D-periodicity of the micro-fibril would position the enzyme so that it may effectively cleave off the N-telopeptide from the adjacent tropocollagen.

Proteolysis experiments on procollagen from human skin revealed that bacterial collagenase cleave at the border between telopeptide and triple helical regions of collagen and at the border between procollagen and telopeptide for both C and N-telopeptides [**44**]. Amino propeptide is thought to fold over on the triple helical portion of collagen increasing the stability of the tropocollagen chain [**44**]. Hydrolysis of the propeptide/telopeptide region by Col G decreases the stability of the tropocollagen which may make it more susceptible to hydrolysis by Col H. Loss of telopeptide would loosen the collagen fibril which in turn would make it easier for Col H to bind and hydrolyze at its specific sites [**4**] [**10**].

4.3c. Non-processive endopeptidase, Col H hydrolyzes specific regions in acid solubilized collagen

Hydrolysis of acid solubilized type I collagen with Col H revealed six additional fragments in addition to the six main isoforms of acid solubilized type I collagen (γ, β11 variant, β11, β12, α1 and α2) (see figure 7A). Understanding the Col H hydrolysis pattern in type I collagen would help to reveal the role it plays in synergistic degradation of collagen. Molecular weights of all Col H bands are estimated using relative mobility of SDS-PAGE bands (see figure &C) [**41**]. After 0.5 hr. fragments 1 (approximately 110 kDa), 2 (approximately 95 kDa), 3 (approximately 83 kDa) and 4 (approximately 53 kDa) appear. After 3 hrs. fragments 5 (approximately 61 kDa) and 6 (approximately 42 kDa). Over the full 21 hrs. intensity for fragment 1 stays consistently low and that intensity for fragments 2 and 3 increase up until 5 hrs. after which they begin to decrease. Intensities for fragments 5 and 6 both increase for the full 21 hrs. of the experiment and the band for fragment 6 appears to shift from approximately 42 kDa to

37 kDa. Fragment 5 plateaus at 800 units and fragment 6 increases drastically up to 4200 units (see figure 7B).

Results from acid solubilized collagen hydrolysis with Col H appear to agree closely with acid solubilized collagen hydrolysis results from French et al. 1987. In results for French et al. 1987 intensity for γ bands is not given [**4**]. It is unknown why γ band intensities are not given. Based on the ordered decrease in intensity for γ, β11/β12, α1 and α2 bands in acid solubilized type I collagen Col H does not hydrolyze telopeptides before initially attacking three quarters region near the N-terminus of collagen. After initial hydrolysis 35 kDa N-terminal fragment and 62 kDa C-terminal fragment are produced. In results from French et al. 1987 N-terminal fragment forms a 70 kDa dimer (F'1/F'2) [**4**]. Assuming that bands from the latest results are running at larger molecular weights we speculate that fragment 2 (approximately 95 kDa) from bovine hydrolysis is equivalent to F'1/F'2 from rat tail collagen hydrolysis. In acid solubilized collagen intermolecular cross-links are most likely hydrolyzed although based on some intermolecular cross-linking may still be taking place. In acid solubilized collagen, aggregates could also be either non-staggered or staggered by a 4D distance [**48**]. Intramolecular cross-linking between α1 and α2 chains at the N-telopeptide would explain the presence of N-terminal 35 kDa dimer [**16**]. Based on the molecular weights and rate at which intensities increase for fragments 3 and 4 these two bands may be analogous to 62 kDa C-terminal fragment (F'3/F'5) and 35 kDa N-terminal fragment (F'7/F'8) monomers from the initial hydrolysis with Col H [**4**].

The approximately 100 kDa fragment 1 from the latest bovine type I collagen hydrolysis results is not present in the experiment from French et al. 1987 [**4**]. Fragment 1 could be from hydrolysis of 4D staggered acid solubilized substrate. C-terminal 62 kDa fragment could cross-link with Lys residues in triple helical region of N-terminal 35 kDa fragment in adjacent

tropocollagen and the resulting product would be approximately 100 kDa [**48**]. Fragment 1 could also be from both inter and intramolecular cross-linking present in non-staggered N-terminal fragment from initial hydrolysis with Col H [**4**] [**48**]. More research will need to be performed to determine the identity of fragment 1. At 3 hrs. the appearance of fragments 5 (approximately 61 kDa) may be from further processing of the 62 kDa C-terminal fragment and is analogous to (F'9/F'10) from French et al. 1987 [**4**]. Fragment 6 (approximately 42 kDa) is not present in results from French et al. 1987. However, it the French group later discovered that N-terminal fragment from initial Col H hydrolysis is further processed and Fragment may be from further processing of N-terminal fragments [**10**].

In rat type I collagen Col H initially hydrolyzes the bond between Arg396 and Gly397 across all α-chains to form the 35 kDa and 62 kDa fragments [**10**]. 62 kDa fragment is next hydrolyzed near the newly formed N-terminus at the bond between Hyp405 and Gly406 to form a 58 kDa fragment [**10**]. 58 kDa fragment is then processed at two sites closer to the C-terminus to give 36 kDa, 32 kDa and 26 kDa fragments [**10**]. Hydrolysis for bovine type I collagen and rat type I collagen are expected to be similar and indeed they were. Col H hydrolysis strategy is thought to be conserved across fibril collagens based on the presence of conserved hyperreactive sites in all type I, II and III collagens [**10**]. 36 kDa, 35 kDa, 32 kDa and 30 kDa may be too small to be efficiently separated using gel electrophoresis. In addition to the F'9, F'10 fragment 6 may be a combination of the 36 kDa, 35 kDa, 32 kDa and 30 kDa fragments which would explain the significant increase in intensity over time (see figure 7B).

While comparison of acid solubilized bovine type I collagen hydrolysis and rat tail type I collagen hydrolysis revealed that Col H mechanism for acid solubilized collagen hydrolysis is conserved. It is currently unknown how Col H hydrolyzes collagen micro-fibrils. Based on acid

solubilized collagen hydrolysis results as well as structure/function results we attempt to tease out a hydrolysis mechanism. The size of envelopes for Col H segments complexed to mini-collagen as well as the orientation of the domains in the full-length enzyme suggests that Col H may be best suited for wedging between tropocollagens within a micro-fibril. PKD2 is most likely used similar to CBD1 in Col G. Col H may scan the surface of collagen micro-fibrils for loosely wound regions to wedge between. Once found the 15 Å distance between PKD2 and CBD may position Col H to clip tropocollagens from within a micro-fibril. Micro-fibril binding results support this ability as addition of N-terminal domains causes an increase in binding affinity consistent with wedging [**49**]. Unpublished high-speed AFM video of free Col H reveals that the enzyme is highly dynamic which may make it easier for the enzyme to intercalate within the interior of a micro-fibril (Appendix video 2). Once bound Col H would then be able to cleave all three chains in tropocollagen. Hydrolysis of tropocollagen at its interior would create a potential new site for Col G to begin hydrolysis. The slow speed for Col H hydrolysis of acid solubilized type I collagen with suggests that not all hyperreactive sites for the enzyme are immediately available for hydrolysis. Hydrolysis with Col G may be required first to loosen the collagen fibril. Synergy between Col G and Col H is discussed further in the next section.

4.3d. Description of collagen degradation synergy between Col G and Col H

Studies from Orgel et al. on the structure of type I collagen fibrils revealed that tropocollagens are arranged to form a supertwisted discontinuous right-handed micro-fibril [**15**]. Tropocollagens within micro-fibril are stabilized through intramolecular and intermolecular cross-links at both the N and C-telopeptides (see figure 3B) [**15**]. Okuyama group disagrees with Orgel fiber diffraction structure based on low resolution, number of reflections and occupancy and they suggest that position of chains remains speculative [**50**]. As it is still the most accepted

model for collagen micro-fibril all proposed methods for collagen hydrolysis are based on the Orgel model. For solubilized collagen, Col G and Col H have the ability to degrade substrate alone however for insoluble collagen micro-fibril perhaps both enzymes are required. Currently the mechanism of synergy between Col G and Col H is currently unknown. Depending on the target cell being isolated, different mixtures of Col G and Col H may be required. Previous research into the isolation and recovery of pancreatic islets show that poorly isolated cells have shorter survival periods post culture and are less likely to survive upon transplant into the recipient's liver [**51**]. Understanding the mechanism by which the enzymes work together to degrade collagen may reveal the most optimal ratio of Col G to Col H necessary for the isolation of cells. Below a method of synergy between Col G and Col H is proposed based on current and previous research results.

In a micro-fibril, due to the quasi-hexagonal packing of tropocollagens in the micro-fibril Col H proteolysis sites on the exterior tropocollagens are more easily accessible for hydrolysis than those in the interior. In a collagen fibril, multiple micro-fibrils are cross-linked, so the initial proteolysis sites on tropocollagens within a micro-fibril are thought to be buried. In addition to Col H proteolysis sites binding sequences for MMP1, MMP13, von Willebrand factor, glycoprotein VI, integrins and fibromodulin are also predicted to be at least partially buried within the core of a collagen fibril [**52**]. Yet, cellular functions regarding these binding sites are still carried out [**52**]. This suggest that collagen micro-fibril be have dynamic character *in vivo*.

Dividing a 300 nm long tropocollagen molecule by the D-period allows one to divide a micro-fibril into five segments (D1-D5) which repeat throughout the fibril [**52**]. To further understand which segments in a collagen fibril are solvent exposed molecular dynamic simulations were performed on previously derived model of collagen fibril with boundaries

placed to replicate the supramolecular arrangement of collagen [**15**] [**53**]. Solvent accessible surface area (SASA) calculations were performed for each segment in a collagen fibril using a 1.4 Å probe and a 5 Å probe. 1.4 Å probe, which is roughly the size of a water molecule, had a high SASA for all segments in the collagen fibril indicating that even the buried tropocollagen molecules are sufficiently hydrated [**52**]. For the 5 Å probe, SASA was measured for both sides of the collagen fibril (A & B) (see figure 8A). For side A SASA for D5 and D4 at the C-terminus were both high but on side B SASA is significantly lower [**52**]. On side A SASA for both D3, D2 and D1 is low but on side B SASA for D3 and D2 increase slightly while D1 increases significantly. For Col H the initial proteolysis site is located at segment D2. Low SASA for D2 at both side A and B suggests that the collagen fibril needs to be initially unwound to expose the proteolysis site for Col H [**52**].

It is believed that Col G is necessary for exposing the main proteolysis site for Col H. Previous NMR and SAXS results revealed that the enzyme binds unidirectionally to the C-terminus of mini-collagen in the absence of undertwisted regions [**37**] [**45**]. D5 segment (C-terminus) of tropocollagen within microfibril is predicted to be packed less tightly on side A [**52**]. Based on B-factor analysis of collagen-like peptide C-terminus of tropocollagen is thought to be less structured [**54**]. Col G will wedge between two adjacent tropocollagens at the D1 segment of a collagen fibril and begin to precessively degrade the substrate while moving towards the N-terminus of the collagen fibril [**36**]. As the enzyme moves towards the N-terminus Col G may work similar to DNA helicase as the collagen fibril becomes slightly unwound. As the fibril becomes slightly unwound proteolysis site for Col H at segment D2 may become more solvent accessible allowing it to be cleaved by Col H. Col H hydrolysis results for acid solubilized bovine and rat tail type I collagen reveals that the enzyme is a non-processive

endoprotease that cleaves only at select sites in the substrate to produce multiple fragments [**4**] [**10**]. After initially cleaving near the N-terminus to produce the 35 kDa and 62 kDa fragments Col H further processes the C-terminal 62 kDa fragment to produce a 59 kDa, 36 kDa, 32 kDa, 30 kDa and 26 kDa fragments [**10**]. N-terminal 35 kDa fragment is also processed further to give a 33 kDa fragment which suggests that the enzyme does not have polarized movement along the collagen fibril and simply cleaves at specific sites [**10**]. This "hit and run" method of hydrolysis for Col H would provide newly exposed regions for Col G to bind to and begin degrading tropocollagens on the exterior of the collagen fibril (see figure 8B). The collagen degrading steps for Col G and Col H are then repeated to hydrolyze the tropocollagens located in the interior of the collagen fibril.

4.3e. Endo/Exo synergy found in cellulose hydrolysis

This method of synergy for Col G and Col H is conserved in a wide variety of enzymes. Synergy between Col G and Col H is similar to endo/exo synergy seen for degradation of a variety of biopolymers such as mucin, chitin and cellulose [**55**] [**56**]. Since cellulose degradation is the most well understand, it's hydrolysis will be described in further detail. However, hydrolysis of mucin and chitin occurs in a similar manner with exoprocessive enzymes hydrolyzing from chain ends and the non-processive endoglucanases hydrolyzing randomly in the interior of the polysaccharide chain.

The main component of cell walls, cellulose is the most abundant polysaccharide in nature and consists of linear chains of multiple glucose units linked through a β-1-4 glycosidic bond [**57**]. Cellulose consists of numerous chains packed together to form fibril like structures similar to collagen which makes both difficult to breakdown. Sugars from the breakdown of cellulose are a potential alternative fuel source so the study of cellulose hydrolysis is of great

importance [**58**]. A group of enzymes including cellobiohydrolases (CBHs), endoglucanases and β-glucosidases work together using endo/exo synergy to efficiently degrade cellulose [**57**]. CBHs are analogous to Col G as they degrade cellulose processively from the chain ends. Endoglucanases are analogous to Col H as they are non-processive and attack at locations in the interior of cellulose chains [**57**]. Product inhibition of CBH is prevented by β-glucosidase hydrolyzing cellobiose into two glucose units [**57**].

Endoglucanase targets amorphous regions of cellulose and which opens up new starting points for CBH. CBH stalls and is released from substrate when reaching these regions [**57**]. Hydrolysis by CBH may in turn reveal new amorphous regions for endoglucanases to target. Similar to endoglucanase, Col H may initially target and cleave amorphous, solvent exposed regions in collagen fibrils. This may be supported by Col H's preference towards collagen like peptides over triple helical collagen [**59**]. In collagen fibril Col G is thought to stall in regions where fibril is disorganized. Small amounts of Col G promote fibril formation so stalling may be due to the enzyme reforming substrate so that hydrolysis can continue [**1**]. Collagen fibril serves as a "railing" for Col G and the same may be true for cellulose fibril with CBH targeting segment. When "railing" is damaged the enzyme can no longer continue on the substrate. Hydrolysis of the damaged "railing" would thus prevent the enzyme from being released. Based on the different domain orientations between bacterial collagenases and cellulases endo/exo synergy shared between both enzymes is the result of convergent evolution.

4.3f. Finding the optimal ratio of Col G to Col H for cell isolation

Finding the optimal ratio of Col G to Col H is important for improving the yield of viable cells during isolation of pancreatic islets and stem cells for purposes of transplantation [**9**] [**7**] [**51**] [**60**]. Pancreatic islet transplantation is an experimental treatment for pancreatic cancer as

well as diabetes while stem cell transplantation may be a potential treatment for a variety of diseases [**7**] [**11**] [**13**]. Auto transplant of cells into a recipient are more likely to succeed as it would prevent an immune response. Previous research into the optimal ratio of Col G to Col H for collagenase synergy revealed that synergy between the two peaks at 25% Col G:75% Col H [**9**]. Synergy was calculated for multiple ratios of Col G to Col H by dividing the collagen degradation activity (CDA) for the Col G/Col H mixture by the sum of the CDA for Col G and Col H individually [**61**] [**9**]. While the two enzymes may be working together the best at this ratio in terms of CDA this ratio was one of the worst. The best ratio for CDA was found to be 90% Col G:10% Col H. This ratio gives the lowest synergy, however this may be due to Col G being the primary workhorse for collagen degradation [**9**]. In one study focusing on the isolation of pancreatic islets from human tissue the best digestion results were obtained with mixture containing the maximum amount of Col G [**62**]. Of the two enzymes Col G must play the primary role in collagen degradation. However, study measuring collagen degradation activity (CDA) of different mixtures of Col G and Col H show that when less Col G is present in the mixture Col H can play a more effective role in collagen degradation [**9**].

Measuring CDA in different mixtures of Col G ΔCBD2 with Col H helped to tease out Col H's ability to degrade collagen. For Col G ΔCBD2 two different forms were used; one where CBD2 was lost by cleavage at Lys896 and one by cleavage Lys908 [**9**]. Previously it was described how CBD2 is necessary for the processivity of Col G [**1**]. Upon loss of CBD2 pancreatic islet isolation by Col G is inhibited [**7**]. For both forms of Col G ΔCBD2:Col H mixtures experimental CDA increases as the amount of Col H is increased with CDA peaking at 25% Col G ΔCBD2:75% Col H [**9**]. For Col G ΔCBD2 cleaved at Lys896 synergy between the two enzymes increases as the amount of Col H increases. However, for Col G ΔCBD2 cleaved at

Lys908 synergy decreases as the amount of Col H increases. The differences for synergy between the two forms of Col G ΔCBD2 are thought to be from differences in purity for the two Col G -ΔCBD2 [**9**].

If the amount of Col G in the collagenase mixture is low Col H may take more of a lead role in hydrolysis. Initial hydrolysis by Col H separates tropocollagen into multiple fragments. Shorter fragments have lower melting temperatures [**63**]. Type I collagen has previously been found to be somewhat thermally unstable at body temperature [**47**]. Col H may hydrolyze collagen into smaller fragments that are more likely to melt away at body temperature disassembling the collagen fibril. This method would not be very affective as the initial proteolysis site for Col H may not always be available which would explain the low CDA value.

Review from McCarthy et al. 2018 states that as long as excess collagenase is present the ratio between Col G and Col H is not important and the type of neutral protease present in the cocktail is what controls the rate of cell isolation [**64**]. Collagenase is proposed to loosen up the ECM revealing protease sensitive sites on ECM proteins key to anchoring cells [**64**]. However, excess neutral protease used for cell isolation has been found to decrease cell survival and viability post isolation and neutral protease by itself has cannot isolate cells from ECM [**8**] [**64**]. Replacement of neutral protease (usually thermolysin) with clostripain was found to greatly increase cell survival post isolation [**8**]. Previously in this chapter it was speculated that clostripain degrades the proteoglycan bridges between collagen fibrils within a fiber. It is currently unknown if it may play another role in dissociation of ECM.

The presence of excess collagenase during cell isolation can cause an immune response and while the neutral protease is necessary as a supplementary enzyme too much can significantly affect cell survival and viability. These findings suggest that for promoting survival

after cell isolation finding the optimal ratio between Col G and Col H is ideal. Based on the proposed mechanism for the synergy between Col G and Col H we believe that the optimal ratio for Col G to Col H in cell isolation should be approximately 3:1 (see figure 8B). Hydrolysis by Col H essentially separates the tropocollagen molecule into three large fragments. For efficient hydrolysis, each fragment should have its own Col G molecule. Previous results for digestion of human pancreas have shown that a 1:1 mixture of Col G to Col H is most effective however, the most Col G heavy ratio that was tested was a 2:1 Col G to Col H ratio. In cell isolation, the non-fibrillar collagens withing the extrafibrillar space must be dismantled before tropocollagens can be exposed for hydrolysis. Col G with its dual CBDs is thought to be best suited in wedging between a heterotypic collagen fibril and the non-fibril type VI collagen it is connected to via the type V collagen propeptide. It may also wedge between the type VI collagen and the type IV collagen of the cell basement membrane. Hydrolysis of this connection between the basement membrane and the collagens of the interstitial space is necessary for efficient cell isolation [**2**] [**31**]. Due to this slightly more Col G may be needed in the collagenase mixture than the 3:1 ratio.

4.4 Conclusion

Using derived structural and binding information for Col G and Col H as well as collagen hydrolysis information, both new and old, a method for collagenolytic synergy between Col G and Col H was derived and an optimal ratio of Col G to Col H for the best cell isolation results was proposed [**4**] [**10**]. Based on Col H hydrolyzing at three specific locations in a tropocollagen for optimal hydrolysis three molecules of Col G will be needed for each molecule of Col H. Fragment pattern for hydrolysis of collagen suggests that enzyme is a non-processive endopeptidase. This ratio will need to be tested to see if it leads to a higher yield pancreatic islet

than what has previously been shown. Much about Col H hydrolysis is still unknown such as the further processing of the 35 kDa N-terminal fragment seen in types I, II and III collagens [**10**]. Further research into the structure and function of Col G and Col H is necessary in order to further optimize the use of Col G/Col H mixtures for medical purposes.

Chapter 4: Figures

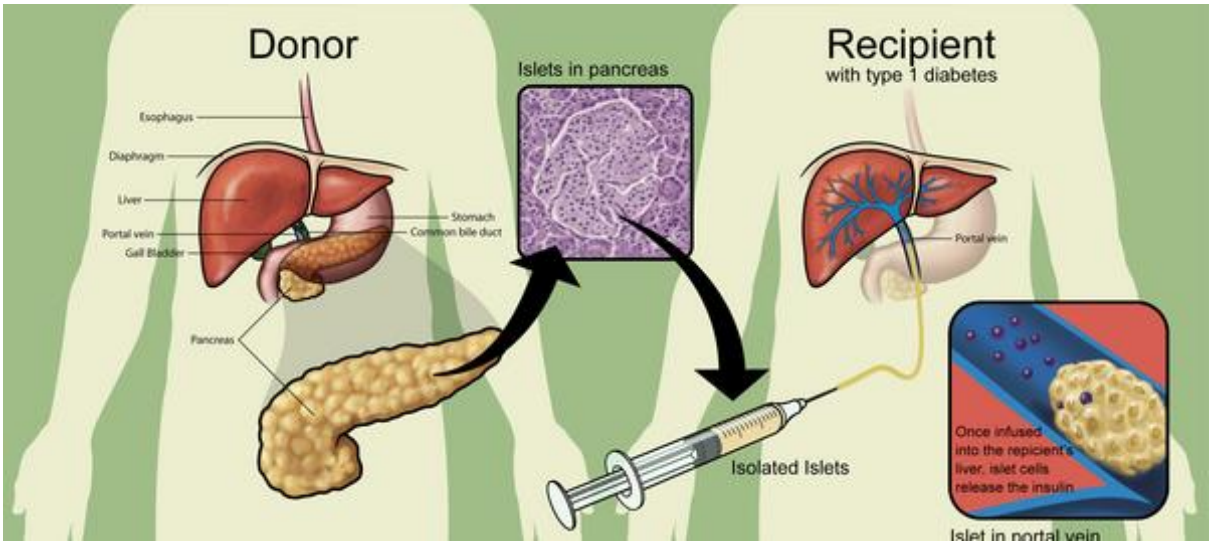


Figure 1: Cartoon outlining pancreatic islet allotransplantation. Islets are injected into liver where they begin to release insulin. Liver is chosen as site for transplantation as it allows for islets to receive sufficient amount of oxygen and nutrition [**3**]

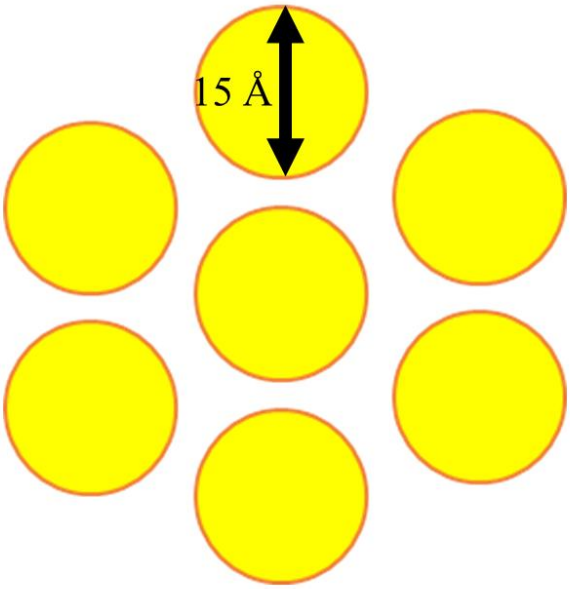


Figure 2: Head-on representation of Orgel micro-fibril. Tropocollagens are packed into a quasi-hexagonal structure with the center-to-center distance being 15 Å.

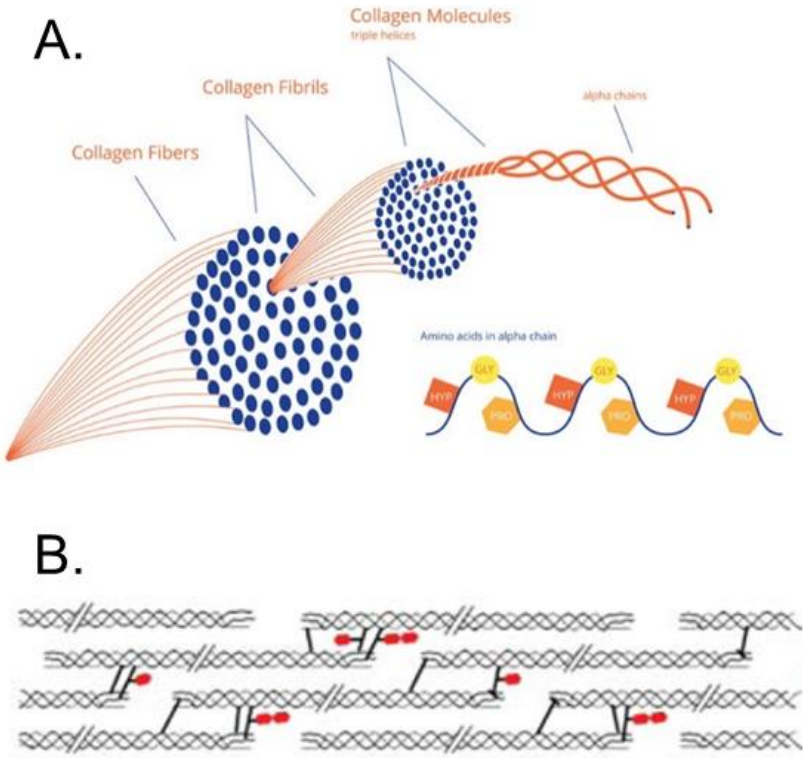


Figure 3: **A**. Cartoon representation of the hierarchical structure of fiber forming collagens. 3 α-chains with composition X-Y-Gly with X and Y most commonly being Pro and Hyp pack together to form triple helical tropocollagen. **B.** Once secreted into the extracellular space tropocollagens pack into micro-fibril. Tropocollagen is both intra and intermolecularly crosslinked at Lys residues in telopeptide and triple helical regions by LOX to give micro-fibril extra tensile strength. Figure 3B is adapted from Yamauchi et al. 2012

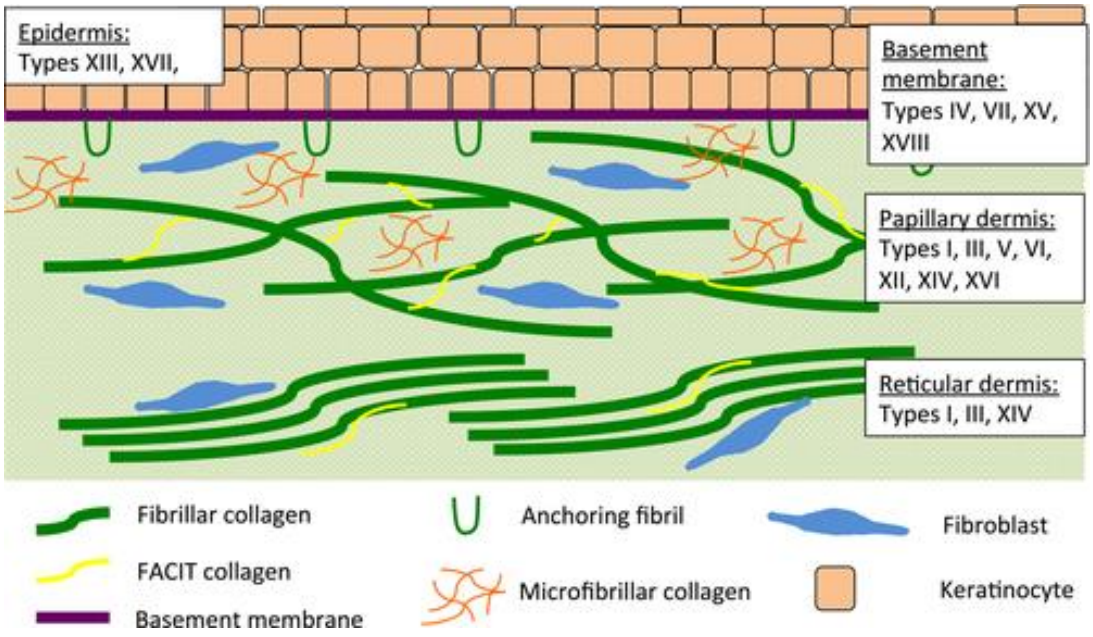


Figure 4: Cartoon representation of basement membranes. Col G, with its dual CBDs may be able to intercalate between fibrillar collagens to degrade FACIT and microfibrillar collagen holding collagen fibers together [**1**].

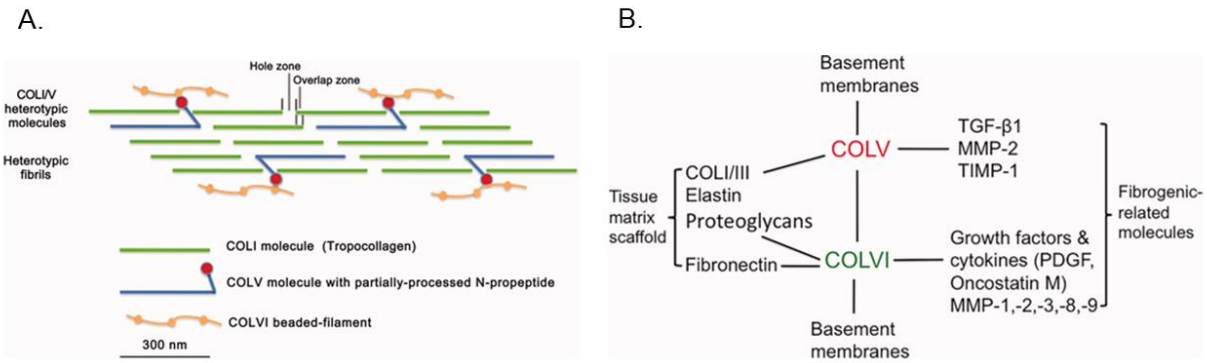


Figure 5: **A.** Representation of heterotypic collagen fibrils containing type V collagen. **B.** Chart showing the role type V collagen plays in keeping cells bound to basement membranes and maintaining tissue matrix scaffold. Figures were adopted from Mak et al.. 2016 [**2**].

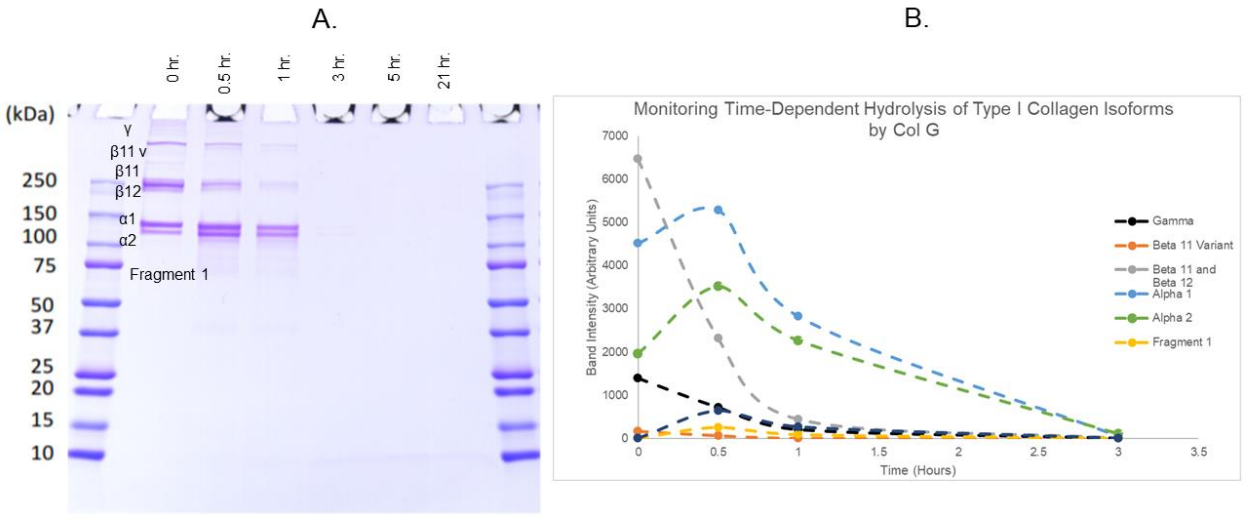
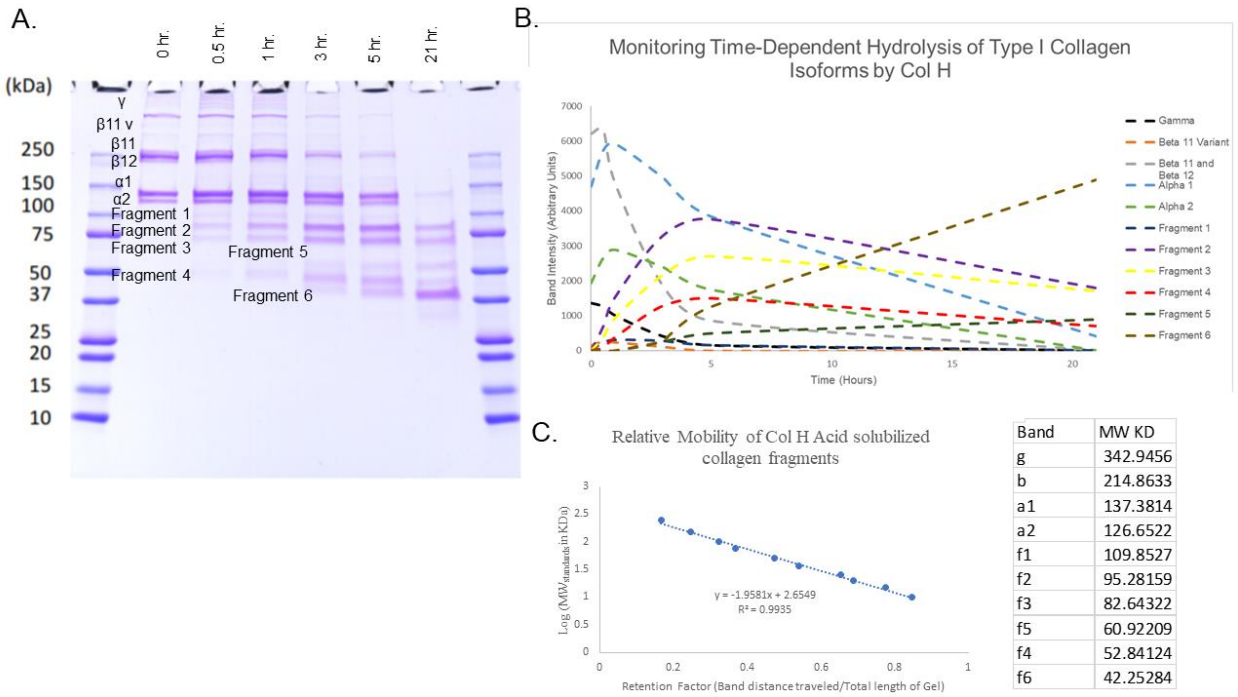


Figure 6: **A.** SDS PAGE results from Col G hydrolysis of bovine type I acid solubilized collagen **B.** Monitoring band intensity from A. over time. Initial increase for α1 and α2 chains is from hydrolysis of γ collagen (two α1 and one α2 chains forming a trimer) and β11/β12 collagens (two α1 chains or one α1 and one α2 chains forming a dimer). At one hour ratio of intensity between α1 and α2 moves from 2:1 to approximately 1:1 indicating an increase in the release of α2 chains. In order to promote release of α2 chain N-telopeptide must be cleaved off by Col G prior to hydrolysis.



Band	MW KD
g	342.9456
b	214.8633
a1	137.3814
a2	126.6522
f1	109.8527
f2	95.28159
f3	82.64322
f5	60.92209
f4	52.84124
f6	42.25284

Figure 7: **A.** SDS PAGE results from Col H hydrolysis of bovine type I acid solubilized collagen. Multiple fragments is indicative of a non-processive endoprotease. **B.** Monitoring band intensity from A. over time. Results tend to agree with French et al. 1987 with the exception of fragment 5 [**4**] **C.** Relative mobility from SDS PAGE was used to calculate molecular weight for each fragment.

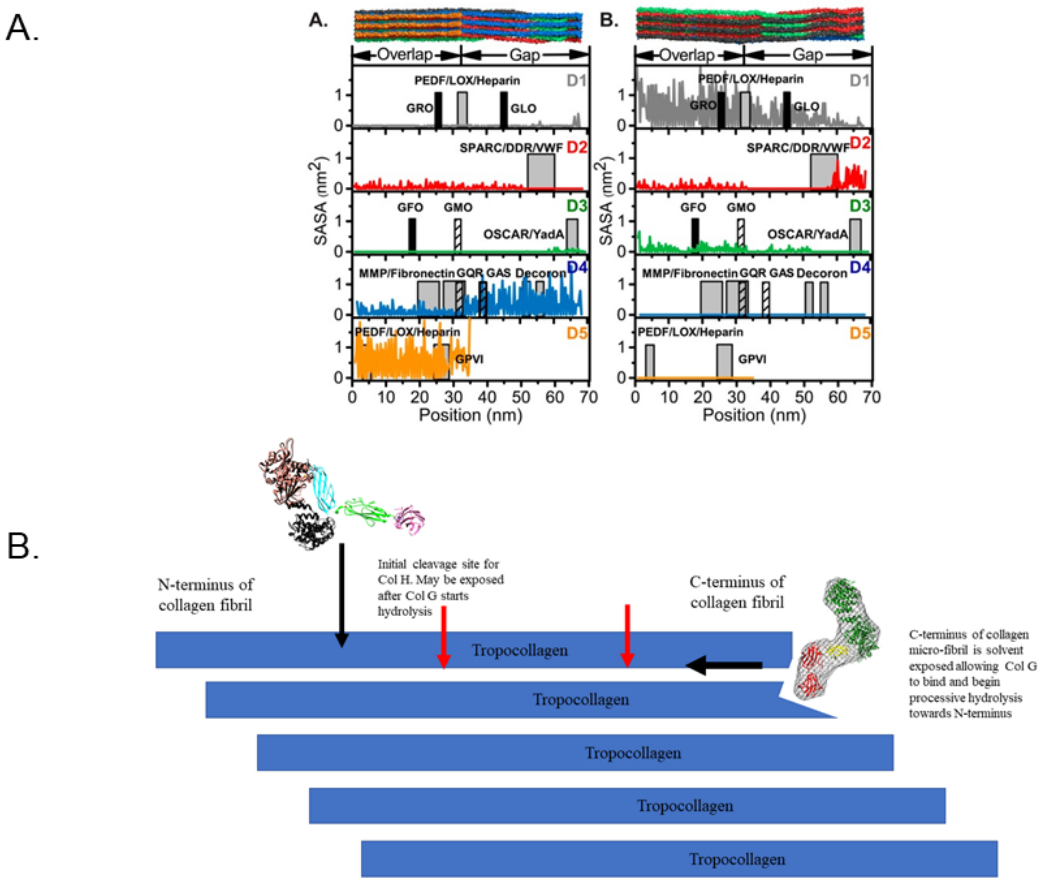


Figure 8: **A.** SASA calculations from Hoop et al. 2017 [**11**]. Calculations revealed that on side A of micro-fibril D5 (C-terminus) is solvent exposed. D2 (near Col H initial hydrolysis site for acid solubilized collagen) on side B is slightly exposed and may become more solvent accessible after Col G starts to unwind micro-fibril. **B.** Cartoon representation of proposed collagen hydrolysis synergy. C-terminus of micro-fibril is packed less tightly than the rest of the molecule. Col G binds and begins hydrolysis while moving toward N-terminus. As the enzymes moves, it would begin to partially unwind the micro-fibril exposing the initial hydrolysis site for Col H. After the initial hydrolysis Col H cleaves at two regions further toward the C-terminus (red arrow) [**14**]. Based on the proposed hypothesis for collagen fibril degradation synergy this research speculates that the optimal ratio for isolation of pancreatic islets is 3:1 Col H to Col G.

Works Cited

1. P. Caviness, R. Bauer, K. Tanaka, K. Janowska, J. R. Roeser, D. Harter, J. Sanders, C. Ruth, O. Matsushita and J. Sakon, *Ca(2+) -induced orientation of tandem collagen binding domains from clostridial collagenase colg permits two opposing functions of collagen fibril formation and retardation*, Febs j **285** (2018), no. 17, 3254-3269.

2. K. M. Mak, C. Y. M. Png and D. J. Lee, *Type v collagen in health, disease, and fibrosis*, The Anatomical Record **299** (2016), no. 5, 613-629.

3. M. R. Rickels and R. P. Robertson, *Pancreatic islet transplantation in humans: Recent progress and future directions*, Endocrine reviews **40** (2019), no. 2, 631-668.

4. M. F. French, K. A. Mookhtiar and H. E. Van Wart, *Limited proteolysis of type i collagen at hyperreactive sites by class i and ii clostridium histolyticum collagenases: Complementary digestion patterns*, Biochemistry **26** (1987), no. 3, 681-687.

5. R. C. McCarthy, A. G. Breite, M. L. Green and F. E. Dwulet, *Tissue dissociation enzymes for isolating human islets for transplantation: Factors to consider in setting enzyme acceptance criteria*, Transplantation **91** (2011), no. 2, 137-145.

6. H. Alipour, A. Raz, S. Zakeri and N. Dinparast Djadid, *Therapeutic applications of collagenase (metalloproteases): A review*, Asian Pacific Journal of Tropical Biomedicine **6** (2016), no. 11, 975-981.

7. M. L. Green, A. G. Breite, C. A. Beechler, F. E. Dwulet and R. C. McCarthy, *Effectiveness of different molecular forms of c. Histolyticum class i collagenase to recover islets*, Islets **9** (2017), no. 6, 177-181.

8. H. Brandhorst, P. R. Johnson, J. Mönch, M. Kurfürst, O. Korsgren and D. Brandhorst, *Comparison of clostripain and neutral protease as supplementary enzymes for human islet isolation*, Cell transplantation **28** (2019), no. 2, 176-184.

9. A. G. Breite, R. C. McCarthy and F. E. Dwulet, *Characterization and functional assessment of clostridium histolyticum class i (c1) collagenases and the synergistic degradation of native collagen in enzyme mixtures containing class ii (c2) collagenase*, Transplantation Proceedings **43** (2011), no. 9, 3171-3175.

10. M. F. French, A. Bhown and H. E. Van Wart, *Identification of clostridium histolyticum collagenase hyperreactive sites in type i, ii, and iii collagens: Lack of correlation with local triple helical stability*, J Protein Chem **11** (1992), no. 1, 83-97.

11. P. Gentile and A. Sterodimas, *Adipose-derived stromal stem cells (ascs) as a new regenerative immediate therapy combating coronavirus (covid-19)-induced pneumonia*, Expert opinion on biological therapy (2020), 1-6.

12. N. Sugita, Y. Moriguchi, M. Sakaue, D. A. Hart, Y. Yasui, K. Koizumi, R. Chijimatsu, S. Shimomura, Y. Ikeda, H. Yoshikawa and N. Nakamura, *Optimization of human mesenchymal stem cell isolation from synovial membrane: Implications for subsequent tissue engineering effectiveness*, Regen Ther **5** (2016), 79-85.

13. K. Ishii, N. Suzuki, Y. Mabuchi, I. Sekiya and C. Akazawa, *Technical advantage of recombinant collagenase for isolation of muscle stem cells*, Regen Ther **7** (2017), 1-7.

14. H. Brandhorst, D. Brandhorst, F. Hesse, D. Ambrosius, M. Brendel, Y. Kawakami and R. G. Bretzel, *Successful human islet isolation utilizing recombinant collagenase*, Diabetes **52** (2003), no. 5, 1143.

15. J. P. R. O. Orgel, T. C. Irving, A. Miller and T. J. Wess, *Microfibrillar structure of type i collagen in situ*, Proceedings of the National Academy of Sciences of the United States of America **103** (2006), no. 24, 9001-9005.

16. M. Yamauchi and M. Sricholpech, *Lysine post-translational modifications of collagen*, Essays Biochem **52** (2012), 113-133.

17. K. Sato, T. Ebihara, E. Adachi, S. Kawashima, S. Hattori and S. Irie, *Possible involvement of aminotelopeptide in self-assembly and thermal stability of collagen i as revealed by its removal with proteases*, J Biol Chem **275** (2000), no. 33, 25870-25875.

18. K. E. Kadler, D. F. Holmes, J. A. Trotter and J. A. Chapman, *Collagen fibril formation*, The Biochemical journal **316 (Pt 1)** (1996), no. Pt 1, 1-11.

19. Y.-Z. Zhang, L.-Y. Ran, C.-Y. Li and X.-L. Chen, *Diversity, structures, and collagen-degrading mechanisms of bacterial collagenolytic proteases*, Applied and environmental microbiology **81** (2015), no. 18, 6098-6107.

20. Y. Geng, D. McQuillan and P. J. Roughley, *Slrp interaction can protect collagen fibrils from cleavage by collagenases*, Matrix Biol **25** (2006), no. 8, 484-491.

21. S. Kalamajski and Å. Oldberg, *The role of small leucine-rich proteoglycans in collagen fibrillogenesis*, Matrix Biology **29** (2010), no. 4, 248-253.

22. D. R. Eyre, M. Weis, D. M. Hudson, J. J. Wu and L. Kim, *A novel 3-hydroxyproline (3hyp)-rich motif marks the triple-helical c terminus of tendon type i collagen*, J Biol Chem **286** (2011), no. 10, 7732-7736.

23. S. Kalamajski, A. Aspberg and A. Oldberg, *The decorin sequence syiriadtnit binds collagen type i*, J Biol Chem **282** (2007), no. 22, 16062-16067.

24. Y. X. Liu, Z. Q. Liu, L. Song, Q. R. Ma, D. Y. Zhou, B. W. Zhu and F. Shahidi, *Effects of collagenase type i on the structural features of collagen fibres from sea cucumber (stichopus japonicus) body wall*, Food Chem **301** (2019), 125302.

25. C. B. Caputo, D. K. MacCallum, J. H. Kimura, J. Schrode and V. C. Hascall, *Characterization of fragments produced by clostripain digestion of proteoglycans from the swarm rat chondrosarcoma*, Arch Biochem Biophys **204** (1980), no. 1, 220-233.

26. W. M. Mitchell, *Cleavage at arginine residues by clostripain*, Methods Enzymol **47** (1977), 165-170.

27. R. V. Iozzo and L. Schaefer, *Proteoglycan form and function: A comprehensive nomenclature of proteoglycans*, Matrix Biology **42** (2015), 11-55.

28. R. L. Stevens, K. Otsu and K. F. Austen, *Purification and analysis of the core protein of the protease-resistant intracellular chondroitin sulfate e proteoglycan from the interleukin 3-dependent mouse mast cell*, J Biol Chem **260** (1985), no. 26, 14194-14200.

29. A. Chakravorty, M. M. Awad, T. J. Hiscox, J. K. Cheung, G. P. Carter, J. M. Choo, D. Lyras and J. I. Rood, *The cysteine protease α-clostripain is not essential for the pathogenesis of clostridium perfringens-mediated myonecrosis*, PloS one **6** (2011), no. 7, e22762-e22762.

30. T. Kobayasi and T. Karlsmark, *Type v and vi collagen for cohesion of dermal fibrillar structures*, J Submicrosc Cytol Pathol **38** (2006), no. 2-3, 103-108.

31. H. Shima, A. Inagaki, T. Imura, Y. Yamagata, K. Watanabe, K. Igarashi, M. Goto and K. Murayama, *Collagen v is a potential substrate for clostridial collagenase g in pancreatic islet isolation*, Journal of diabetes research **2016** (2016), 4396756-4396756.

32. D. E. Birk, *Type v collagen: Heterotypic type i/v collagen interactions in the regulation of fibril assembly*, Micron **32** (2001), no. 3, 223-237.

33. S. Symoens, M. Renard, C. Bonod-Bidaud, D. Syx, E. Vaganay, F. Malfait, S. Ricard-Blum, E. Kessler, L. Van Laer, P. Coucke, F. Ruggiero and A. De Paepe, *Identification of binding partners interacting with the α1-n-propeptide of type v collagen*, Biochem J **433** (2011), no. 2, 371-381.

34. R. R. Bruns, W. Press, E. Engvall, R. Timpl and J. Gross, *Type vi collagen in extracellular, 100-nm periodic filaments and fibrils: Identification by immunoelectron microscopy*, The Journal of cell biology **103** (1986), no. 2, 393-404.

35. R. J. Wenstrup, J. B. Florer, E. W. Brunskill, S. M. Bell, I. Chervoneva and D. E. Birk, *Type v collagen controls the initiation of collagen fibril assembly*, J Biol Chem **279** (2004), no. 51, 53331-53337.

36. T. Watanabe-Nakayama, M. Itami, N. Kodera, T. Ando and H. Konno, *High-speed atomic force microscopy reveals strongly polarized movement of clostridial collagenase along collagen fibrils*, Sci Rep **6** (2016), 28975.

37. S. T. L. Philominathan, T. Koide, K. Hamada, H. Yasui, S. Seifert, O. Matsushita and J. Sakon, *Unidirectional binding of clostridial collagenase to triple helical substrates*, Journal of Biological Chemistry **284** (2009), no. 16, 10868-10876.

38. T. Toyoshima, O. Matsushita, J. Minami, N. Nishi, A. Okabe and T. Itano, *Collagen-binding domain of a clostridium histolyticum collagenase exhibits a broad substrate spectrum both in vitro and in vivo*, Connect Tissue Res **42** (2001), no. 4, 281-290.

39. H. Furthmayr, H. Wiedemann, R. Timpl, E. Odermatt and J. Engel, *Electron-microscopical approach to a structural model of intima collagen*, The Biochemical journal **211** (1983), no. 2, 303-311.

40. C. A. Schneider, W. S. Rasband and K. W. Eliceiri, *Nih image to imagej: 25 years of image analysis*, Nat Methods **9** (2012), no. 7, 671-675.

41. L. Rittié, *Type i collagen purification from rat tail tendons*, Methods Mol Biol **1627** (2017), 287-308.

42. D. M. Neville, Jr., *Molecular weight determination of protein-dodecyl sulfate complexes by gel electrophoresis in a discontinuous buffer system*, J Biol Chem **246** (1971), no. 20, 6328-6334.

43. O. Matsushita, T. Koide, R. Kobayashi, K. Nagata and A. Okabe, *Substrate recognition by the collagen-binding domain of clostridium histolyticum class i collagenase*, J Biol Chem **276** (2001), no. 12, 8761-8770.

44. N. P. Morris, L. I. Fessler and J. H. Fessler, *Procollagen propeptide release by procollagen peptidases and bacterial collagenase*, J Biol Chem **254** (1979), no. 21, 11024-11032.

45. S. T. L. Philominathan, T. Koide, O. Matsushita and J. Sakon, *Bacterial collagen-binding domain targets undertwisted regions of collagen*, Protein Science **21** (2012), no. 10, 1554-1565.

46. G. B. Fields, *A model for interstitial collagen catabolism by mammalian collagenases*, J Theor Biol **153** (1991), no. 4, 585-602.

47. E. Leikina, M. V. Mertts, N. Kuznetsova and S. Leikin, *Type i collagen is thermally unstable at body temperature*, Proceedings of the National Academy of Sciences **99** (2002), no. 3, 1314.

48. K. Kobayashi, T. Ito and T. Hoshino, *Electron microscopic demonstration of acid-labile, 4d-staggered intermolecular association of collagen formed in vitro*, Coll Relat Res **5** (1985), no. 3, 253-260.

49. O. Matsushita, C. M. Jung, J. Minami, S. Katayama, N. Nishi and A. Okabe, *A study of the collagen-binding domain of a 116-kda clostridium histolyticum collagenase*, J Biol Chem **273** (1998), no. 6, 3643-3648.

50. K. Okuyama, H. P. Bächinger, K. Mizuno, S. Boudko, J. Engel, R. Berisio and L. Vitagliano, "Re: Microfibrillar structure of type i collagen in situ," *Acta crystallogr d biol crystallogr*, vol. 65, United States, 2009, pp. 1007-1008; author reply 1009-1010.

51. A. N. Balamurugan, M. L. Green, A. G. Breite, G. Loganathan, J. J. Wilhelm, B. Tweed, L. Vargova, A. Lockridge, M. Kuriti, M. G. Hughes, S. K. Williams, B. J. Hering, F. E. Dwulet and R. C. McCarthy, *Identifying effective enzyme activity targets for recombinant class i and class ii collagenase for successful human islet isolation*, Transplantation direct **2** (2015), no. 1, e54-e54.

52. C. L. Hoop, J. Zhu, A. M. Nunes, D. A. Case and J. Baum, *Revealing accessibility of cryptic protein binding sites within the functional collagen fibril*, Biomolecules **7** (2017), no. 4, 76.

53. I. Streeter and N. H. de Leeuw, *Atomistic modeling of collagen proteins in their fibrillar environment*, J Phys Chem B **114** (2010), no. 41, 13263-13270.

54. R. Berisio, L. Vitagliano, L. Mazzarella and A. Zagari, *Crystal structure of the collagen triple helix model [(pro-pro-gly)(10)](3)*, Protein science : a publication of the Protein Society **11** (2002), no. 2, 262-270.

55. X. Sun, Y. Li, Z. Tian, Y. Qian, H. Zhang and L. Wang, *A novel thermostable chitinolytic machinery of streptomyces sp. F-3 consisting of chitinases with different action modes*, Biotechnology for Biofuels **12** (2019), no. 1, 136.

56. S. J. Horn, A. Sørbotten, B. Synstad, P. Sikorski, M. Sørlie, K. M. Vårum and V. G. Eijsink, *Endo/exo mechanism and processivity of family 18 chitinases produced by serratia marcescens*, Febs j **273** (2006), no. 3, 491-503.

57. J. Jalak, M. Kurašin, H. Teugjas and P. Väljamäe, *Endo-exo synergism in cellulose hydrolysis revisited*, J Biol Chem **287** (2012), no. 34, 28802-28815.

58. M. E. Himmel, S. Y. Ding, D. K. Johnson, W. S. Adney, M. R. Nimlos, J. W. Brady and T. D. Foust, *Biomass recalcitrance: Engineering plants and enzymes for biofuels production*, Science **315** (2007), no. 5813, 804-807.

59. G. H. Wolters, G. H. Vos-Scheperkeuter, H. C. Lin and R. van Schilfgaarde, *Different roles of class i and class ii clostridium histolyticum collagenase in rat pancreatic islet isolation*, Diabetes **44** (1995), no. 2, 227-233.

60. A. N. Balamurugan, A. G. Breite, T. Anazawa, G. Loganathan, J. J. Wilhelm, K. K. Papas, F. E. Dwulet, R. C. McCarthy and B. J. Hering, *Successful human islet isolation and transplantation indicating the importance of class 1 collagenase and collagen degradation activity assay*, Transplantation **89** (2010), no. 8, 954-961.

61. R. C. McCarthy, B. Spurlin, M. J. Wright, A. G. Breite, L. K. Sturdevant, C. S. Dwulet and F. E. Dwulet, *Development and characterization of a collagen degradation assay to assess purified collagenase used in islet isolation*, Transplant Proc **40** (2008), no. 2, 339-342.

62. T. Kin, X. Zhai, D. O'Gorman and A. M. Shapiro, *Detrimental effect of excessive collagenase class ii on human islet isolation outcome*, Transpl Int **21** (2008), no. 11, 1059-1065.

63. K. L. Gorres and R. T. Raines, *Prolyl 4-hydroxylase*, Critical reviews in biochemistry and molecular biology **45** (2010), no. 2, 106-124.

64. R. C. McCarthy, M. L. Green and F. E. Dwulet, *Evolution of enzyme requirements for human islet isolation*, OBM Transplantation **2** (2018), no. 4.

Chapter 5: *Vibrio alginolyticus* REC domain consists of two domains with two-fold symmetry connected by a flexible linker

Abstract

Currently attempts at crystallization of REC domain have proven to be non-successful. In this chapter we sought to understand why attempts have proven unsuccessful. Homology model of REC domain from Itasser revealed a two-domain structure (I and II), with both domains containing a Rossman fold, connected by 21-residue flexible linker. Circular dichroism and Tryptophan fluorescence emission spectra gave support to the homology model. Limited proteolysis revealed that the linker region is susceptible to hydrolysis. While the structures for subdomains I and II are structurally homologous (R.M.S.D. = 1.9 Å) subdomain I from REC is missing key residues present in the active site of subdomain II. The potential origin of two domain REC incorporated into VarS sensor histidine kinase is also investigated in this chapter.

5.1 Introduction

In general, two-component systems are comprised of a sensor protein (Histidine Kinase) and a corresponding response regulator. Since each particular system is specialized to respond to a specific environmental signal there may be multiple systems present in a single cell [**2**]. Upon receiving a signal in the periplasmic region, the sensor kinase forms a dimer and phosphate is transferred from ATP to a conserved histidine. There are two different classes of histidine kinases (see figure 1A and 1B). Class I kinases, of which VarS is a member, have the CA and dHp domains directly linked to each other. In Class II kinases the CA domain is separated from the domain containing the phosphorylated histidine by a hinge like domain important for signaling [**3**]. There are two types of Class I kinases [**3**]. The first type is known as a simple two component in which the phosphate bound to the histidine in the CA domain is immediately

transferred to a response regulator. The second type of two-component system is known as a phosphorelay, in which the phosphate at the CA domain is first transferred to a conserved aspartate located on an intermediate receiver (REC) domain and then transferred to a conserved histidine located on the histidine phosphor-transfer (HPT) domain before finally being transferred to a response regulator

REC domains are approximately 120 residues long with a Rossman like fold consisting of five-stranded parallel β-sheet structure with each β-sheet connected by an α-helix [**2**]. REC domains are known as the common regulatory module for all response regulators [**4**]. These domains fold easily, are stable and compact and undergo conformation changes upon phosphorylation [**5**]. In response regulators binding at the N-terminal REC domain causes the protein to dimerize which allows it to bind to DNA [**5**]. REC domains have a small set of highly conserved residues located within the active site. The most conserved residue of course being the phosphate accepting Asp. High energy acyl phosphate may provide energy coupling to drive a conformational change within REC [**2**]. In addition to the phosphate accepting Asp, the following key residues in the active site are conserved; an Asp residue in the adjacent β-sheet, two additional Asp or Glu residues in the active site that coordinate a Mg^{2+} and a conserved Lys residue that forms a salt bridge with the phosphate bound to the active site [**2**]. In addition to the active site residues Phe/Tyr are highly conserved on β-sheet 5 while Ser/Thr are highly conserved on β-sheet 5. In activated response regulators Phe/Tyr interact with DNA binding effector domain while Ser/Thr help to coordinate the conserved aromatic residues [**2**].

Some response regulators contain two REC domains. The first domain contains Asp residue which receives Phosphate while the second domain has the Rossman-like fold characteristic of the first domain, however it is non-functional suggesting that second domain

may have arisen from gene duplication event [**1**] [**6**]. In PleD presence of secondary REC adaptor domain helps in intra and intermolecular dimer formation via hydrophobic interactions and salt bridge formation with the first domain [**1**]. PleD dimer is involved in the formation of cyclic di-guanosine monophosphate (c-di-GMP), an important signaling molecule from two GTP molecules. Dimer formation in PleD allows GTP loaded DGC domains from both subunits to come together and catalyze the formation of c-di-GMP [**1**]. Addition of REC adaptor domain may help promote dimer formation however, there are multiple DGC domains at the C-terminus of only one REC domains indicating that adaptor domain is not necessary [**1**].

Nearly 25% of histidine kinases incorporate an internal REC domain to form phosphor-relay histidine kinases. REC domains are incorporated into histidine kinases at the C-terminus of the dHp domain via gene fusion [**7**]. Currently crystal structure of internal REC domain from VarS has yet to be solved. Solving structure would help to continue understand the different structural nuances of sensor histidine kinases. Based on the number of residues in the protein, REC from VarS is predicted to be a multi-domain protein. Understanding how the multi-domain structure of REC evolved as well as understanding how gene fusion between VarS and REC occurred would help explain the structural differences between the two domains. Understanding these structural differences would help in planning new approach for crystallization of the domain.

5.2 Methods

5.2a. Expression and Purification of VarS REC domain

The genes for REC has been previously identified and isolated [**8**]. In order to construct the expression vector for REC the gene was ligated into the pHSG398T plasmid using EcoRI and HinDIII. Plasmid was then transformed into *E. coli* BL21 cells for expression. Nickel-affinity

chromatography was used to purify the expressed protein. Once purified the protein was dialyzed into 1 x PBS for an extra round of purification as well as for storage. Production and purification of REC domain was accomplished by Dr. Takehiko Mima at Okayama University of Medicine as the Sakon lab at the University of Arkansas does not have access to gene for VarS from *V. alginolyticus*.

5.2b. Itasser Homology Modeling of REC

Due to difficulty in crystallization of REC and HPT domains homology modeling was used to give a probable structure of the proteins based on sequence homology in proteins whose structures have been solved. Random mutagenesis using a transponson based system alongside sequence alignment studies were used to find the nucleotide and amino acid sequences for REC (residue 535 – 781) [**8**]. Modeling was accomplished by inputting sequence into the Itasser (Iterative Threading Assembly Refinement) program developed by the Zhang lab at the University of Michigan [**9**]. For REC, homology model for sub-domain I was produced by deleting coordinates in the original Itasser derived homology model that correspond to sub-domain II and the linker region. Homology model for REC sub-domain II was produced in the same manner with coordinates corresponding to sub-domain I being deleted.

5.2c. Circular Dichroism and Tryptophan Emission Spectra of REC Domain

Circular Dichroism and Tryptophan Fluorescence emission were used to check if expressed REC domain secondary structure and fold matched the Itasser homology models. For both Circular Dichroism and Fluorescence experiments proteins were dialyzed into 30 mM Hepes pH 7.5, 50 mM KCl, 1 mM TCEP. For Circular Dichroism experiments best results are obtained when the protein is in a low salt buffer. For the CD experiment proteins were diluted

with the new buffer to 1.5 ml with a final concentration of 0.2 mg/ml. For Tryptophan emission experiments proteins from the CD experiments were diluted with buffer as needed. Circular Dichroism and Tryptophan emission spectra of REC domains was found using a Jasco CD spectrometer at room temperature.

5.2d. Limited Proteolysis of REC domain

Proti-Ace screen from Hampton Research was used for proteolysis of REC domain because it contains multiple enzymes that could potentially target the dynamic loop region [**10**]. Initial digestion was performed as described in the kit. Briefly, 100 µl of deionized water was added to each Protease in the kit to create a 1 mg/ml stock. From each 1 mg/ml stock a 1:1000 dilution was performed (0.001 mg/ml); each protease was diluted with pre-supplied 10 mM Hepes pH 7.5, 500 mM Sodium chloride dilution buffer. 10 µl of REC (1 mg/ml) were mixed with 10 µl of each 1:1000 protease stock and incubated at 37° C for 1 hour. After 1 hour the reaction was stopped by adding SDS-PAGE sample buffer to each sample and the samples were ran on a 15% SDS PAGE gel for analysis.

5.2e. Time Dependent Proteolysis of REC Domain with Elastase

Time Dependent digestion of REC was performed identical to initial proteolysis however the concentration of Elastase was increased 1:100 dilution (0.01 mg/ml). 30 µl of REC was mixed with 30 µl of the 1:100 dilution of elastase and the reaction was incubated for 1 hour at 37° C. Every 10 minutes 7.5 µl of the mixture was aliquoted and the reaction was stopped by adding an equal amount of SDS sample buffer. Samples were frozen to prevent potential disulfide bond cross linking from occurring. The samples were run on a 15% SDS PAGE gel to visualize the appearance of a new sub-domain.

5.2f. Identifying if multi-domain structure is present in crystal structures of related REC domains

To identify if the multi-domain structure from the REC homology is conserved in related proteins, the homology model was loaded onto the DALI server and matched against the full Protein Databank [**11**]. The structures with the top ten Z-scores were chosen and superimposed with the REC homology model using PyMol to identify if multi-domain structure with flexible linker is conserved in solved crystal structures of related REC domains. Only non-redundant structures were chosen from the Protein Databank. Conserved structural elements other than linker were noted as well. To investigate the origin of both sub-domains from REC homology model the newly created homology models for REC sub-domain I and II were loaded onto DALI server and matched against the full Protein Databank Non-redundant structures with the top ten Z-scores were chosen for both domains I and II and superimposed with PyMol.

5.3 Results and Discussion

5.3a. Description of Itasser derived REC domain model

Due to difficulties with crystallization Itasser was used to construct a model of REC domain using homologous structures as a template [**9**]. This model revealed a structure that consists of two domains (I and II) with a pseudo two-fold symmetry similar to receiver domain in PleD response regulator [**1**]. Domain I and II both have a Rossman fold common amongst REC domains [**12**]. Domain I contains four α-helices and five β-sheets and domain II contains five α-helices and five β-sheets. The domains are connected by a 21-residue linker (residues 114-133) that is predicted to highly flexible based on B-factors from the homology model [**13**]. In addition to the loop connecting domain I and II the following regions in the model are also predicted to be highly flexible: residues 64-69 connecting α-helix b to β-sheet D and residues

100-114 which make up α-helix d. The active site of REC is found in domain II. Key Mg^{2+} chelating Asp residues 140 and 141 are found on the loop connecting β-sheet A' and α-helix a'. Asp184, which receives the phosphate from the HisKa domain, is found nearby Asp140 and Asp141 on β-sheet C' (see figure 2). Key Lys residue 236, responsible for stabilizing phosphate via salt bridge is found on the loop connecting β-sheet E' and α-helix e' [**2**]. Intriguingly, LTKP motif containing Lys236 is the only site responsible for phosphate transfer conserved in both domain I and II.

5.3b. Circular Dichroism and Tryptophan Emission Spectra support REC homology model structure

Circular dichroism and tryptophan emission spectra were found to give support to the REC homology model. Circular dichroism spectrum was normalized and smoothed using exponential smoothing function from excel so that noise at lower wavelengths is minimized. However, data below 195 nm was still too noisy to be read. Starting at 250 nm REC spectrum decreases smoothly until reaching approximately 230 nm where the ellipticity decreases more rapidly. Rapid decrease continues until a minimum at 208 nm is reached and then ellipticity increases until a maximum at 195 nm is reached (see figure 3A). For a protein that is primarily α-helical circular spectra contains local minima at 223 nm and 208 nm and a protein that is primarily β-sheet has minimum at 218 nm and a maximum at 195 nm [**14**]. Circular dichroism spectra for REC domain is consistent with a protein whose secondary structure contains both α-helices and β-sheets.

Since REC domain contains only a single Trp residue; Trp emission spectrum was used to determine if the protein is folded identical to the Itasser derived homology model. Absorbance wavelength was 280 nm and the emission wavelength for the protein peaked at 343 nm (see

figure 3B). This is consistent with Trp61 being exposed to solvent [**15**]. Trp that is fully exposed to solvent has an emission wavelength maximum at 350 nm. In the homology model the Trp61 is found on α-helix b near the loop connecting α-helix b to β-sheet D. Interactions between Trp61, Ile6, Leu33, Asn36, Val43, Leu45, Val57, Pro60, Met68 and Pro70 form a hydrophobic pocket that may need to be stabilized if crystallization of REC will occur (see figure 4). In addition to protein dynamics and flexibility, amino acid composition plays a large role in crystallization. The presence of hydrophobic residues on the surface of a molecule would cause entropy of the solvent to become more negative. Hydrophobic effect of Trp61 causes water molecules in the solvent to become more ordered [**16**]. Mutation of Trp61 to Tyr may keep this proposed hydrophobic pocket stable however the presence of a hydroxyl group from Tyr would keep water molecules from associating together due to the hydrophobic effect.

5.3c. Proteolysis of REC domain confirms the existence of flexible loop in the protein

Limited Proteolysis of REC domain was performed to better probe the existence of dynamic loop connecting domain I and II. After one hour the digestion was stopped, and all samples were run on a 15% SDS PAGE gel. The loop contains Arg131 which is a potential site for trypsin. The loop contains the residues Ala122, Leu120 and Leu132 for elastase to potentially attack. Chymotrypsin and subtilisin have previously been known to cleave the c-terminal side of Leu. Papain could potentially cleave the same site in the loop as trypsin. Endoproteinase Glu-C cleaves the peptide bond c-terminal to Glu which there are three of present in the loop region (Glu115, Glu128 and Glu130). Only chymotrypsin, elastase and subtilisin were able to digest REC domain after 1 hour (see figure 5A). Molecular weight of wild-type REC domain is approximately 28 kDa and on the gel undigested REC shows up at approximately 35 kDa. REC

domain is most likely running larger on the 15% due to the protein not being fully unfolded or basic amino acids (Arg and Lys) interfering with SDS binding to the protein [**17**].

After digestion, the new band appearing at approximately 27 kDa for elastase appeared to be most stable as it lacks the long smear trail seen in the new bands for α-chymotrypsin and subtilisin. Time dependent proteolysis of REC with elastase was performed with a fresh batch of REC domain to better isolate the new lower molecular weight species seen in the initial digestion. After 10 minutes, a new band appeared between 27 kDa and 17 kDa. After 60 minutes, this band is still present which indicates that this new species must be stable (see figure 5B).

Potential high molecular weight contaminants are present in the initial proteolysis and time dependent proteolysis experiments. These contaminants are most likely from the initial expression and purification of REC domain from *E. coli* but could be the result of soluble REC aggregates from Cys disulfide bond formation. REC domain contains 3 Cys residues. The apparent molecular weight of the contaminants (80 kDa and 100 kDa) are not consistent with that of a REC dimer however it may be consistent with a REC trimer (REC monomer is approximately 27 kDa). Band will need to be digested and analyzed to identify the contaminant. If bands are the result of soluble REC aggregates, protein will need to be purified to remove aggregates and proteolysis experiment will need to be repeated.

5.3d. Flexible loop between domain I and II is inhibiting Crystallization of REC domain

Protein crystals form when over time proteins in solution come together through intermolecular contacts [**16**]. The formation of these intermolecular contacts, like all things, is driven by change in Gibbs free energy (ΔG°) [**16**]. Briefly, a reaction with a negative value for

$\Delta G°$ is likely to proceed spontaneously. For protein crystallization the Gibbs free energy equation can be written as

$$\Delta G°_{Crys} = \Delta H°_{Crys} - T(\Delta S°_{Protein} + \Delta S°_{Solvent})$$

It has been theorized that a change in $\Delta H°_{Crys}$ is not important for crystallization which means that crystallization by change in $\Delta S°$ [**16**]. Imposing order upon a three-dimensional biomolecule results in the loss of six translational and rotational degrees of freedom per molecule and results in a massive negative change in entropy for $\Delta S°_{Protein}$ and is only slightly compensated by newly created vibrational degrees of freedom [**18**]. To overcome this negative $\Delta S°_{Protein}$ ordered water molecules are released into the solvent as protein crystal contacts form to give a large positive value for $\Delta S°_{Solvent}$ [**16**].

Highly dynamic and flexible regions in proteins (such as loops) would cause $\Delta S°_{Protein}$ to have a large negative value. As such, proteins with more ordered structures and less flexible regions are more likely to form quality crystals for x-ray crystallography. In the context of REC domain, removal of the loop region (crystalize just the active site domain) or truncation of the loop region may allow the protein to crystallize easier. To determine if this multi-domain structure connected by a flexible loop region is conserved in proteins related to REC domain, the homology model was loaded onto the DALI server and matched against the full Protein Databank [**11**].

The structures with the top 10 Z-scores were chosen and aligned using PyMol and it was found that all but two of the structures contained only the active site domain [**19**] [**20**] [**21**]. Figures are summarized in figure 6A. For the other two structures the loop region is present, but it is highly truncated (see figure 6B and 6C). Chosen structures superimposed with homology

model of REC nicely; all with R.M.S.D values < 1 Å [**22**]. The active site triad key for phosphorylation in REC domains is highly conserved in all the structures (see figure 6D) [**2**]. In addition to the active site acidic residues, key Lys residue responsible for forming a salt bridge with active site phosphate is conserved in all DALI structures [**2**].

In the REC homology model Leu134, Leu153, Val157, Val160, Leu180 and Leu247 appear to form hydrophobic pocket at the N-terminus of domain II. This pocket appears to be conserved throughout the DALI structures with homologous residues being in place of the Leu and Val residues (see figure 6E). To make crystallization of REC domain easier a truncated version of the protein containing just domain II was expressed. To keep the N-terminal hydrophobic pocket stabilized the new N-terminus was made to be Leu134. REC domain containing just domain II was found to express rather poorly. Currently it is unknown if lowering the temperature during expression or if increasing the expression time would prevent the formation of inclusion bodies.

If expression levels for REC domain containing only domain II cannot be improved, shortening of the loop between domain I and II may provide an alternative method for decreasing $\Delta S°_{Protein}$ and increasing the chance of crystallization (see figure 2) [**16**]. Based on the B-factors for the loop region we predict that deletion of residues 117 to 129 from the loop region would sufficiently decrease the $\Delta S°_{Protein}$ while still leaving enough residues to efficiently link the two domains. Glu115 and Leu132 lie at either end of the loop between domain I and II with 7.9 Å distance between the two. The addition of Ser116 and Arg130 to both sides of the loop would be enough to sufficiently bridge the 7.9 Å gap between the two domains.

5.3e. Gene duplication may be responsible for two REC sub-domains

The primary known function of most REC domains is to act as a phosphorylation mediated switch for response regulators [**12**]. While REC domains are commonly attached to DNA binding domains they are also found on a variety of output domains including enzymatic, RNA-binding and ligand binding domains [**5**]. However, some REC domains are not attached to any output domain and exist independently. Approximately 50% of response regulators in Archaea consist of only a single REC domain with no output domain [**12**]. Currently it is unknown how REC domains began to be incorporated into different histidine kinase forming Phospho-relay two component systems. Incorporation of the REC domain into a histidine kinase would present an additional regulatory checkpoint for environmental stimulus response that is not present for the response regulator [**2**]. REC domains in response regulators and phosphor-relays have similar structures and can contain either one or two sub-domains [**1**] [**2**]. In response regulators binding at the N-terminal REC domains causes the protein to dimerize which allows the C-terminal effector domain to bind to DNA [**5**]. It is currently unknown if internal REC domains from histidine kinase forms a dimer as well. Below we investigate the potential origin of REC domain incorporation into histidine kinases.

In REC homology model residues corresponding to domain II and the linker region were deleted to create a model for just domain I. For domain II, this process was repeated with domain I and linker region being deleted. The newly derived homology models for domain I and II from REC were loaded onto the DALI server and matched against each other. Structures for domains I and II were found to be highly conserved with an R.M.S.D = 1.9 Å (see figure 7A). The homology model of REC domain I is missing the key Asp residue that is phosphorylated during histidine kinase signaling. Rec domain I is also missing one of the acidic residues necessary for

Mg^{2+} chelation. However, as stated earlier in the chapter key Lys residue responsible for stabilizing the phosphate is conserved in domain I (see figure 7B) [**2**]. REC domain II was found to contain all the necessary motifs necessary for phosphate transfer.

In addition to being matched against each other, the two derived homology models for REC domain I and II were matched against the full Protein Databank. For both domain I and II the structures with the top 10 Z-scores were chosen and aligned using PyMol. For REC domain I it was found that the structures superimposed well with the of domain I homology model, all with R.M.S.D. values < 2.5 Å (see figure 7C). For REC domain II it was found that the structures superimposed well with R.M.S.D. values < 1.9 Å (see figure 7D). However, the percent sequence identity for domain II was much higher than that for domain I. For domain II, the chosen structures had percent sequence identities higher than 30% while for domain I chosen structures had percent sequence identities < 20%. Domain II of REC had a higher sequence identity when matched against the entire Protein Databank compared to domain I, which indicates that domain II must be more closely related to this single domain common ancestor [**11**]. Due to this domain II must have evolved in REC before domain I.

DALI results for REC homology model revealed that structures containing two REC domains are from PLED like response regulators with GGDEF effector domains [**1**]. The presence of dual REC domains in PLED response regulators may give a hint to the origin of REC domain from VarS. Based on their presence on a variety of output domains response regulatory REC domains may have evolved from single domain response regulators from Archaea [**23**]. Incorporation of REC domain would give early bacterial ancestors the ability to affectively respond to different stimuli and promote survival via phosphate transfer [**23**]. It is currently unknown by which horizontal gene transfer mechanism this single domain common

ancestor was incorporated into prokaryotes. It is speculated that many eukaryotic or archaeal systems evolved through horizontal gene transfer [**24**]. Archaea like single domain response regulators are still used in prokaryote signal response. *E. coli* CheY protein which regulates chemotaxis is a known relative of single domain response regulators [**23**].

While REC domains consist of approximately a single domain with approximately 120 residues PleD-like REC domains contain two sub-domains and approximately 240 residues. Gene duplication may explain the presence of dual REC domains in PleD like response regulators. In two component systems from the tomato plant, gene duplication events primarily occur in the cytoplasmic response regulator [**25**]. After gene duplication, pre-existing functions are not as likely to be conserved which may explain domain I's lack of residues involved in phosphate transfer [**6**]. PleD-like response regulators may have evolved two REC domains two assist in dimer formation of GGDEF effector domains (see figure 8). Crystal structure of PleD reveals that REC-like adaptor domain assist in homodimer formation which is essential for GGDEF dimer formation. Dimer formation is necessary to promote formation of cyclic di-guanosine monophosphate (c-di-GMP) from two molecules of GTP [**1**]. To explain the presence of two domains in VarS, gene for PleD-like response regulator may have been incorporated into gene for VarS histidine kinase via gene fusion as genes for PleD family of response regulators are found in *V. alginolyticus*. [**7**]. REC-like adaptor domain assists in dimer formation for PleD and may assist in dimer formation for VarS [**1**]. REC domain could only be incorporated into the gene for VarS by insertion at the C-terminus as insertion at the N-terminus would disrupt the transmembrane region of the new protein causing it to be unable to fold correctly [**26**]. Once being incorporated into VarS, REC must have lost GGDEF effector domain over time.

Incorporation of REC domain into the histidine kinase may explain some of the differences in residue content between domain I and II. Hydrophobic pocket present in domain I is also present in domain II however the pocket in domain II is not as hydrophobic. Trp61 from domain I is His65 in domain II, Met68 from domain I is Asn72 in domain II and Pro70 from domain I is Asp74 in domain II. The other residues that make up the hydrophobic pocket are homologous for both domains. Presence of surface exposed Trp residues is highly unlikely however, in the full length VarS histidine kinase surface exposed hydrophobic pocket could be buried. After being incorporated into the VarS histidine kinase domain I may have evolved to select for more hydrophobic residues in this pocket (see figure 4). This increase in hydrophobicity may allow this pocket in domain I to serve as a contact point for REC with the HATPase domain. Increased hydrophobic contacts between HisKA and REC domain may keep the VarS sensor histidine kinase folded correctly. Just like the increase in hydrophobic residues for domain I can be explained by REC being incorporated into the gene for VarS so to can the decrease in phosphate transfer motifs be explained. In domain I active site phosphate accepting Asp residue is not present. In domain II Asp140 is replaced with Glu7 in domain I while Asp141 is replaced with Pro8. Residues responsible for phosphate transfer in domain I may have been lost due to pressure to prevent phosphate transfer into domain I. If phosphate were to be transferred into domain I it may not be able to be affectively transferred further downstream.

Prevention of phosphate transfer would stop bacteria from responding to outside stimulus. For different bacteria this would lead to different outcomes. In *Vibrio alginolyticus* this would prevent the bacteria from efficiently degrading substrate stopping the uptake of necessary nutrients for survival. Understanding how evolution leads to differences between the two domains in REC is key to the development of strategies for crystallization of the domain.

Crystallization of REC domain is necessary for the development of broad-spectrum antibiotics that would target conserved structural motifs in REC domains from different species. Drug resistant "superbugs" are a growing problem due to incorrect use of anti-biotics. Recently developed anti-biotics all tend to target the same structures and pathways decreasing their effectiveness [**27**]. Gene expression analysis of drug resistant tuberculosis isolates have shown that two component response systems can enhance drug resistance [**28**]. Targeting certain two component systems may be an efficient way to increase drug sensitivity of bacteria.

5.4 Conclusion

In *V. alginolyticus*, Itasser derived homology model of REC domain reveals a structure with two sub-domains connected by a flexible linker. Circular dichroism and tryptophan emission spectra provided support to the model and limited proteolysis confirmed the presence of flexible linker. Based on the two-domain structure of REC a theory on the evolution of VarS incorporated REC domain from PleD-like response regulator via horizontal gene transfer and gene fusion was proposed. Based on its role in the REC-like adaptor domain, non-active site domain in REC may help promote dimer formation in VarS during signal activation of the sensor histidine kinase [**1**]. Crystallization efforts for REC domain have so far been unsuccessful. In order to crystallize REC domain flexible linker was removed. Expression plasmid containing REC sub-domain II (sub-domain II contains the active site) was transformed into *E. coli* and the protein was expressed. However, expression level for the protein was poor. If expression levels for REC sub-domain II cannot be improved REC domain with a shortened linker may need to be expressed [**16**]. Based on the REC homology model it is speculated that deletion of residues 117 to 129 from the linker should still allow the C-terminus of sub-domain I to be efficiently linked with the N-terminus of sub-domain II.

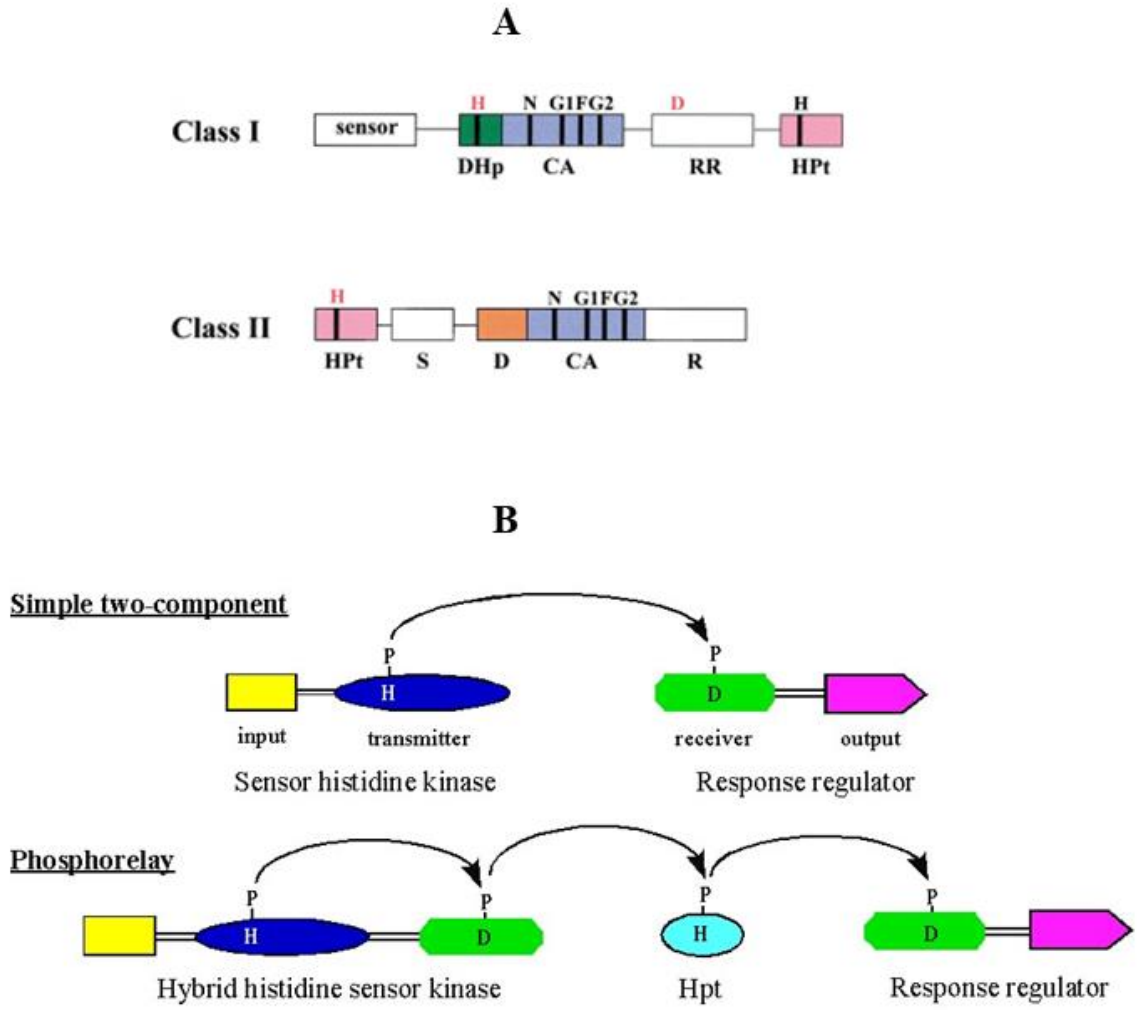


Figure 1: **A.** Cartoon representing two classes of histidine kinases. **B.** Cartoon showing the differences between simple two-component system and a phosphorelay.

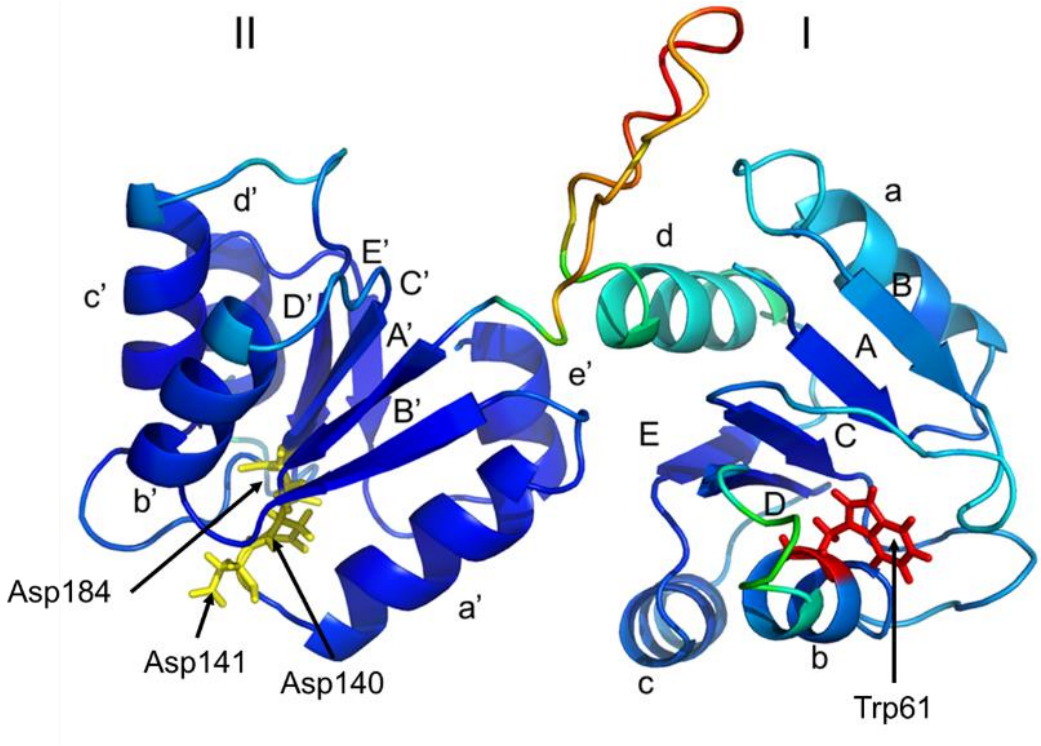


Figure 2: Itasser derived homology model of REC domain. B-factors are colored using PyMol (regions in blue are highly ordered, regions in red are highly disordered). Key Asp residues for Phosphorylation are colored yellow and surface exposed Trp61 is colored red.

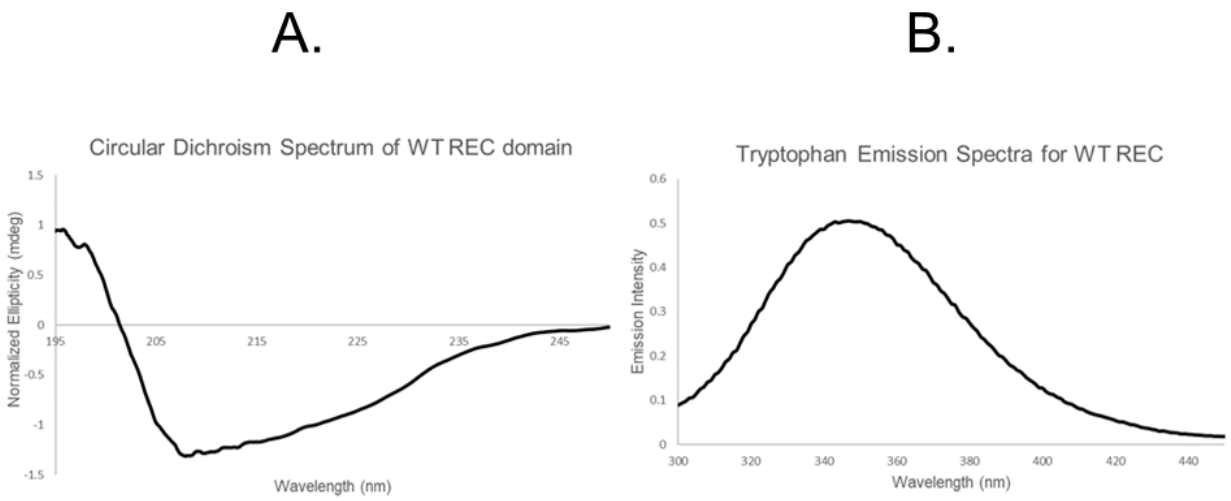


Figure 3: **A.** Circular dichroism spectrum for REC domain. A minimum at 208 nm and a maximum at 195 nm is consistent with a protein that contains both α-helices and β-sheets. **B.** Trp emission spectrum for REC domain. Maximum at 343 nm is most likely due to Trp residue being surface exposed

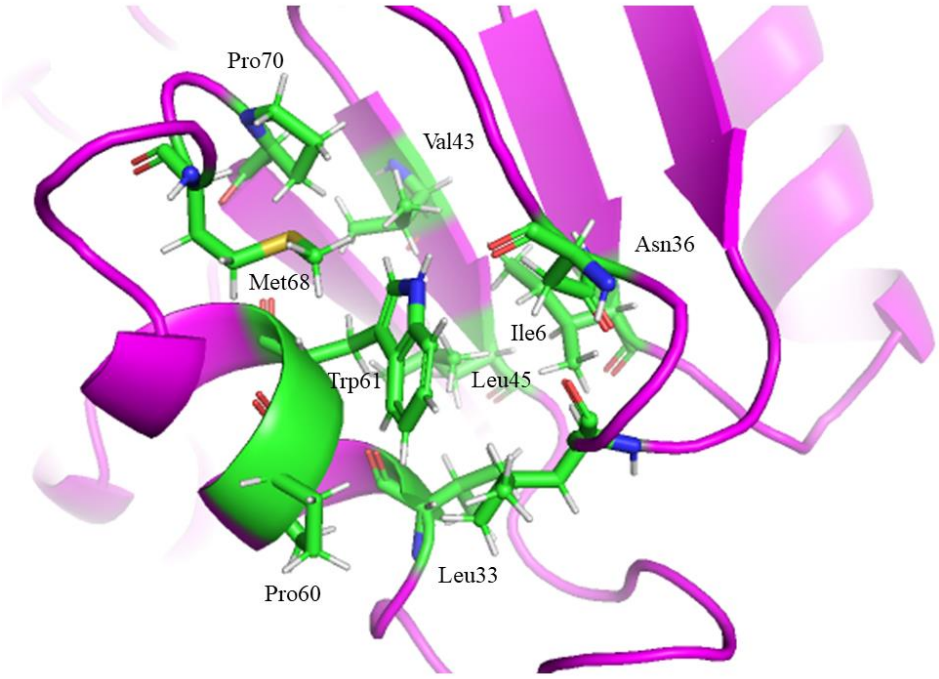


Figure 4: Hydrophobic pocket from REC sub-domain I. Pocket may give a contact point for the domain to be incorporated into full-length VarS.

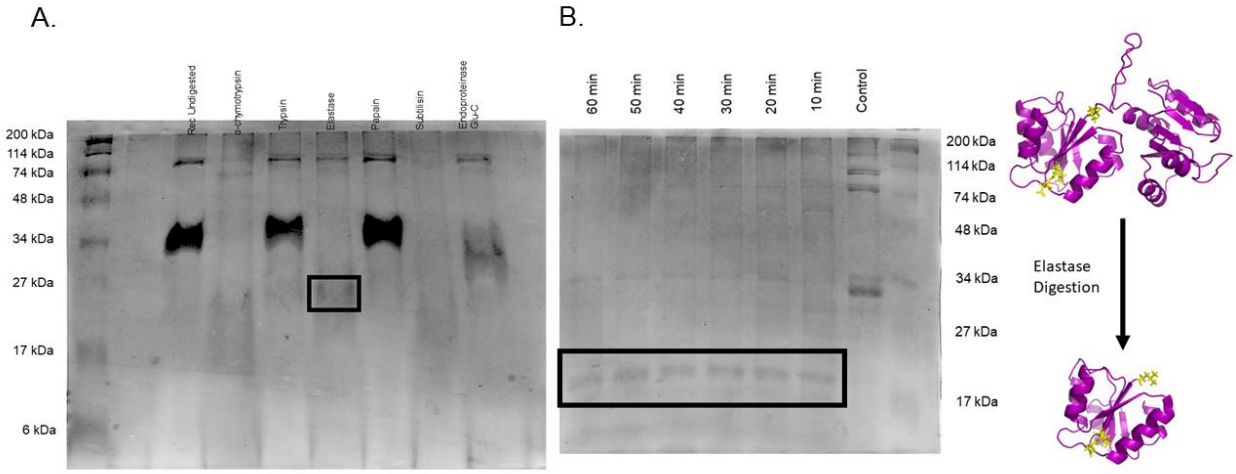


Figure 5: **A.** Initial proteolysis of REC domain using Proti-ace kite from Hampton research. Digestion from Elastase appears to produce a stable product. **B.** Time dependent proteolysis of REC domain using Elastase. Stable product appears after 10 mins.

A. Table summarizing DALI results

PDB Code	R.M.S.D. with Homology Model	# of domains	Color
3MF4	0.617 Å	2	Green
3LUF	0.619 Å	2	Blue
6IFH	0.817 Å	1	Salmon
2ZWM	0.751 Å	1	Orange
3T6K	0.828 Å	1	Black
4D6X	0.845 Å	1	Red
2A9O	0.893 Å	1	Cyan
6EBR	0.862 Å	1	Magenta
1YS6	0.920 Å	1	Brown
1NXS	0.927 Å	1	Deep Purple

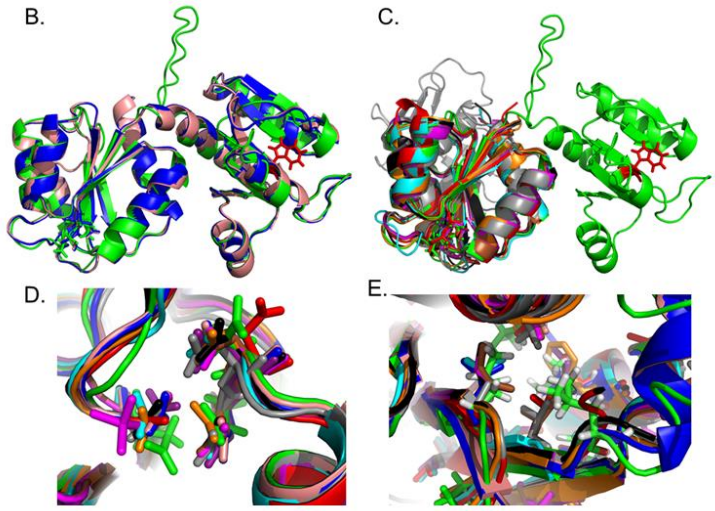
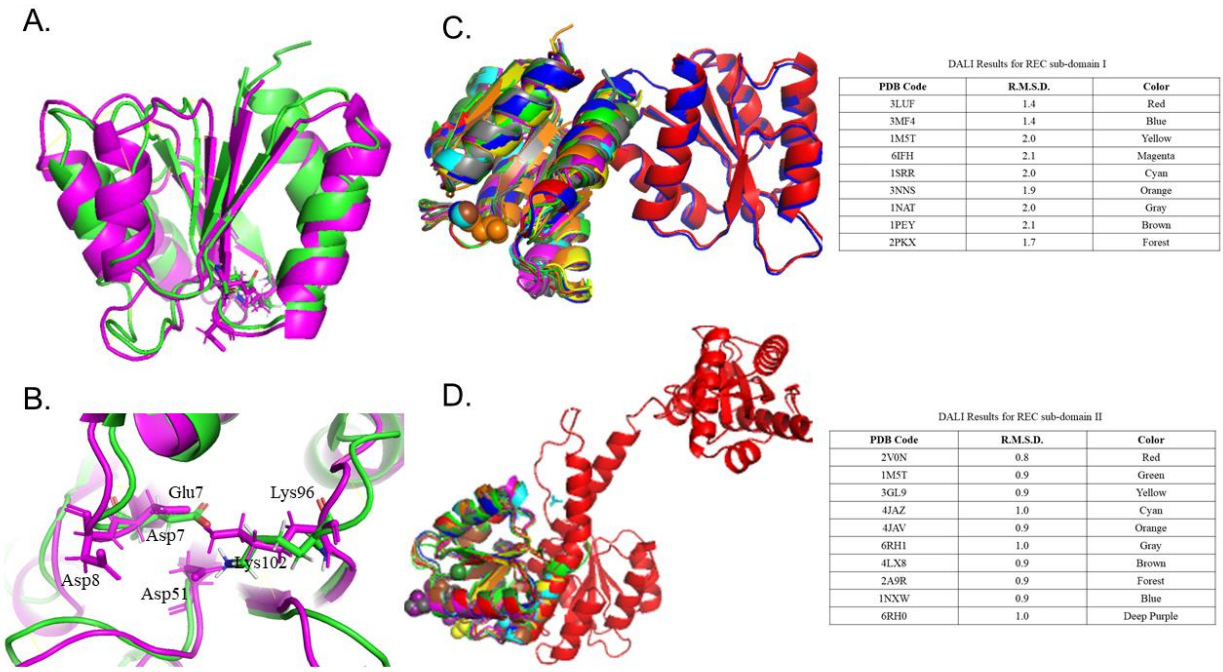


Figure 6: **A.** Table summarizing DALI results for REC domain **B.** REC domain homology model superimposed with multi-domain DALI structures. **C.** REC domain homology model superimposed with single domain DALI structures. **D.** Superimposed active site region for REC domain and all DALI results. **E.** Superimposed hydrophobic region at the N-terminal region of domain II for REC domain and all DALI structures.



DALI Results for REC sub-domain I

PDB Code	R.M.S.D.	Color
3LUF	1.4	Red
3MF4	1.4	Blue
1M5T	2.0	Yellow
6IFH	2.1	Magenta
1SRR	2.0	Cyan
3NNS	1.9	Orange
1NAT	2.0	Gray
1PEY	2.1	Brown
2PKX	1.7	Forest

DALI Results for REC sub-domain II

PDB Code	R.M.S.D.	Color
2V0N	0.8	Red
1M5T	0.9	Green
3GL9	0.9	Yellow
4JAZ	1.0	Cyan
4JAV	0.9	Orange
6RH1	1.0	Gray
4LX8	0.9	Brown
2A9R	0.9	Forest
1NXW	0.9	Blue
6RH0	1.0	Deep Purple

Figure 7: **A.** Superimposed structures for REC sub-domain I (green) and REC subdomain (II) magenta **B.** Active site region from A. zoomed in. Sub-domain is missing the majority of residues necessary for phosphate transfer **C.** Superimposed DALI results for REC sub-domain I. **D.** Superimposed DALI results for REC sub-domain II. Smaller R.M.S.D. values (in table) indicate REC sub-domain II is more closely related to REC common ancestor.

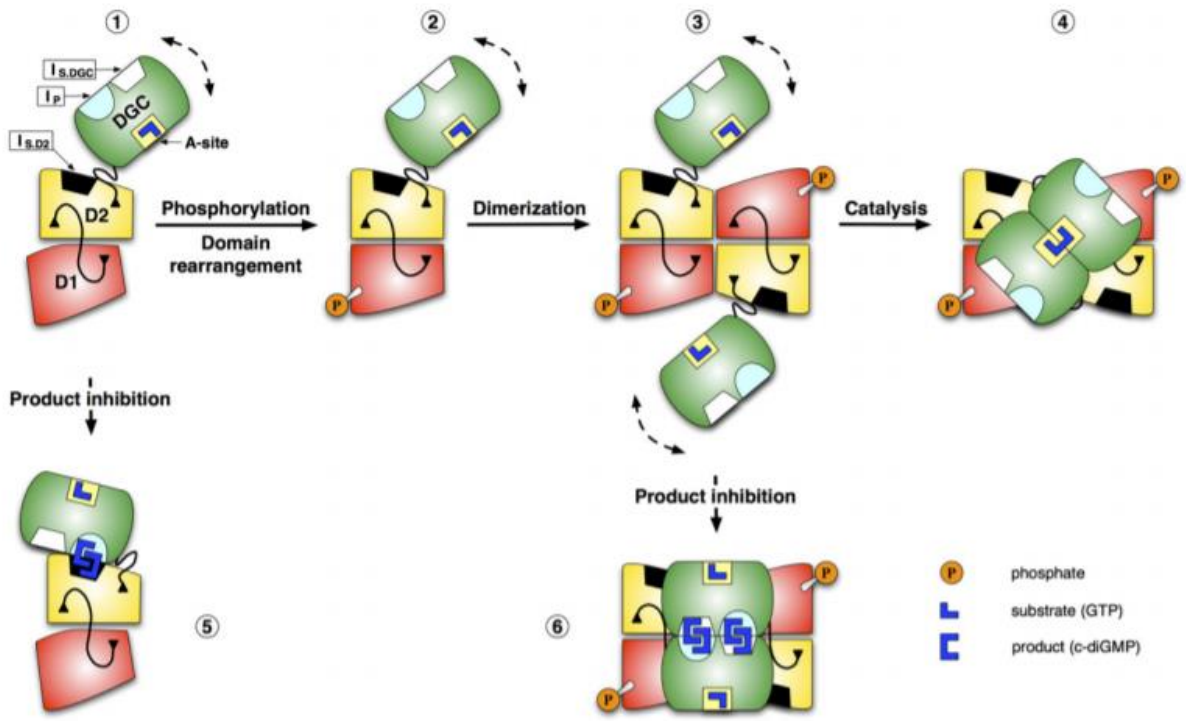


Figure 8: Cartoon representation of the role the adaptor domain plays in activating PleD-like response regulators. Adaptor domain was most likely kept after being incorporated into gene for VarS. Cartoon is from Wassman et al. 2007 [**1**].

Works Cited

1. P. Wassmann, C. Chan, R. Paul, A. Beck, H. Heerklotz, U. Jenal and T. Schirmer, *Structure of bef3- -modified response regulator pled: Implications for diguanylate cyclase activation, catalysis, and feedback inhibition*, Structure **15** (2007), no. 8, 915-927.

2. R. Gao and A. M. Stock, *Biological insights from structures of two-component proteins*, Annual review of microbiology **63** (2009), 133-154.

3. R. Dutta, L. Qin and M. Inouye, *Histidine kinases: Diversity of domain organization*, Mol Microbiol **34** (1999), no. 4, 633-640.

4. B. T. Nixon, C. W. Ronson and F. M. Ausubel, *Two-component regulatory systems responsive to environmental stimuli share strongly conserved domains with the nitrogen assimilation regulatory genes ntrb and ntrc*, **83** (1986), no. 20, 7850-7854.

5. M. Y. Galperin, *Structural classification of bacterial response regulators: Diversity of output domains and domain combinations*, **188** (2006), no. 12, 4169-4182.

6. E. Alm, K. Huang and A. Arkin, *The evolution of two-component systems in bacteria reveals different strategies for niche adaptation*, PLoS Comput Biol **2** (2006), no. 11, e143.

7. S. Pasek, J. L. Risler and P. Brézellec, *Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins*, Bioinformatics **22** (2006), no. 12, 1418-1423.

8. T. Mima, K. Gotoh, Y. Yamamoto, K. Maeda, T. Shirakawa, S. Matsui, Y. Murata, T. Koide, H. Tokumitsu and O. Matsushita, *Expression of collagenase is regulated by the vars/vara two-component regulatory system in vibrio alginolyticus*, J Membr Biol **251** (2018), no. 1, 51-63.

9. A. Roy, A. Kucukural and Y. Zhang, *I-tasser: A unified platform for automated protein structure and function prediction*, Nature protocols **5** (2010), no. 4, 725-738.

10. A. Dong, X. Xu, A. M. Edwards, C. Chang, M. Chruszcz, M. Cuff, M. Cymborowski, R. Di Leo, O. Egorova, E. Evdokimova, E. Filippova, J. Gu, J. Guthrie, A. Ignatchenko, A. Joachimiak, N. Klostermann, Y. Kim, Y. Korniyenko, W. Minor, Q. Que, A. Savchenko, T. Skarina, K. Tan, A. Yakunin, A. Yee, V. Yim, R. Zhang, H. Zheng, M. Akutsu, C. Arrowsmith, G. V. Avvakumov, A. Bochkarev, L.-G. Dahlgren, S. Dhe-Paganon, S. Dimov, L. Dombrovski, P. Finerty, S. Flodin, A. Flores, S. Gräslund, M. Hammerström, M. D. Herman, B.-S. Hong, R. Hui, I. Johansson, Y. Liu, M. Nilsson, L. Nedyalkova, P. Nordlund, T. Nyman, J. Min, H. Ouyang, H.-w. Park, C. Qi, W. Rabeh, L. Shen, Y. Shen, D. Sukumard, W. Tempel, Y. Tong, L. Tresagues, M. Vedadi, J. R. Walker, J. Weigelt, M. Welin, H. Wu, T. Xiao, H. Zeng, H. Zhu, G. Midwest Center for Structural and C. Structural Genomics, *In situ proteolysis for protein crystallization and structure determination*, Nature Methods **4** (2007), no. 12, 1019-1021.

11. L. Holm, *Dali and the persistence of protein shape*, Protein Science **29** (2020), no. 1, 128-140.

12. R. B. Bourret, *Receiver domain structure and function in response regulator proteins*, Current opinion in microbiology **13** (2010), no. 2, 142-149.

13. J. Yang, Y. Wang and Y. Zhang, *Resq: An approach to unified estimation of b-factor and residue-specific error in protein structure prediction*, J Mol Biol **428** (2016), no. 4, 693-701.

14. N. J. Greenfield, *Using circular dichroism spectra to estimate protein secondary structure*, Nature Protocols **1** (2006), no. 6, 2876-2890.

15. C. P. Pan, P. L. Muiño, M. D. Barkley and P. R. Callis, *Correlation of tryptophan fluorescence spectral shifts and lifetimes arising directly from heterogeneous environment*, J Phys Chem B **115** (2011), no. 12, 3245-3253.

16. Z. S. Derewenda and P. G. Vekilov, *Entropy and surface engineering in protein crystallization*, Acta Crystallogr D Biol Crystallogr **62** (2006), no. Pt 1, 116-124.

17. E. Kaufmann, N. Geisler and K. Weber, *Sds-page strongly overestimates the molecular masses of the neurofilament proteins*, FEBS Lett **170** (1984), no. 1, 81-84.

18. A. V. Finkelstein and J. Janin, *The price of lost freedom: Entropy of bimolecular complex formation*, Protein Eng **3** (1989), no. 1, 1-3.

19. C. W. Lee, S. H. Park, S. G. Lee, S. C. Shin, S. J. Han, H. W. Kim, H. H. Park, S. Kim, H. J. Kim, H. Park and J. H. Lee, *Crystal structure of the inactive state of the receiver domain of spo0a from paenisporosarcina sp. Tg-14, a psychrophilic bacterium isolated from an antarctic glacier*, J Microbiol **55** (2017), no. 6, 464-474.

20. I. Fernández, L. H. Otero, S. Klinke, M. D. C. Carrica and F. A. Goldbaum, *Snapshots of conformational changes shed light into the ntrx receiver domain signal transduction mechanism*, J Mol Biol **427** (2015), no. 20, 3258-3272.

21. C. J. Bent, N. W. Isaacs, T. J. Mitchell and A. Riboldi-Tunnicliffe, *Crystal structure of the response regulator 02 receiver domain, the essential yycf two-component system of streptococcus pneumoniae in both complexed and native states*, J Bacteriol **186** (2004), no. 9, 2872-2879.

22. I. Kufareva and R. Abagyan, *Methods of protein structure comparison*, Methods in molecular biology (Clifton, N.J.) **857** (2012), 231-257.

23. U. Jenal and M. Y. Galperin, *Single domain response regulators: Molecular switches with emerging roles in cell organization and dynamics*, Current opinion in microbiology **12** (2009), no. 2, 152-160.

24. E. J. Capra and M. T. Laub, *Evolution of two-component signal transduction systems*, Annual review of microbiology **66** (2012), 325-347.

25. Y. He, X. Liu, L. Ye, C. Pan, L. Chen, T. Zou and G. Lu, *Genome-wide identification and expression analysis of two-component system genes in tomato*, International journal of molecular sciences **17** (2016), no. 8, 1204.

26. P. J. A. Cock and D. E. Whitworth, *Evolution of prokaryotic two-component system signaling pathways: Gene fusions and fissions*, Molecular Biology and Evolution **24** (2007), no. 11, 2355-2357.

27. H. B. Felise, H. V. Nguyen, R. A. Pfuetzner, K. C. Barry, S. R. Jackson, M. P. Blanc, P. A. Bronstein, T. Kline and S. I. Miller, *An inhibitor of gram-negative bacterial virulence protein secretion*, Cell host & microbe **4** (2008), no. 4, 325-336.

28. G. Yu, Z. Cui, X. Sun, J. Peng, J. Jiang, W. Wu, W. Huang, K. Chu, L. Zhang, B. Ge and Y. Li, *Gene expression analysis of two extensively drug-resistant tuberculosis isolates show that two-component response systems enhance drug resistance*, Tuberculosis (Edinburgh, Scotland) **95** (2015), no. 3, 303-314.

Chapter 6: Potential role of conserved Cys in *Vibrio alginolyticus* HPT domain for regulation of collagenase expression under oxidative stress

Abstract

While attempting to study the structure for the cytoplasmic region of full length VarS histidine kinase (residues 197 – 932) it was discovered the protein was partially degraded and only appeared to be in its native state due to the presence of two disulfide bonds in the histidine kinase (Cys697/Cys789 and Cys883/Cys892). Due to the location of Cys883/Cys892 disulfide bridge within the HPT domain (residues 797 – 932) its potential role as a redox sensitive regulator for collagenase expression was investigated using homology modelin, Pz-peptidase assay and qRT-PCR. qRT-PCR revealed that sRNA2 transcription is only slightly downregulated in the presence of Menadione. Potential role of Cys883/Cys892 will need to be investigated further.

6.1 Introduction

Two-component systems serve as a basic stimulus response mechanism allowing organisms to sense and respond to different environmental changes. A brief outline of their mechanism is as follows; first a signal is received which causes the protein to dimerize, next ATP donates a phosphate to a conserved His residue, finally the phosphate is passed (either in a *cis* or *trans* method) to a conserved Asp residue located on the Response Regulator. Response regulator then leads to either the expression or repression of certain genes through transcription of sRNA which binds to repressor proteins [**3**]. VarS, histidine kinase pair from the VarS/VarA two component system, is a phosphor-relay type histidine kinase where the phosphate is passed through an internal REC domain and then a C-terminal HPT domain before finally passing it on to the VarA response regulator. The VarS/VarA system or closely related systems are used to

control virulence factors for a variety of bacteria. *V. cholera* uses a system closely related to VarS/VarA to control the expression of ToxR, a transcriptional activator that is essential for *V. cholera* virulence [**4**]. *S. aureus* uses the YycG/YycF two-component regulatory system to control the expression of genes involved in cell wall metabolism, membrane bound transport systems and pathogenicity [**5**].

The majority of HPT proteins are all α-helical and contain a conserved four helical bundle as the minimal core structure (see figure 1) [**2**] [**6**]. In most HPT domains this bundle is assembled as a monomer but in some cases, such as for ChpT from *C. crescentus* and Spo0B from *B. subtilis* it can be assembled as a dimer [**2**] [**7**]. The two forms do not have detectable sequence similarity and are most likely not evolutionary related but are examples of convergent evolution [**2**]. Outside of the conserved four helical bundle HPT domains can have different lengths or contain different structural elements [**2**]. The solvent exposed Histidine located on the second αB accepts phosphate from REC domain and transfer it to response regulator leading to gene transcription and protein expression. Below the general structure for phosphor-relay HPT domains is described.

Ypd1 from S. cerevisiae is a monomeric HPT domain used in regulating *S. cerevisiae* response to osmotic stress [**8**]. Due to the many studies on the structure of Ypd1 it is considered a model HPT protein [**2**] [**6**]. HPT from *V. alginolyticus* has sequence similarity with Ypd1 from *S. cerevisiae* indicating that its four-helix bundle is assembled as a monomer (see figure 1). Structure function analysis of Ypd1 indicate that several residues located near the key His residue that are important for phosphate transfer. Alanine substitution of Lys67 was found to have the greatest effect on phosphate transfer [**2**]. Basic amino acids are found to be highly conserved at this position in the HPT family. In REC domains conserved basic residues in the

active site help to coordinate phosphate binding by forming a salt bridge with oxygen from the phosphate [**6**]. Conserved basic residues may be playing the same role for HPT. Conserved Gly four residues from the phosphate accepting His residue towards the C-terminus was also found to be important for phosphate transfer. Mutation of this Gly residue to Gln results in reduced levels of phosphorylation [**9**]. Substitution of Gly68 from Ypd1 with bulkier residues inhibits transition state formation with response regulator Sln1 [**10**]. Crystal structure of Ypd1:Sln1 complexed with beryllium fluoride ion revealed that conserved hydrophobic and hydrogen bond interactions on three α-helices from the four helix bundle are necessary for forming the complex [**1**]. There is high sequence similarity at these same positions for the HPT domain of *V. alginolyticus*. Internalized REC and HPT domains do not have any effect on phosphate transfer or response regulator output but simply allow for extra regulatory checkpoints along the phosphorylation pathway. Intermediate REC and HPT domains can be targeted by specific phosphatases for an additional gene transcription regulation step [**6**].

During host invasion *V. alginolyticus* is thought to encounter oxidative stress environments, either from reactive oxygen species (ROS) from host immune system or H_2O_2 diffused from intestinal epithelial cells [**11**] [**12**]. Host immune systems generate ROS to kill invading pathogens [**11**]. ROS can damage bacteria in multiple ways; by disrupting oxidative phosphorylation, by damaging the bacterial membrane through lipid peroxidation, by oxidizing purines in DNA preventing replication and by protein carbonylation [**13**] [**14**] [**15**]. ROS sensing can be used by bacteria to adopt to and avoid host immune system [**16**]. To conserve energy needed for oxidative stress repair pathways bacteria may have evolved methods to sense intracellular ROS and up/down regulate expression of different proteins. Cys residues have high redox sensitivity which makes them excellent sensors for oxidative stress. In this chapter it is

discussed whether HPT disulfide bond between Cys883 and Cys892 discovered using mass spectrometry helps to downregulate expression of collagenase during oxidative stress. In certain response regulators intramolecular disulfide bond formation creates steric hindrance between the protein and DNA preventing sRNA transcription and perhaps a similar mechanism is occurring for VarS HPT [**3**] [**17**] [**16**]. Disulfide bond could also be intermolecular. In RegB from *R. capsulatus*, a histidine kinase responsible for many cellular processes, conserved Cys in dHp domain forms intermolecular disulfide bond in the presence of oxidizer. This Cys is highly conserved in RegB analogs however, Cys from RegB is not found in an analogous position in VarS (see figure 2) [**18**]. RegB is a simple histidine kinase while VarS is a phosphorelay-type histidine kinase. Currently it is unknown if this redox sensing Cys motif is present in phosphorelay-type histidine kinase.

6.2 Methods

6.2.a. Expression and Purification of VarS HPT domain and Cytoplasmic VarS

The gene for HPT has previously identified and isolated [**19**]. In order to construct the expression vector for HPT the genes was ligated into the pHSG398T plasmid using EcoRI and HinDIII. Plasmid was then transformed into *E. coli* BL21 cells for expression. Nickel-affinity chromatography was used to purify the expressed protein. Once purified the proteins were dialyzed into 1 x PBS for an extra round of purification as well as for storage.

The cytoplasmic region of VarS (residues 197-932) was expressed and purified as follows. A DNA fragment encoding $VarS_{197-932}$ was inserted in a pET-16b expression vector. Plasmid for N-terminal His10-tagged $VarS_{197-932}$ was transformed into *E. coli* BL21(DE3) cells and expressed by IPTG. Newly expressed $VarS_{197-932}$ was purified by affinity chromatography with Ni-NTA agarose. $VarS_{197-932}$ protein was further purified by ion exchange chromatography

with Q sepharose. Once purified, the protein was dialyzed into the following buffer for storage 30 mM HEPES pH 7.5, 50 mM KCl, 5 mM MgCl2, 1 mM DTT, 0.1 mM EDTA, 40% glycerol, 10% ethylene glycol. Proteins were shipped from Okayama University. Production and purification of all HPT and $VarS_{197-932}$ was accomplished by Dr. Takehiko Mima at Okayama University of Medicine since the Sakon lab at the University of Arkansas does not have access to the gene for VarS from *V. alginolyticus*.

6.2b. Itasser Homology Modeling of HPT

Due to difficulty in crystallization of HPT domain homology modeling was used to give a probable structure of the proteins based on sequence homology in proteins whose structures have been solved. Random mutagenesis using a transponson based system alongside sequence alignment studies were used to find the nucleotide and amino acid sequences for HPT (residue 797 to 932) [**19**]. Modeling was accomplished by inputting sequence into the Itasser (Iterative Threading Assembly Refinement) program developed by the Zhang lab at the University of Michigan [**20**]. Homology model for REC sub-domain II containing linker to HPT domain was produced with Itasser using REC sub-domain II sequence with linker to HPT to identify if Cys697/Cys789 disulfide bond is possible *in vivo*.

6.2c. Circular Dichroism and Tryptophan Emission Spectra of HPT Domain

Circular Dichroism and Tryptophan Fluorescence emission were used to check if expressed HPT secondary structure and fold matched the Itasser homology models. For both Circular Dichroism and Fluorescence experiments proteins were dialyzed into 30 mM Hepes pH 7.5, 50 mM KCl, 1 mM TCEP. For Circular Dichroism experiments best results are obtained when the protein is in a low salt buffer. For the CD experiment proteins were diluted with the

new buffer to 1.5 ml with a final concentration of 0.2 mg/ml. For Tryptophan emission experiments proteins from the CD experiments were diluted with buffer as needed. Circular Dichroism and Tryptophan emission spectra of HPT was found using a Jasco CD spectrometer at room temperature.

6.2d. Identifying Apparent Molecular Weight of $VarS_{197-932}$

Size exclusion chromatography (G-200 sephadex) with reference proteins were used to determine the apparent molecular weight of $VarS_{197-932}$. 1 g of G-200 sephadex was mixed with 60 ml of elution buffer (30 mM Hepes pH 7.5 50 mM KCl) and swollen at 90° C for 5 hrs. The slurry was degassed by bubbling with inert N_2 gas for 10 min and gravity flow was used to pack the column. Blue Dextran (MW: 2,000 kDa) was eluted to find the void volume of the column. Gravity filtration was used to run the following reference proteins through the column; Bovine Gamma Globulin (MW: 150 kDa), Bovine Serum Albumin (MW: 66.5 kDa) and Lysozyme (MW: 14 kDa). All reference proteins were eluted through the column in triplicate. $VarS_{197-932}$ was eluted through column normally and after being incubated with TCEP. $VarS_{197-932}$ was treated with TCEP by incubating equal volumes of protein with 10 mM TCEP for 1 hr. at 4°C. As the protein elutes from the column 0.5 ml volumes were collected and a micro-bradford assay was used to determine when each protein was eluted. Absorbance at 625 nm was used to reveal when blue dextran was eluted. Log (MW) vs Ve/Vo was plotted, and the linear trend line was used to find the apparent molecular weight of $VarS_{197-932}$.

6.2e. Identification of VarS$_{197-932}$ disulfide bonds using trypsin digestion coupled with MALDI – TOF

12% SDS PAGE gels were used to check if addition of TCEP causes VarS$_{197-932}$ dimer to dissociate into the monomer. Protein samples were run in the absence and presence of .2 M TCEP. For the reduced sample equal volumes of 2 mg/ml VarS$_{197-932}$ and .2 M TCEP were incubated together at room temperature for 30 minutes before electrophoresis. In the non-reduced sample .2 M TCEP was replaced with an equal volume of water.

Dr. Rohanna Liyanage in the mass spectrometry department at the University of Arkansas performed all trypsin digestion and mass spectrometry analysis. 5 µl of VarS$_{197-932}$ was digested in its native state (without reducing agent) by adding 10 µl Mass Spectrometry grade trypsin (Promega) at a concentration of 10 ng/µl in ammonium carbonate buffer. Mixture was incubated overnight at 37°C. After incubation 100 µl of 5% Formic Acid (FA) in 60% acetonitrile was added to quench trypsin and break down ammonium bicarbonate. Digest mixtures were then dried and reconstituted in 50 µl of 0.1% FA. A HPLC-20A/LC-30A, coupled to a Shimadzu electro spray ionization quadrupole ion trap-time of flight mass spectrometer (LC-ESI-IT-TOF-MS) was used to obtain high resolution accurate intact masses of the digestions with less than 10 ppm error. The digest was subjected to LC separations using a Bio wide Pore C18 reverse phase column (4.6 mm × 15 cm, 5 µm) (Supelco, St. Louis, MO) with a solvent flow rate of 0.8 mL/min in 0.1% formic acid/acetonitrile gradient of 5–100%, over a 60 min period. Shimadzu LC-MS solution version 3.81 was used to process peaks to obtain mass spectra for each chromatographic peak. Then the monoisotopic mass obtained from the isotopically resolved high resolution multiple charged ions were matched against the Insilco tryptic digest masses

obtained by Bruker Daltonics sequence editor software version 3.2 (Bruker Daltonics GmbH) based on all possible disulfide linkages.

6.2f. Identification of conserved structural and sequential elements in related HPT domains

To identify if structural elements within the VarS HPT homology model are conserved homology model was loaded onto DALI server and matched against the entire Protein Databank. Non-redundant structures with the top ten Z-scores were chosen and aligned with the HPT homology model using PyMol to identify key structural elements. The presence of Cys residues were noted for each structure from DALI. If a structure contained multiple conserved Cys residues the distance between Cα's was noted. To identify conserved sequence motifs for the DALI structures FASTA sequences were aligned using Clustal omega online server [**21**]. The positions of conserved hydrophobic, basic and acidic residues near the active site His residue were noted [**2**]. The positions of Cys residue for each FASTA sequence were also noted.

6.2g. Effect of VarS Cys883Ser mutation on Pz-peptidase activity

All experiments regarding the effects of VarS C883S mutation on Pz-peptidase activity were performed by Dr. Takehiko Mima at Okayama University. A gene expression vector possessing wild type VarS and a gene expression vector possessing the mutant VarS C883S were transformed into a *V. alginolyticus* VarS deletion mutant. A gene expression vector possessing no VarS gene was used as a negative control. Expression vector for VarS C883S was constructed using site directed mutagenesis with expression vector for wild type VarS being used as the template.

The primers used for mutagenesis were (5'-TCATATTCCGGAGTGCCTAGGCTGAA) and (5'-CACTCCGGAATATGAGCTACTGCCGT) with underlined regions corresponding to

the new S883 residue. Introduction of mutation was confirmed by Sanger sequencing. Newly transformed cells were grown in NZ1/2SW medium (1% NZ amine, 1.67% sea salts) containing 0.2% gelatin at 25°C. Expression of collagenase was induced by adding 50 µl of overnight culture to each well of a 96 well plate containing 50 µl of an insoluble collagen solution and incubating at 37°C for 2 hrs [**19**]. After induction collagenase activity of bacterial culture was measured colorimetrically using 4-phenylazobenzyloxy-carbonyl-Pro-Leu-Gly-Pro-D-Arg (Pz-peptide, Sigma-Aldrich) as described by Wünsch and Heidrich [**22**].

6.2h. Effect of Oxidative stress on expression of sRNA2

To measure the potential effect of oxidative stress on the activity of *V. alginolyticus* wild type cells were grown as previously described [**19**]. When the OD_{600} of culture reached 0.5, 25 µM menadione or 2 mM H_2O_2 were added, and the cultures were incubated for an additional 2 hrs. or 7 min. respectively. Total RNA was isolated using RNeasy Protect Bacteria Mini kit (Qiagen), and reverse-transcribed with a ReverTra Ace qPCR RT Master Mix (TOYOBO, Japan) according to the manufacturers' instructions. The relative sRNA quantity was measured using a KAPA SYBR Fast qPCR kit (NIPPON Genetics, Japan). qRT-PCR was carried out using a CFX Connect Real-Time PCR Detection System (BioRad), and data were analyzed using the supplied software. Expression levels for sRNA2 were normalized to that of 23S rRNA. (5'-AAGCCTAATGGACGAGGTATCTTCA-3') is the sRNA2 forward primer used and (5'-ATCCTGACGGGCGAACTCTATC-3') is the reverse primer used. 23S rRNA forward primer (5'-TTACCTACTCTTGACATC-3') and (5'-GACTTAACCCAACATTTC-3') 23S rRNA reverse primer were used to quantify the amount of 23S rRNA. All experiments regarding the effect of oxidative stress on sRNA2 expression were performed by Dr. Takehiko Mima at

Okayama University of Medicine as the University of Arkansas does not have access to qRT-PCR thermocycler.

6.3 Results and Discussion

6.3a. VarS cytoplasmic domains are cleaved off during expression from E. coli

Initially, size exclusion chromatography was used to identify if the full length VarS histidine kinase undergoes dimerization *in vitro* similar to other histidine kinases [**23**]. Due to the molecular weight of the $VarS_{197-932}$ (monomer: 82 kDa, dimer: 164 kDa) 1 g of G-200 sephadex was swollen in 30 mM Hepes pH 7.5, 50 mM KCl. Blue dextran was eluted through the column in triplicate and the void volume (Vo) of the column was found to be 4.5 ml. Bovine Gamma Globulin (MW: 150 kDa), Bovine Serum Albumin (MW: 66.5 kDa) and Lysozyme (MW: 14 kDa) were also eluted in triplicate and elution peaks were averaged for each standard to give the average elution volume (Ve). Log(MW) vs. Ve/Vo for the three standards was plotted to make a calibration curve with the equation $y = -0.8476x + 6.3074$ where y is the Log(MW) of the standards and x is the Ve/Vo for each standard. For $VarS_{197-932}$ elution peak for the protein was found at 4.5 ml (see figure 3A). When this value is entered into the calibration curve equation an apparent molecular weight of 257 kDa is given. To investigate whether disulfide bond formation is responsible for soluble aggregate formation $VarS_{197-932}$ was treated with TCEP and chromatography experiment was repeated in duplicate. For TCEP treated $VarS_{197-932}$ elution peak was found at 5 ml which gives an apparent molecular weight of 204 kDa (see figure 3B). SEC derived molecular weights are inaccurate when compared to known molecular weight calculated from Expasy ProtParam tool (monomer: 82 kDa, dimer: 164 kDa). 0.5 ml fractions collected for both native and TCEP treated $VarS_{197-932}$ were most likely too large to give an accurate

molecular weight for $VarS_{197-932}$ and the standards. However, this experiment was able to reveal that the protein undergoes a conformational change after treatment with TCEP.

In order to investigate this conformational change SEC elution fractions from the reduced protein were ran on a 12% SDS PAGE gel alongside native $VarS_{197-932}$ and TCEP treated $VarS_{197-932}$ (see figure 4A). Native $VarS_{197-932}$ sample ran as a single band with a MW of approximately 80 kDa while the addition of TCEP caused the appearance of three bands with approximate MW of 80, 58 and 25 kDa. TCEP treated SEC $VarS_{197-932}$ samples from 4.5, 5.0 and 5.5 ml fractions all ran as single bands with MW of approximately 58 kDa. While it was not discovered if full length VarS is a dimer *in vitro* it was discovered that $VarS_{197-932}$ was partially degraded during its expression from *E. coli* and that intramolecular disulfide bonds were keeping the protein together (see figure 4B). For a protein that exist in the cytosol, a primarily reducing environment, the ability to form disulfide bonds is an intriguing discovery [**24**].

6.3b. ESI Mass Spectrometry analysis revealed Cys residues involved in formation of disulfide bridges

To determine the location of disulfide bonds within $VarS_{197-932}$, ESI mass spectrometry experiment was planned. $VarS_{197-932}$ was digested by trypsin in its native state and the presence of disulfide bridges were confirmed by mass spectrometry analysis. After LC-MS, chromatographic peaks were processed to obtain the monoisotopic masses for multiple charged ions. The experimentally obtained masses were matched against *in silico* obtained digest masses from Bruker daltonics sequence editor version 3.2 based on all possible disulfide linkages. ESI mass spectrometry experiment was performed as a collaboration by our lab and Dr. Rohana

Liyanage in the statewide mass facility at the University of Arkansas. Mass spectrometry analysis was able to confirm the presence of two disulfide bridges in native $VarS_{197-932}$.

The first disulfide bridge was found to be formed between Cys697 in the REC domain and Cys789 in the linker region between REC and HPT domains. Region containing disulfide bridge was hydrolyzed with trypsin at Arg690, Lys710, Arg758 and Lys795. Chromatographic peak corresponding to the two tryptic peptides was analyzed and mass spectra for +6 and +7 charged ions (m/z) were used to confirm the existence of this disulfide bridge. +6 and +7 spectra were chosen as they were the easiest to interpret. The percent difference between experimental and theoretical mass spectra for +6 and +7 ions were found to be less than 0.1% (see figure 5 and 6).

The second disulfide bridge was found to be formed between Cys883 and Cys892 in the HPT domain. Region containing disulfide bridge was formed by hydrolysis with trypsin by at Lys875, Lys887, Lys889 and Arg900. Chromatographic peak corresponding to the two HPT tryptic peptides was analyzed and mass spectra for +3 and +4 charged ions (m/z) were used to confirm the existence of this disulfide bridge. +3 and +4 spectra were chosen as they were the easiest to interpret. The percent difference between experimental and theoretical mass spectra for +3 and +4 ions were also found to be less than 0.1% (see figure 7 and 8).

Linker database reveals that Cys residues are unlikely to be present in linker region of multi-domain proteins. Itasser derived homology model of REC domain containing the additional residues from the linker region revealed that Cα distance between Cys697 and Cys789 is approximately 25 Å which is not within the appropriate distance for disulfide bond formation (see figure 9) [**25**]. Since Cys697/Cys789 distance is being estimated from homology model the position of linker α-helix could be incorrect and may be much shorter. Using nomenclature from

chapter 3, loop connecting β-sheet E to linker α-helix may be oriented so that linker α-helix comes closer to Cys789. While this may be occurring *in vitro*, it would not be beneficial for the enzyme if this disulfide bond is occurring *in vivo*. Cys789 is located near the active site of REC domain and if linker α-helix were brought near enough that Cys697/Cys789 disulfide bond could form it would potentially block active site from accepting phosphate. Disulfide bond could be the result of intermolecular interactions. Sensor histidine kinases form dimers both *in vivo* and *in vitro* [**26**] [**23**]. Dimer is formed primarily by hydrophobic interactions and currently it is unknown if these interactions are strengthened by disulfide bond formation [**27**]. Based on the location of both Cys residues lying within the HPT sub-domain and their proximity to the key His residue in the primary structure of HPT the disulfide bond between Cys883 and Cys892 was analyzed further (see figure 10).

6.3c. Description of Itasser derived HPT domain homology model

Similar to REC domain, troubles with crystallization of HPT forced the use of homology modeling to derive a structure for the domain. Sequence for HPT was uploaded to the Itasser server and model was computed based on crystal structures of homologous proteins [**20**]. The Itasser derived homology model of HPT revealed a structure that consists of five α-helices; the core of the protein is arranged in a four helix bundle motif with the remaining two α-helices at the N-terminus of the model (see figure 11A). Four helix bundle motifs are highly conserved amongst HPT domains [**2**]. 20 residue N-terminus of the homology model is predicted to be highly flexible based on Cα B-factors [**28**]. Some hydrophobic regions in the N-terminus are predicted to be solvent exposed. For crystallization of HPT domain to be successful, truncation of N-terminus may be required [**29**]. Other than the N-terminus the model is predicted to be quite ordered. In the homology model solvent exposed His874 is located on α-helix c with key

phosphate coordinating Lys875 and hydrophobic residues necessary for interaction with REC domain are also on this helix. Hydrophobic and polar residues necessary for complex formation with REC are conserved throughout helices b and d as well [**1**]. Cys883 is located on the loop connected α-helix c to α-helix d and Cys892 is located on α-helix d. In homology model disulfide bond between Cys883 and Cys892 is not predicted to occur.

6.3d. Circular Dichroism and Tryptophan Emission Spectra support HPT homology model structure

Circular dichroism (CD) and tryptophan emission spectra were used to support the HPT homology based structure. Raw data for circular dichroism spectrum was normalized and smoothed using exponential smoothing function in excel. Noise at low wavelengths due to excess salt in the buffer was minimized from this method however, data below 195 nm was still too noisy to be read. Ellipticity from HPT spectrum decreased rapidly until reaching 220 nm and then the ellipticity began to decrease slowly until a minimum was reached at 206 nm. After the minimum was reached ellipticity began to increase until reaching the maximum at 195 nm (see figure 11B). HPT CD spectrum is indicative of a protein that is primarily α-helical which is consistent with known structure for other HPT domains and with the homology model [**30**] [**2**].

Since HPT domain contains only a single Trp residue the Trp emission spectrum was used to determine if the protein was folded identical to the Itasser derived homology model. Absorbance wavelength was 280 nm and the emission wavelength for the protein peaked at 340 nm. Similar to REC domain the emission spectrum is consistent with Trp27 being at least partially exposed to the solvent (see figure 11C) [**31**]. In the homology model Trp27 is located on α-helix b near the potentially dynamic loop connecting α-helix a to α-helix b. Hydrophobic interactions between Trp27, Ala30, Leu31, Ala41, Lys42, Leu45, Tyr133 and Leu134 and

electrostatic interaction between Trp27 side chain nitrogen and side chain carboxy group from Glu38 and form a pocket that may need to be stabilized for potential crystallization attempts of HPT.

6.3e. Sequence and Structural alignment reveal that Cys892 is highly conserved across HPT domains

To determine if Cys883 and Cys892 are conserved in structural homologs of HPT homology model Itasser derived HPT domain was loaded onto DALI server and matched against the entire Protein DataBank [**32**]. The structures with the top 10 Z-scores were chosen and aligned using PyMol (see figure 12A and B). Structures from DALI search are listed in table 1. Chosen structures from DALI were found to have lower percent sequence identity that those chosen from the non-redundant database. Structures were found to align well with the homology model all with R.M.S.D. less than or equal to 3.2 Å. For the DALI structures percent sequence identity with HPT homology model was not higher than 20% with the exception of *E. coli* HPT protein from BarA (45%) and *Shewanella* HPT domain (38%). For the itasser HPT homology model the flexible N-terminus is found only in two of the DALI structures, that for *S. cerevisiae* Ypd1 (PDB: 5KBX) and that for *E. coli* HPT protein from BarA (PDB: 3IQT). When searching the homology model across the entire Protein Databank Cys883 and Cys892 were found to be only partially conserved. Some structures contain both Cys residues at locations identical to that of the homology model. Most of the structures contain only one or no conserved Cys residues. Clustal omega server was used to align the FASTA sequences from DALI structures with the sequence for VarS HPT domain (see figure 12C) [**33**]. In structures with only one Cys residue it was found that the position for Cys892 was more conserved than that for Cys883. The position

of key Gly residues as well as hydrophobic, polar, basic and acidic residues near phosphorylated His were noted (see figure 12C).

6.3f. Potential cytoplasmic Cys883/Cys892 disulfide bond does not affect Pz-peptidase activity

To investigate if the disulfide bond between Cys883 and Cys892 in VarS could modulate the function of VarS Dr. Takehiko Mima at the University of Okayama mutated Cys883 to Ser and Pz-peptidase activity of *V. alginolyticus* mutant was checked compared to the wild type [**19**]. Experiment was performed at the University of Okayama due to our lab not having the proper materials for growing mutant *V. alginolyticus* strains and being closed due to Covid-19. Overnight cultures for wild type and mutant *V. alginolyticus* were incubated with insoluble collagen solution to induce expression of *Vibrio* collagenase [**19**] [**34**]. Pz-peptidase activity of newly expressed collagenase was measured over 25 hrs. for each culture [**19**] [**22**]. *V. alginolyticus* culture containing expression vector with no VarS gene did not express collagenase. However, for *V. alginolyticus* containing expression vector for wild types and Cys883Ser mutant Pz-peptidase activity was found to be roughly the same for both (see figure 13). These results appear to indicate that disulfide bond between Cys883 and Cys892 most likely does not exist in native *V. alginolyticus* and potential Cys883/Cys892 disulfide bond is not involved in collagenase expression.

6.3g. Menadione effects the expression of sRNA2 responsible for Vibrio collagenase expression

Results from previous section revealed that disulfide bond between Cys883 and Cys892 is not necessary for the catalytic mechanism behind phosphate transfer in VarS. However, Cys residues may still serve as regulatory sensors for oxidative stress in the cell [**35**]. To test if collagenase expression pathways are turned off in the presence of oxidative stress *V.*

alginolyticus wild type cells were grown by Dr. Takehiko Mima at the University of Okayama as previously described and incubated normally or with oxidizing agent (25 mM menadione or 2 mM H_2O_2) [**19**]. qRT-PCR was used to quantify the amount of transcribed sRNA2. Like section 4.6, experiment was performed at University of Okayama as our lab does not have access to wild type *V. alginolyticus* cells or to a real time PCR detection system. It was found that addition of 25 mM menadione leads to an approximate 30% down regulation in sRNA2 transcription levels while addition of 2 mM H_2O_2 does not cause any change in sRNA2 transcription (see figure 12). Unexpectedly it was also discovered that sRNA2 transcription is turned on for *V. alginolyticus* in just L-broth. More experiments will need to be performed to determine if trace elements in L-broth are responsible for activating VarS sensor histidine kinase.

6.3h. Protein disulfide isomerases are not necessary for correctly folding HPT

The presence of disulfide bonds in the region of VarS that is expected to be in the cytoplasm is an unexpected finding. Cytosol in prokaryotes and eukaryotes is a highly reducing environment which normally prevents disulfide bond formation [**36**]. Reducing environment in bacterial cytosol is mostly maintained by superoxide dismutase, glutathione peroxidase, thioredoxin/thioredoxin reductase pathway and the glutathione/glutaredoxin pathways. Pathways most likely evolved to prevent reactive oxygen species (ROS) damage within the cell Superoxide dismutase and glutathione peroxidase work to directly reduce ROS to water [**37**]. Thioredoxin and glutathione work to reduce cytosolic protein disulfide bonds from ROS. After reducing disulfide bond conserved Cys residues are oxidized and must be reduced by thioredoxin and glutaredoxin before they can act on another protein [**36**].

While Cys residues can be present in proteins within the reducing cytosol of bacteria usually they are involved in oxidative stress repair pathways similar to thioredoxin/thioredoxin

reductase and glutathione/glutaredoxin. For some proteins, disulfide bond formation is necessary for allowing the protein to fold properyly [**38**]. Protein disulfide isomerases (PDI) are proteins that catalyze the formation of disulfide bonds necessary for proteins secreted into the extracellular space [**39**]. While primarily they exist in the periplasm, some PDIs catalyze the formation of disulfide bonds from proteins located in the cytoplasm [**40**] [**41**]. Pz-peptidase assay for *V. alginolyticus* Cys883Ser revealed that a potential cytoplasmic disulfide bond between Cys883 and Cys892 in VarS is not important for maintaining function to regulate expression of collagenase (see figure 13). Since function is maintained for both the wild type and mutant *V. alginolyticus* structure function relationship suggests Cys883 and Cys892 do not form a disulfide bond in native *V. alginolyticus* and PDIs are not needed in folding of the HPT subdomain of VarS. It could be that Cys883 and Cys892 may be forming a disulfide bond only *in vitro*. However, it seems unlikely that evolution would allow for VarS to conserve Cys residues that play no discernable role in protein function. VarS must have been evolutional pressured to keep Cys883 and Cys892 indicating that these residues must play a critical role in the histidine kinase. For multiple redox proteins and transcription factors Cys residues serve as oxidative stress sensors and lead to promotion of redox processes within the cell [**35**] [**36**] [**16**] [**42**] [**17**]. For VarS Cys883 and Cys892 may play the same role.

6.3i. Cys883/Cys892 disulfide bond partially downregulates sRNA2 expression during oxidative stress

Previously it was thought that two component systems and oxidative stress sensing pathways are independent events [**16**]. However, Cys residues, with their high redox sensitivity, within sensor histidine kinases could potentially serve as internal oxidative stress sensors [**43**]. In the presence of oxidative stress disulfide bond formation may stop or slow down phosphate

transfer in histidine kinase allowing the bacteria to focus more energy towards turning on oxidative stress repair pathways. Bacteria using disulfide bonds to sense oxidative stress is not without precedent. In *S. aureus* transcription of *arg* gene leads to the expression of quorum sensing AgrC histidine kinase as well as its response regulator AgrA and the precursor to AgrC quorum signal AgrD [**44**] [**45**]. ArgA is activated by autoinducing peptide which leads to AgrA being phosphorylated. Once phosphorylated, AgrA regulates transcription of genes encoding different metabolic factors as well as phenol-soluble modulin (PSM) peptides including virulence δ-toxin [**16**] [**42**]. PSM peptide expression is controlled by AgrA in an sRNA independent manner. It is speculated that AgrA binds directly to δ-toxin promoter region, however this has yet to be demonstrated [**46**]. AgrA also controls the transcription of RNAIII. RNAIII is a sRNA that is responsible for controlling the expression of additional virulence factors in *S. aureus* via interactions with *rot* repressor protein [**42**]. Molecular simulations revealed that in oxidative stress conditions the presence of intramolecular Cys199/Cys228 disulfide bond present in AgrA prevents the response regulator from interacting with gene for coding for RNAIII leading to downregulate of *S. aureus* virulence factors [**16**]. Mutation of Cys199 to Ser was found to make AgrA more susceptible to oxidative stress (see figure 15) [**16**]. AgrA has been previously shown to downregulate the expression of oxidative stress defense enzyme glutathione peroxidase either directly or through RNAIII transcription [**47**]. In the presence of H_2O_2, AgrA being unable to interact with gene for RNAIII leads to the upregulation of glutathione peroxidase making *S. aureus* more resistant to oxidative stress [**16**]. Analogous Cys199 and Cys228 residues are found to be conserved among AgrA homologs indicating that this method for sensing oxidative stress is conserved in some response regulators [**16**].

In *V. cholerae* response regulator AphB responsible for the expression of key virulence factors also contains key Cys residues that allow the bacteria to respond appropriately to oxidative stress conditions [**17**]. Under oxygen-limiting conditions *V. cholerae* virulence genes are found to be highly expressed [**17**]. In AphB, Cys235 was found to be the key residue in responsible for allowing the protein to respond to oxidative stress. Mutation of Cys235 to Ser made AphB active even in highly aerobic conditions [**17**]. Dimerization is known to be important for the activity of AphB and that Cys235Ser mutant was able to form a dimer under both aerobic and anaerobic conditions. While oxygen reacting at Cys235 was shown to inhibit dimer formation. Liquid chromatography coupled with tandem mass spectrometry revealed that Cys235 forms a disulfide bond with either Cys76 or Cys94 during aerobic conditions. In the compact crystal structure of AphB the distance between the Cαs of Cys235 and Cys94 and Cys235 and Cys94 were found to be too large to indicate disulfide bond formation [**48**]. Currently it is unknown how disulfide bond formation may occur in AphB. Addition of DTT to WT AphB was found to promote dimer formation and DNA binding activity of AphB [**17**]. Similar to AgrA with gene for glutathione peroxidase, active AphB was found to repress the gene for *ohrA* which encodes for a specific organic peroxide-specific peroxiredoxin by binding directly to promoter region [**49**].

Similar to *S. aureus* and *V. cholerae, V. alginolyticus* most likely encounters oxidative stress during the later portions of infection and due to this VarS/VarA two component system may have evolved the same catalytic machinery to regulate gene transcription in the presence of oxidative stress [**49**]. While oxidative stress regulation mechanisms are present in response regulators AgrA and AphB; response regulator VarA contains only a single Cys residue at position 43 and based on AgrA and AphB having multiple Cys residues is likely not involved in

oxidative stress-based transcription regulation. It is known that some sensor histidine kinases are specifically activated by ROS [**50**] [**51**]. However, non ROS activated histidine kinases with redox regulatory control motifs are less common. Below we describe a potential mechanism for oxidative stress-based transcription regulation in the histidine kinase portion of a two-component system. Depending on the stress encountered, bacteria may have evolved different regulation mechanisms based on the most encountered stressors for the bacteria.

DALI results indicate that potential mechanism for gene regulation in the presence of oxidative stress may be conserved in some histidine kinase HPT sub-domains. Structure of HPT from *Shewanella* (PDB:3MYF) and HPT from *E. coli* BarA histidine kinase (PDB:3IQT) show that two Cys residues are in the same location as the HPT homology model Cα distance for both structures is approximately 11 Å not within the appropriate distance for disulfide bond formation. However, piston-like movements common in four helix bundle motifs may position the two Cys residues within 2 Å for formation of disulfide bridge [**25**]. HPT domains from Maize (PDB:1WN0), *M. truncatula* (PDB:3US6) (PDB:4G78) have Cys892 conserved near the HPT active site and multiple others clustered nearby all within potential disulfide bond length. Location of the specific Cys residues involved in disulfide bond formation may be different for each species. For protein with multiple Cys residues SS bonds are not listed. It may be that disulfide bond formation is intermolecular and not intramolecular as only Cys892 is highly conserved throughout the DALI structures. In *R. capsulatus* sensor histidine kinase RegB plays a role in control of many cellular processes. It is a simple histidine kinase that has evolved an Arg-Cys-Arg motif downstream from dHp domain that forms an intermolecular disulfide bond in oxidizing conditions. Intermolecular disulfide bond converts RegB from active dimer to an inactive tetramer [**18**]. Switch to from dimer to tetramer may disrupt binding interface between

CA and dHp domain. Something similar may be occurring for VarS HPT domain where intermolecular disulfide bond disrupts binding interface between REC and HPT. More experimentation will be needed to determine if disulfide bond is intra or intermolecular.

For VarS, we speculate that Cys883/Cys892 disulfide bond inhibits phosphate transfer similar to how disulfide bond present in AgrA or AphB inhibits DNA binding [**16**] [**17**]. In the presence of oxidative stress, phosphate transfer may be stopped or slowed by disulfide bond formation. Based on comparison with crystal structure of YPD1:SNL1 complex with beryllium fluoride and Mg^{2+} the active site interface between HPT domain and REC consists of hydrophobic and hydrophilic interactions from α-helices b, c, d and e from the HPT homology model and α-helices a' and e' from loop between α-helix e' and β-sheet E' and loop between β-sheet B' and α-helix b' from the REC homology model [**1**]. Intra/intermolecular disulfide bond formation between Cys883 and Cys892 may partially disrupt the interface between HPT and REC domain which may partially inhibit phosphate transfer between REC and HPT (see figure 16).

Partial disruption of the REC:HPT interface would explain the qRT-PCR results where transcription of sRNA2 is only partially downregulated when *V. alginolyticus* culture is incubated with 25 μM menadione prior to RNA isolation (see figure 14). The difference in sRNA2 transcription after addition of 25 μM menadione or 2 mM H_2O_2 to *V. alginolyticus* culture may be due to the different lifetimes of their initial redox biproducts. Menadione is a quinone that undergoes catalytic electron redox cycling to reduce molecular O_2 to superoxide anion which has a half-life of 1 to 15 mins. While H_2O_2 reacts with metals to produce metal or hydroxyl radicals which have half-lives of about 10^{-9} s [**52**]. In addition to its shorter half-live, H_2O_2 is known to react slowly with free thiols and is driven by the pKa of sulfhydryl group from

the thiol [**53**]. pH of VarS during qRT-PCR experiment is 7.5 which mean that only a small portion of sulfhydryl groups are deprotonated. While H_2O_2 is a stronger oxidizing agent than superoxide anion its short half-life may explain its inability to affect sRNA2 transcription in *V. alginolyticus*.

VarS/VarA two component system directly regulates the transcription of sRNA2. Newly transcribed sRNA2 interacts with RNA binding protein CsrA which is bound to mRNA for Col A *vibrio* collagenase. Removal of CsrA allows for the gene encoding Col A collagenase to be expressed [**3**]. Slowing down sRNA2 transcription in the presence of oxidative stress would slow down expression of *vibrio* collagenase and potentially allow for gene for oxidative stress repair pathways to be turned on. Currently it is unknown if sRNA2 represses the expression of oxidative stress repair enzymes similar to *S. aureus* with RNAIII and glutathione peroxidase [**16**]. Future experiments would be needed to confirm if this is the case.

6.4 Conclusion

For HPT domain, Itasser derived homology model gives a structure that is primarily α-helical containing the common four helical bundle consistent with all HPT domains. Structural comparison of Itasser derived homology model with results from DALI search revealed that active site Cys residues close to phosphate accepting His are relatively conserved throughout HPT domains. Pz-peptidase assay for VarS Cys883Ser reveals that potential cytoplasmic disulfide bond is not necessary for promoting phosphate transfer. Based on the presence of Cys in a region of the protein that is cytoplasmic as well as oxidative stress based mechanism for shutting off response regulators it is speculated that these residues play a role in allowing VarS to sense oxidative stress and stop phosphate transfer/gene transcription in its presence. qRT-PCR showed that sRNA2 transcription was only slightly downregulated in the presence of menadione

and not affected at all in the presence of H_2O_2. Further experiments will need to be performed to see if this potential oxidative stress gene regulation system is present for VarS.

For HPT, homology model presents a structure with a flexible N-terminus. Truncation of the 20 residue N-terminus may lower entropy of the protein enough so that crystallization becomes a spontaneous reaction [**29**]. To investigate the potential role of active site Cys residues for HPT the potential role of sRNA2 in repressing oxidative stress repair pathways will need to be determined. For fully active AgrA and AphB response regulators repression of glutathione peroxidase and peroxiredoxins was seen. The same correlation may be seen for VarS/VarA two-component system. To further understand the roles Cys play in sRNA2 transcription qRT-PCR expression may need to be repeated with Cys mutated to Ser. Mutating both Cys to Ser would prevent disulfide bond formation from occurring in the protein regardless of oxidative stress. For wild type VarS, in the presence of 25 μM menadione sRNA2 transcription was downregulated however, it is unknown if Cys residues played a role in stopping transcription. If sRNA2 transcription for the VarS Cys883Ser Cys892Ser in the presence of 25μM menadione is identical to the wild type it would reveal that Cys residues are necessary for shutting down sRNA2 transcription in the presence of oxidative stress.

Chapter 6: Figures

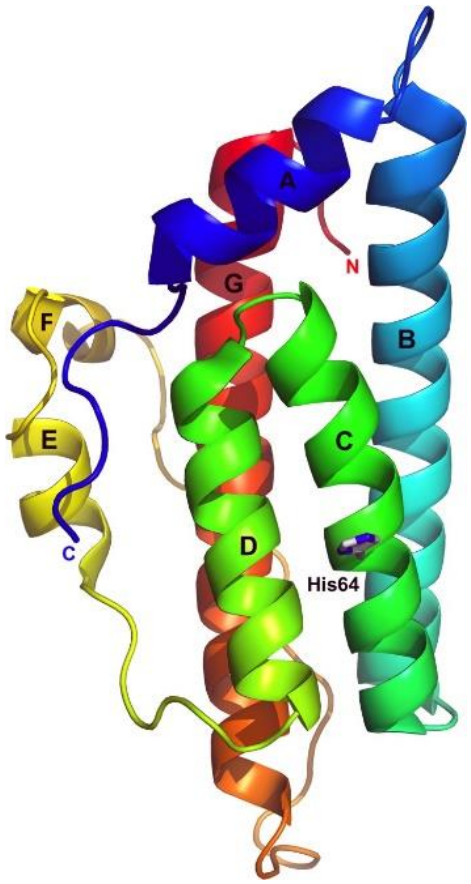


Figure 1: Crystal structure of Ypd1. Structure for Ypd1 is well known so it is considered a model structure for HPT domains. Primary structural motif is the four helix bundle *(αB-αC-αD-αG)*. Phosphorylated His residue is shown as a stick figure. α-helices A, B, C and D are responsible for interacting with the response regulator Sln1. Figure is from Fassler et al. 2013

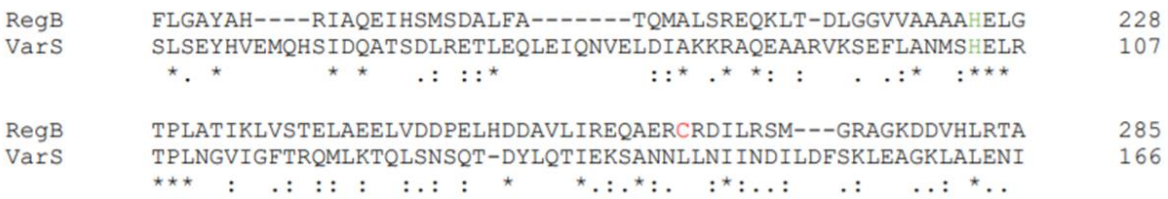


Figure 2: Sequence alignment of RegB with VarS. Phosphate accepting His is in green while RegB Cys is in red. Cys present in HPT domain of VarS may play the same role as Cys in RegB.

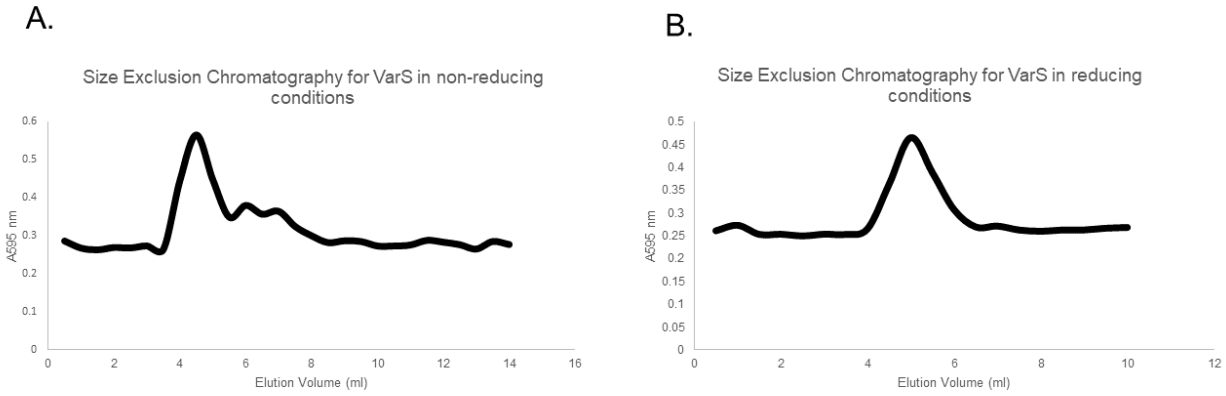


Figure 3: **A.** SEC results for VarS sensor histidine kinase in buffer without reducing agent. **B.** SEC results for VarS after treatment with TCEP. Shift in elution peaks indicates structural change that occurs upon addition of a reducing agent

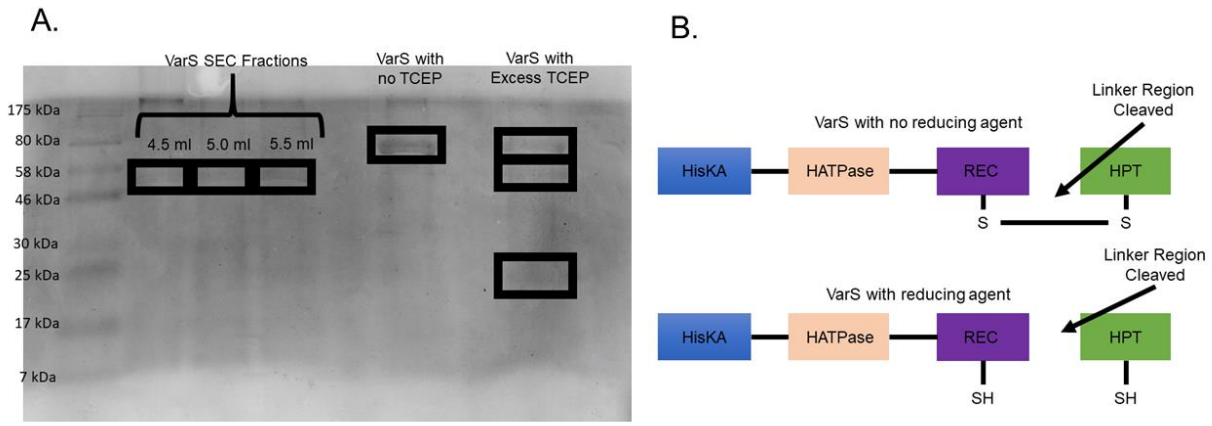


Figure 4: **A.** SDS PAGE results for TCEP treated VarS SEC fragments. VarS with and without TCEP were included as controls. Reducing the protein with TCEP causes it to break into 3 bands **B.** Cartoon representation of explanations for TCEP treated VarS.

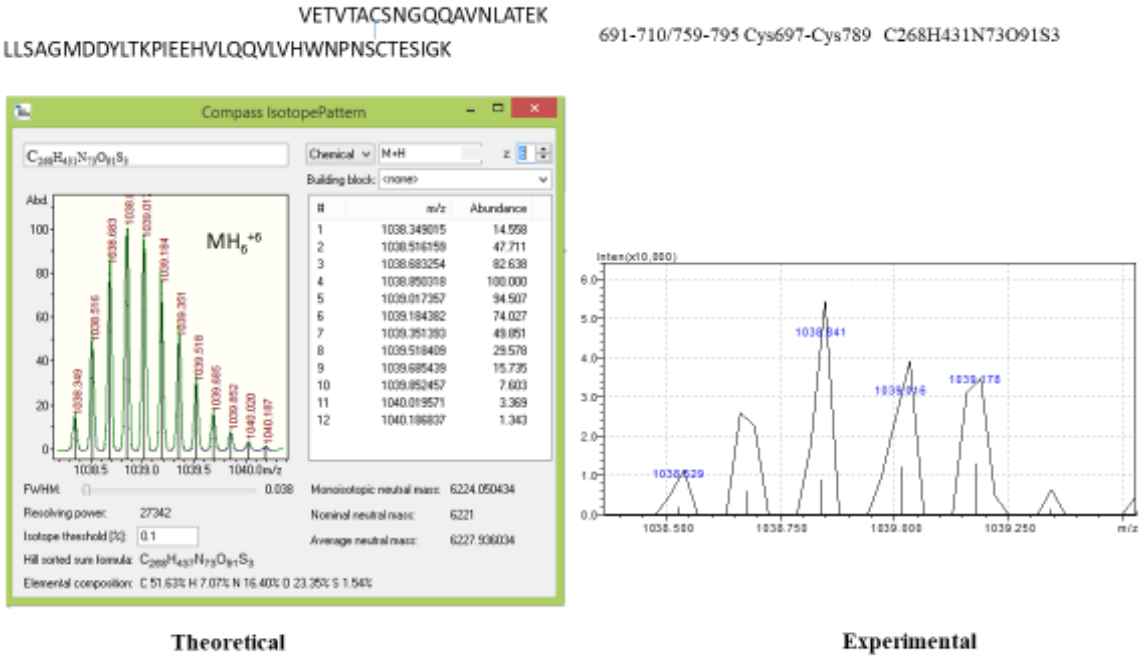
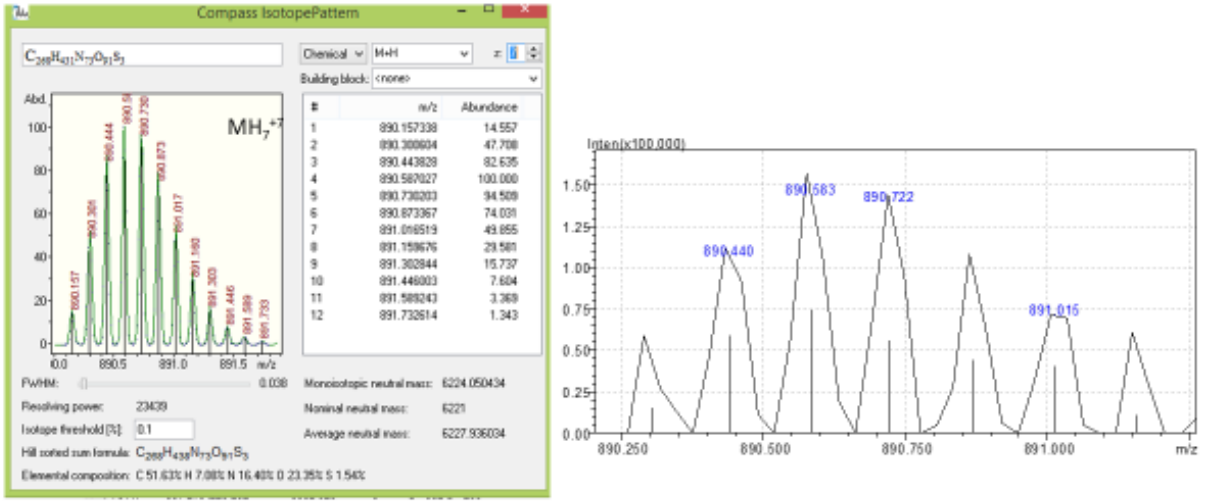


Figure 5: Mass spectra used to confirm +6 charged disulfide bridged trypsin digest fragments 691-710/759-795 from VarS.

VETVTACSNGQQAVNLATEK 691-710/759-795 Cys697-Cys789 C268H431N73O91S3

LLSAGMDDYLTKPIEEHVLQQVLVHWNPNSCTESIGK



Theoretical **Experimental**

Figure 6: Mass spectra used to confirm +7 charged disulfide bridged trypsin digest fragments 691-710/759-795 from VarS.

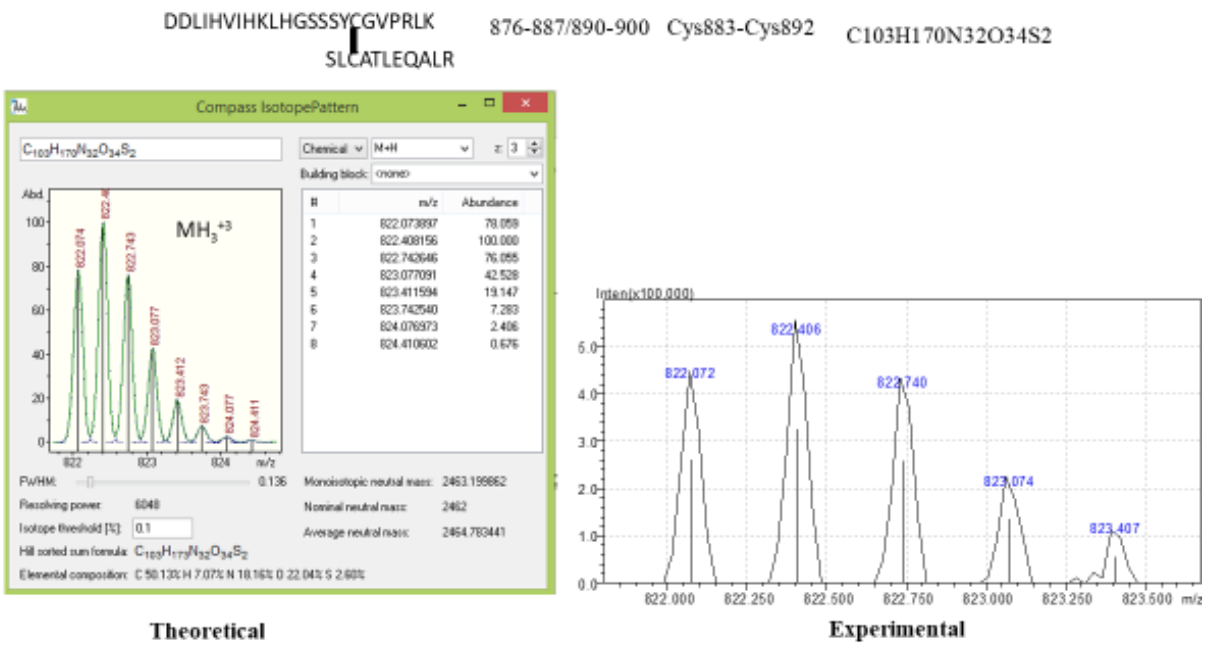


Figure 7: Mass spectra used to confirm +3 charged disulfide bridged trypsin digest fragments 876-887/890-900 from VarS.

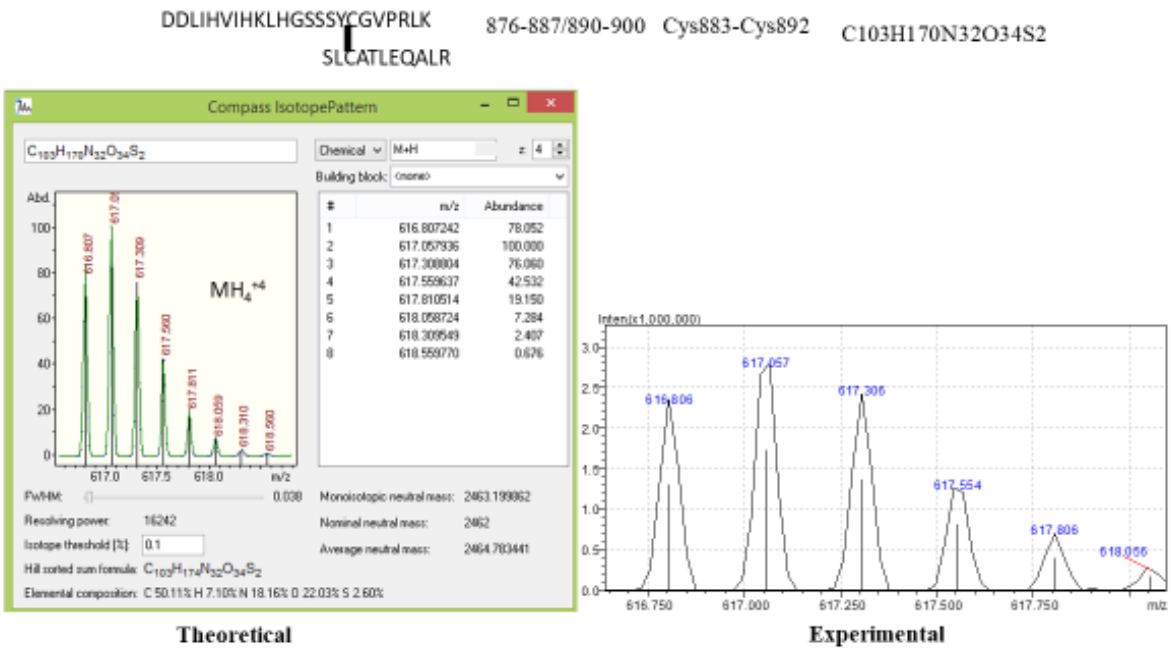


Figure 8: Mass spectra used to confirm +4 charged disulfide bridged trypsin digest fragments 876-887/890-900 from VarS.

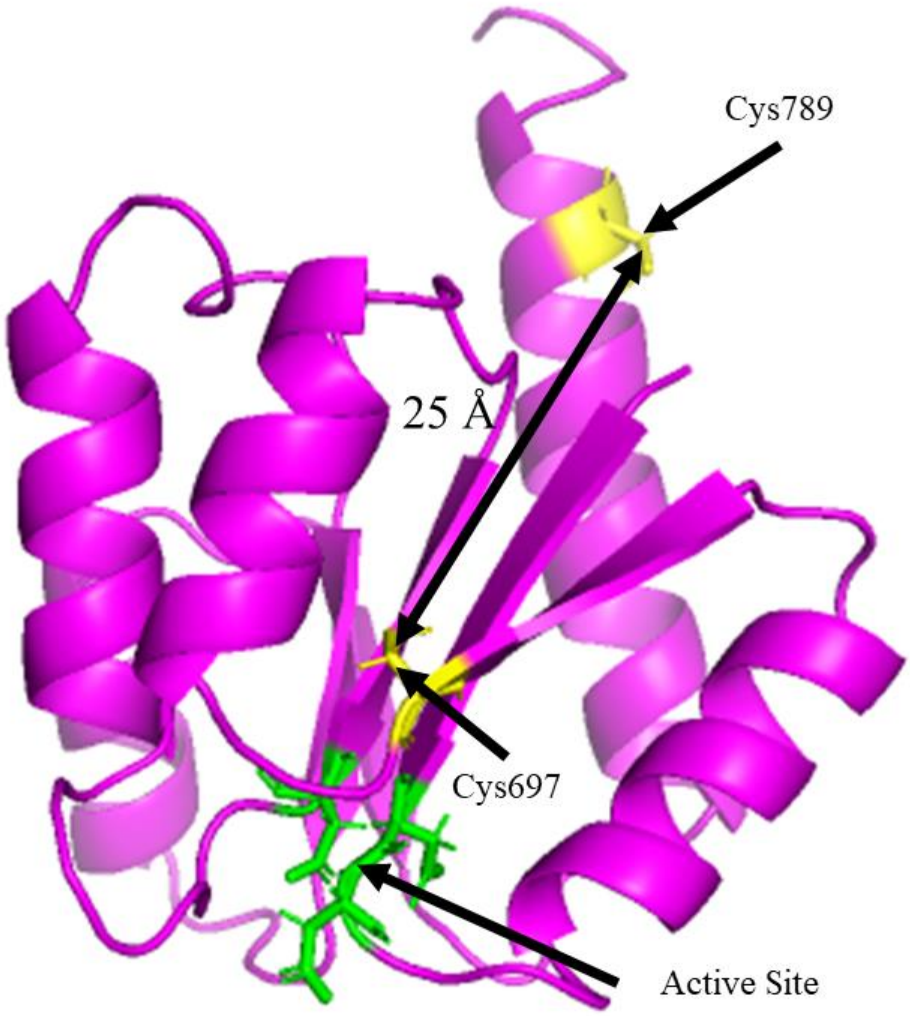


Figure 9: Homology model of REC sub-domain II containing linker region to HPT. Distance between Cys697 and Cys789 in the model is approximately 25 Å which is way longer than disulfide bond in nature. Position of linker containing Cys789 may not be accurate and it may be closer to Cys697. Repositioning linker so that it is closer would bring it closer to the active site. Repositioning linker closer to active site would not make sense as it may inhibit phosphate transfer. Cys697/Cys789 disulfide bridge may only be present *in vitro*.

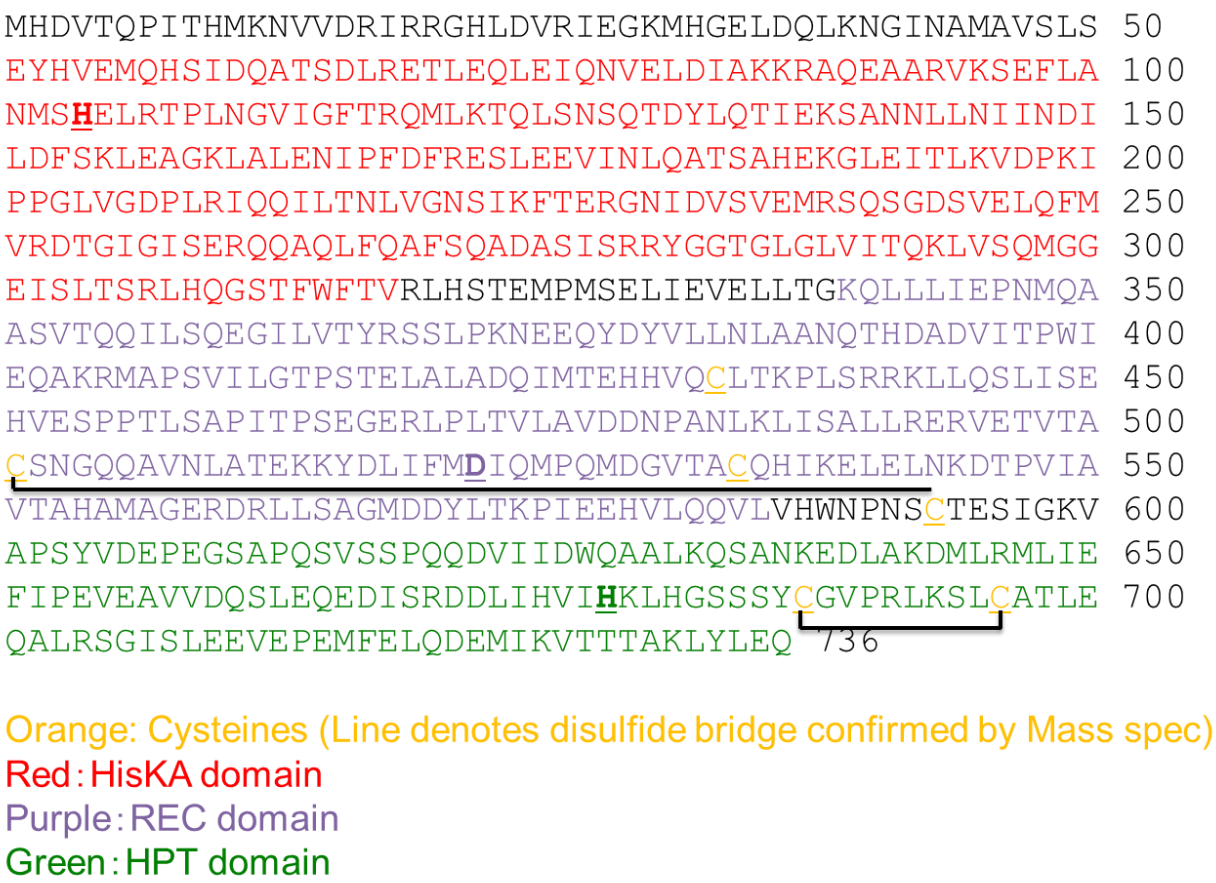


Figure 10: Sequence for cytoplasmic portion of VarS sensor histidine kinase. Cys residues and disulfide bridges are denoted. Due to its close proximity to phosphate accepting His residue (underlined) the disulfide bond between Cys883 and Cys892 was chosen for further study.

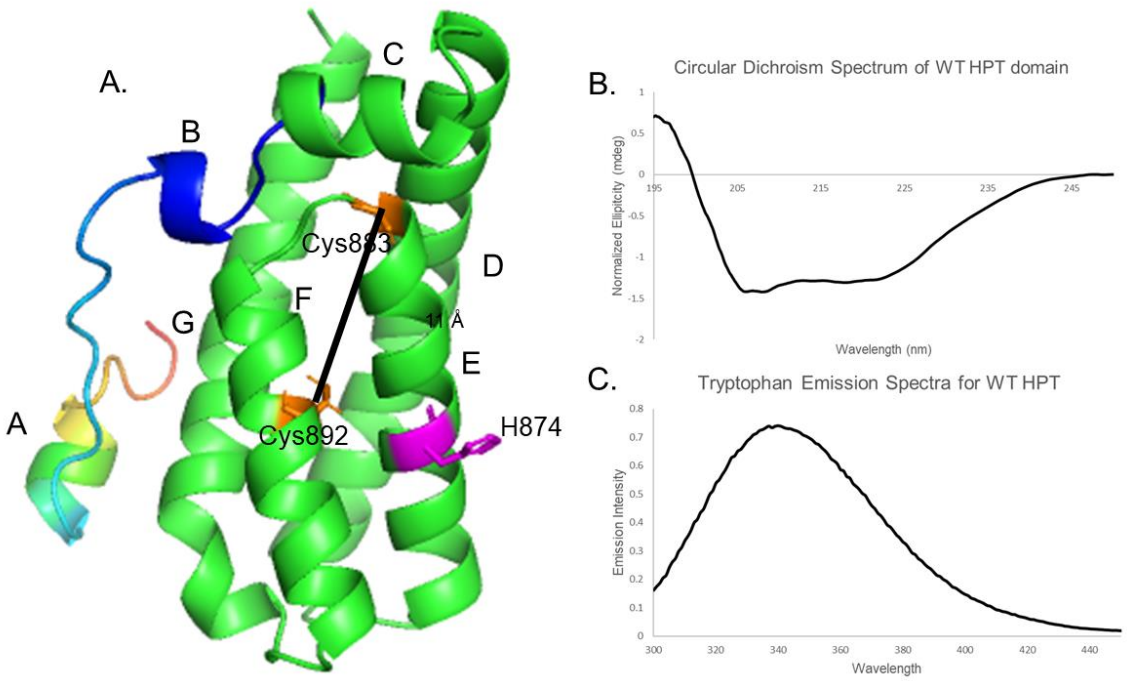
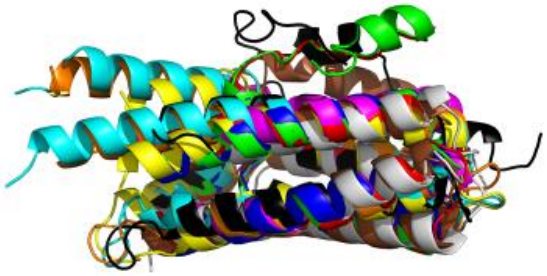


Figure 11: **A.** Homology model of VarS HPT domain derived from Itasser. Phosphate accepting His is shown in purple while Cys residues are orange. Four helix bundle is consistent with HPT domains in other species. **B.** CD spectrum of wild type HPT domain. Minima at 225 nm and 205 nm are consistent with a protein that is primarily α-helical. **C.** Tryptophan emission spectra for wild type HPT. Max at 340 nm indicates that Trp residue at N-terminus must be solvent exposed.

A.

Model	Color	# of active site Cys	R.M.S.D.
HPT homology model	Green	2	N/A
3IMF	Blue	2	0.7
3IQT	Red	3	1.3
1WN0	Yellow	2	2.5
3US6	Orange	2	2.6
1BDJ	Magenta	0	3.1
4G78	Cyan	2	2.9
5KBX	Black	1	3.2
1TQG	Gray	1	2.9
1I5N	Wheat	0	1.8
3KY1	Brown	1	2.4

B.



C.

```
VarS HPT                                        AVVDQSLEQ---EDISRD---DLIHVIHKLHGSSSYCGVPRLKSLCATLEQALRS----G 106
Shewanella Hybrid Histidine kinase HPT          EKIQTALGQ---N--DQA---TMLSTIHKLHGASCYCGVPTTQRLCQEIESALKR----Q 85
E. Coli BarA HPT                                NKVEEQLVG---E--NPE---GLVDLIHKLHGSCGYSGVPRMKNLCQLIEQQLRS----G 94
M. truncatula MtHPT1                            KDLSFAVDQ---QSIDFK---KVDAHVHQFKGSSASIGAQRVKNSCVAFRNFCEEQNIDA 111
M. truncatula MtHPT2                            NNMSRALEQ---VPVNFK---QIDAHAHQQKGSSASVGAARVKNVCGTFRNFCEAQNLEG 112
S. cerevisiae HPT signalling protein YPD1        AQMQRQLDG---E-KNLT---ELDNLGHFLKGSSAALGLQRIAWVCERIQNLGR------ 90
Maize ZmHP2                                     SELAALLDQ---PIVDFD---KVDAYVHQLKGSSASVGAQKVKFTCMQFRQLCQDKNRDG 112
T. maritima CheA phosphotransferase domain      QNLNDTLLELEKNPEDMELINEAFRALHTLKGMAGTMGFSSMAKLCHTLENILDKARNSE 78
E. Coli ArcB HPT                                SVLESNLTA---Q--DKK---GIVEEGHKIKGAAGSVGLRHLQQLGQQIQSPDLP----- 91
R. sphaeroides Putative histidine protein kinase DAMEASLLALQAGEDAAAHVGPLFRAVHTFKGNSRVLGLSVVESRAHLCEDLIGLVRDAG 84
S. typhimurium CHEMOTAXIS PROTEIN CHEA      ADMEQHLLDLVPESPDAEQLNAIFRAAHSIKGGAGTFGFTILQETTHLMENLLDEARRGE 80
                                                :  :                  * :* .  *          ..
```

Figure 12: **A.** Table summarizing DALI results **B.** Homology model for HPT superimposed with DALI matches **C.** Sequence alignment of results from B using Clustal omega. Cys residues are colored red, key His is blue, key Gly are green, key basic residues are orange, key hydrophobic residues are purple and key acidic residues are pink. Cys892 appears to be far more conserved Cys883.

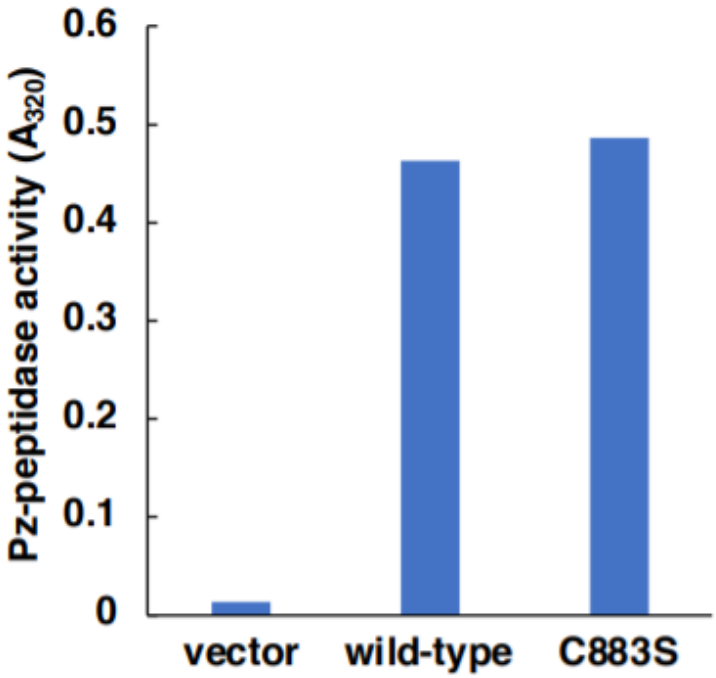


Figure 13: Pz-peptidase activity for *V. alginoltyicus* with wild type VarS, C883S VarS and no VarS (negative control). Potential cytoplasmic disulfide bridge between Cys883 and Cys892 does not appear to be necessary for maintaining structure or function of VarS.

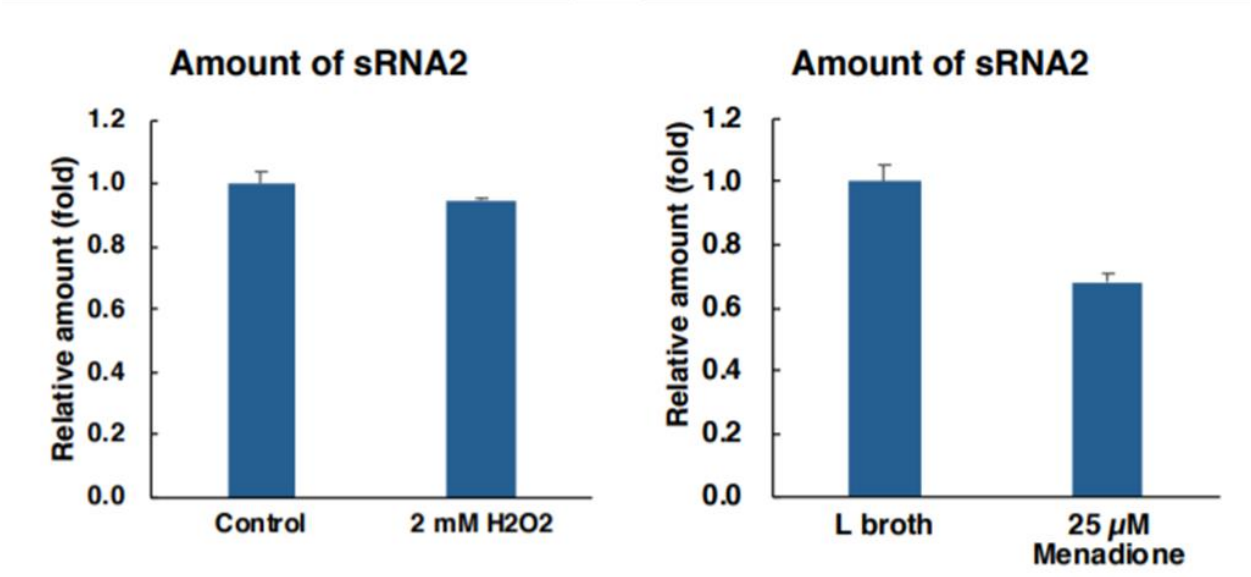


Figure 14: qRT-PCR experiment was performed to monitor the transcription of sRNA2 necessary for collagenase expression during normal and oxidative stress conditions. For oxidative stress conditions either 2 mM H_2O_2 or 25 μM Menadione were added to *V. alginolyticus* cultures during incubation. After RNA isolation it was shown that transcription was only affected by 25 μM Menadione (transcription downregulated by approximately 30%). More experiments need to be performed to determine the true effect of oxidative stress on collagenase expression.

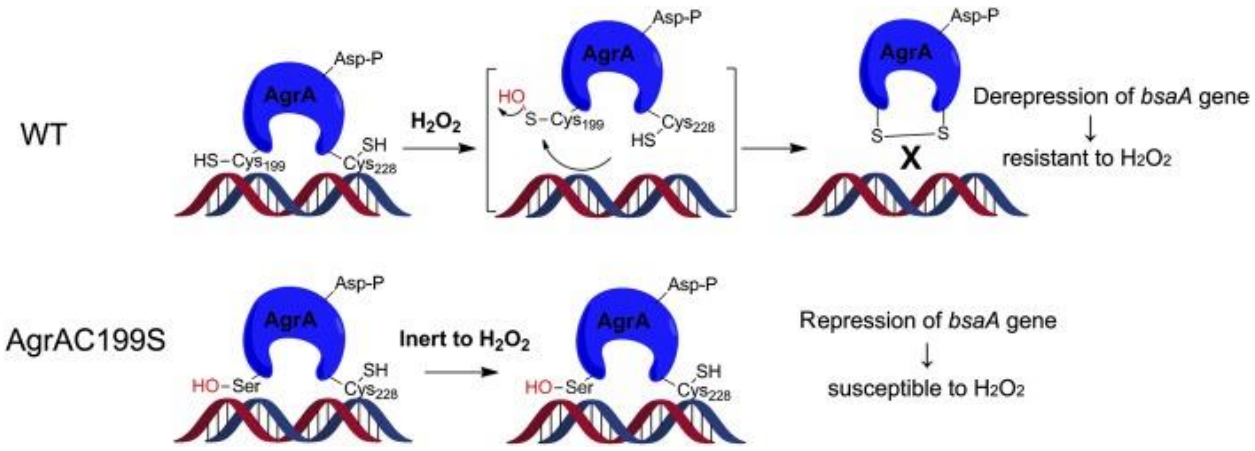


Figure 15: Cartoon representation of results for Sun et al. 2012. Disulfide bond formation between Cys199 and Cys228 in AgrA prevents the response regulator from binding to DNA which prevents downregulations glutathione peroxidase. Mutation of Cys199 to Ser prevents disulfide bond formation and makes *S. aureus* more susceptible to oxidative stress.

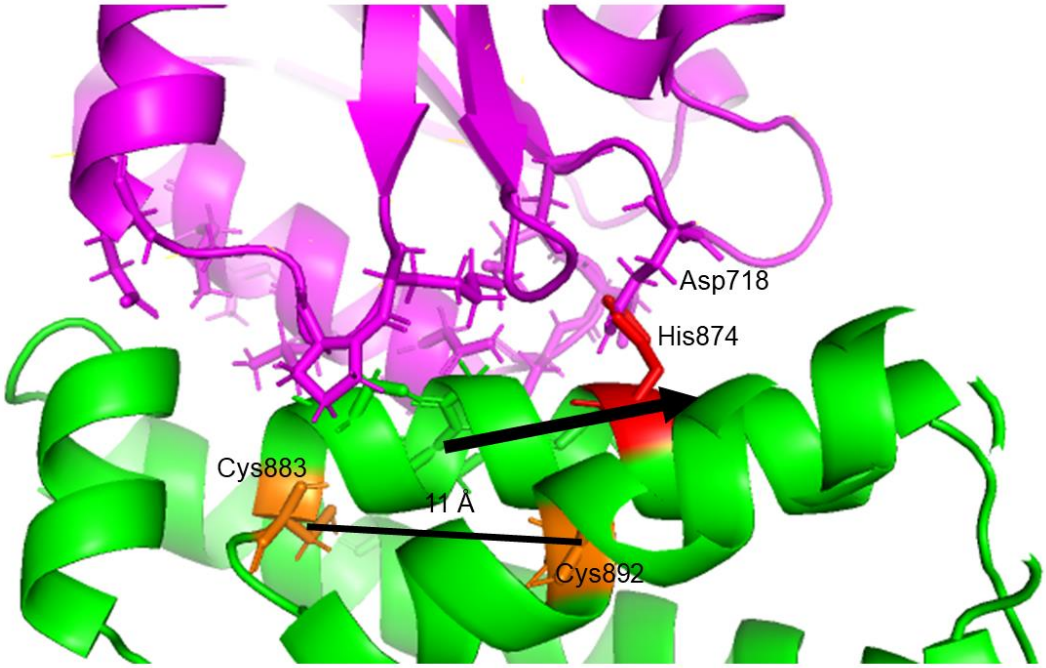


Figure 16: Model interface between HPT and REC in VarS. Model was constructed by superimposing HPT domain (R.M.S.D. = 4.041 Å) and REC domain (R.M.S.D. = 0.671 Å) with YPD1:SLN1 crystal structure (PDB: 2R25) [**1**]. Disulfide bond formation in HPT may be caused by piston-movements. Piston like movement may disrupt the binding interface between HPT and REC lowering the rate of phosphate transfer.

Works Cited

1. X. Zhao, D. M. Copeland, A. S. Soares and A. H. West, *Crystal structure of a complex between the phosphorelay protein ypd1 and the response regulator domain of sln1 bound to a phosphoryl analog*, J Mol Biol **375** (2008), no. 4, 1141-1151.

2. J. S. Fassler and A. H. West, *Histidine phosphotransfer proteins in fungal two-component signal transduction pathways*, Eukaryotic cell **12** (2013), no. 8, 1052-1060.

3. K. Lapouge, M. Schubert, F. H. Allain and D. Haas, *Gac/rsm signal transduction pathway of gamma-proteobacteria: From rna recognition to regulation of social behaviour*, Mol Microbiol **67** (2008), no. 2, 241-253.

4. A. R. Mey, H. A. Butz and S. M. Payne, *Vibrio cholerae csra regulates toxr levels in response to amino acids and is essential for virulence*, Mbio **6** (2015), no. 4, 11.

5. S. Dubrac and T. Msadek, *Identification of genes controlled by the essential yycg/yycf two-component system of staphylococcus aureus*, Journal of bacteriology **186** (2004), no. 4, 1175-1181.

6. R. Gao and A. M. Stock, *Biological insights from structures of two-component proteins*, Annual review of microbiology **63** (2009), 133-154.

7. K. I. Varughese, Madhusudan, X. Z. Zhou, J. M. Whiteley and J. A. Hoch, *Formation of a novel four-helix bundle and molecular recognition sites by dimerization of a response regulator phosphotransferase*, Molecular Cell **2** (1998), no. 4, 485-493.

8. T. Maeda, S. M. Wurgler-Murphy and H. Saito, *A two-component system that regulates an osmosensing map kinase cascade in yeast*, Nature **369** (1994), no. 6477, 242-245.

9. F. Janiak-Spens and A. H. West, *Functional roles of conserved amino acid residues surrounding the phosphorylatable histidine of the yeast phosphorelay protein ypd1*, Mol Microbiol **37** (2000), no. 1, 136-144.

10. E. N. Kennedy, S. D. Hebdon, S. K. Menon, C. A. Foster, D. M. Copeland, Q. Xu, F. Janiak-Spens and A. H. West, *Role of the highly conserved g68 residue in the yeast phosphorelay protein ypd1: Implications for interactions between histidine phosphotransfer (hpt) and response regulator proteins*, BMC Biochemistry **20** (2019), no. 1, 1.

11. J. A. Imlay, *Where in the world do bacteria experience oxidative stress?*, Environmental microbiology **21** (2019), no. 2, 521-530.

12. M. G. Espey, *Role of oxygen gradients in shaping redox relationships between the human intestine and its microbiota*, Free Radical Biology and Medicine **55** (2013), 130-140.

13. G. Wang, Y. Hong, M. K. Johnson and R. J. Maier, *Lipid peroxidation as a source of oxidative damage in helicobacter pylori: Protective roles of peroxiredoxins*, Biochimica et Biophysica Acta (BBA) - General Subjects **1760** (2006), no. 11, 1596-1603.

14. B. Van Houten, G. A. Santa-Gonzalez and M. Camargo, *Dna repair after oxidative stress: Current challenges*, Current opinion in toxicology **7** (2018), 9-16.

15. Y. Hong, J. Zeng, X. Wang, K. Drlica and X. Zhao, *Post-stress bacterial cell death mediated by reactive oxygen species*, Proc Natl Acad Sci U S A **116** (2019), no. 20, 10064-10071.

16. F. Sun, H. Liang, X. Kong, S. Xie, H. Cho, X. Deng, Q. Ji, H. Zhang, S. Alvarez, L. M. Hicks, T. Bae, C. Luo, H. Jiang and C. He, *Quorum-sensing agr mediates bacterial oxidation response via an intramolecular disulfide redox switch in the response regulator agra*, Proceedings of the National Academy of Sciences of the United States of America **109** (2012), no. 23, 9095-9100.

17. Z. Liu, M. Yang, G. L. Peterfreund, A. M. Tsou, N. Selamoglu, F. Daldal, Z. Zhong, B. Kan and J. Zhu, *Vibrio cholerae anaerobic induction of virulence gene expression is controlled by thiol-based switches of virulence regulator aphb*, Proceedings of the National Academy of Sciences of the United States of America **108** (2011), no. 2, 810-815.

18. L. R. Swem, B. J. Kraft, D. L. Swem, A. T. Setterdahl, S. Masuda, D. B. Knaff, J. M. Zaleski and C. E. Bauer, *Signal transduction by the global regulator regb is mediated by a redox-active cysteine*, The EMBO journal **22** (2003), no. 18, 4699-4708.

19. T. Mima, K. Gotoh, Y. Yamamoto, K. Maeda, T. Shirakawa, S. Matsui, Y. Murata, T. Koide, H. Tokumitsu and O. Matsushita, *Expression of collagenase is regulated by the vars/vara two-component regulatory system in vibrio alginolyticus*, J Membr Biol **251** (2018), no. 1, 51-63.

20. A. Roy, A. Kucukural and Y. Zhang, *I-tasser: A unified platform for automated protein structure and function prediction*, Nature protocols **5** (2010), no. 4, 725-738.

21. F. Madeira, Y. M. Park, J. Lee, N. Buso, T. Gur, N. Madhusoodanan, P. Basutkar, A. R. N. Tivey, S. C. Potter, R. D. Finn and R. Lopez, *The embl-ebi search and sequence analysis tools apis in 2019*, Nucleic Acids Research **47** (2019), no. W1, W636-W641.

22. E. Wuensch and H. G. Heidrich, *[on the quantitative determination of collagenase]*, Hoppe Seylers Z Physiol Chem **333** (1963), 149-151.

23. O. Ashenberg, K. Rozen-Gagnon, M. T. Laub and A. E. Keating, *Determinants of homodimerization specificity in histidine kinases*, Journal of molecular biology **413** (2011), no. 1, 222-235.

24. K. Linke and U. Jakob, *Not every disulfide lasts forever: Disulfide bond formation as a redox switch*, Antioxid Redox Signal **5** (2003), no. 4, 425-434.

25. R. Bhattacharyya, D. Pal and P. Chakrabarti, *Disulfide bonds, their stereospecific environment and conservation in protein structures*, Protein Engineering, Design and Selection **17** (2004), no. 11, 795-808.

26. T. Affandi and M. M. McEvoy, *Mechanism of metal ion-induced activation of a two-component sensor kinase*, Biochem J **476** (2019), no. 1, 115-135.

27. D. Xing, M. B. Ryndak, L. Wang, I. Kolesnikova, I. Smith and S. Wang, *Asymmetric structure of the dimerization domain of phor, a sensor kinase important for the virulence of mycobacterium tuberculosis*, ACS omega **2** (2017), no. 7, 3509-3517.

28. J. Yang, Y. Wang and Y. Zhang, *Resq: An approach to unified estimation of b-factor and residue-specific error in protein structure prediction*, J Mol Biol **428** (2016), no. 4, 693-701.

29. Z. S. Derewenda and P. G. Vekilov, *Entropy and surface engineering in protein crystallization*, Acta Crystallogr D Biol Crystallogr **62** (2006), no. Pt 1, 116-124.

30. N. J. Greenfield, *Using circular dichroism spectra to estimate protein secondary structure*, Nature Protocols **1** (2006), no. 6, 2876-2890.

31. C. P. Pan, P. L. Muiño, M. D. Barkley and P. R. Callis, *Correlation of tryptophan fluorescence spectral shifts and lifetimes arising directly from heterogeneous environment*, J Phys Chem B **115** (2011), no. 12, 3245-3253.

32. L. Holm, *Benchmarking fold detection by dalilite v.5*, Bioinformatics **35** (2019), no. 24, 5326-5327.

33. F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson and D. G. Higgins, *Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega*, Mol Syst Biol **7** (2011), 539.

34. V. Keil-Dlouha, R. Misrahi and B. Keil, *The induction of collagenase and a neutral proteinase by their high molecular weight substrates in achromobacter iophagus*, Journal of Molecular Biology **107** (1976), no. 3, 293-305.

35. R. C. Cumming, N. L. Andon, P. A. Haynes, M. Park, W. H. Fischer and D. Schubert, *Protein disulfide bond formation in the cytoplasm during oxidative stress*, J Biol Chem **279** (2004), no. 21, 21749-21758.

36. A. Rietsch and J. Beckwith, *The genetics of disulfide bond metabolism*, Annual Review of Genetics **32** (1998), no. 1, 163-184.

37. C. Staerck, A. Gastebois, P. Vandeputte, A. Calenda, G. Larcher, L. Gillmann, N. Papon, J.-P. Bouchara and M. J. J. Fleury, *Microbial antioxidant defense enzymes*, Microbial Pathogenesis **110** (2017), 56-65.

38. J.-M. Lv, S.-Q. Lü, Z.-P. Liu, J. Zhang, B.-X. Gao, Z.-Y. Yao, Y.-X. Wu, L. A. Potempa, S.-R. Ji, M. Long and Y. Wu, *Conformational folding and disulfide bonding drive distinct stages of protein structure formation*, Scientific reports **8** (2018), no. 1, 1494-1494.

39. C. W. Gruber, M. Cemazar, B. Heras, J. L. Martin and D. J. Craik, *Protein disulfide isomerase: The structure of oxidative folding*, Trends Biochem Sci **31** (2006), no. 8, 455-464.

40. H. Ali Khan and B. Mutus, *Protein disulfide isomerase a multifunctional protein with multiple physiological roles*, Frontiers in chemistry **2** (2014), 70-70.

41. H. Kadokura and J. Beckwith, *Mechanisms of oxidative protein folding in the bacterial cell envelope*, Antioxidants & redox signaling **13** (2010), no. 8, 1231-1246.

42. S. Y. Queck, M. Jameson-Lee, A. E. Villaruz, T.-H. L. Bach, B. A. Khan, D. E. Sturdevant, S. M. Ricklefs, M. Li and M. Otto, *Rnaiii-independent target gene control by the agr quorum-sensing system: Insight into the evolution of virulence regulation in staphylococcus aureus*, Molecular cell **32** (2008), no. 1, 150-158.

43. M. Abo and E. Weerapana, *Chemical probes for redox signaling and oxidative stress*, Antioxidants & redox signaling **30** (2019), no. 10, 1369-1386.

44. E. Geisinger, T. W. Muir and R. P. Novick, *Agr receptor mutants reveal distinct modes of inhibition by staphylococcal autoinducing peptides*, Proc Natl Acad Sci U S A **106** (2009), no. 4, 1216-1221.

45. P. Mayville, G. Ji, R. Beavis, H. Yang, M. Goger, R. P. Novick and T. W. Muir, *Structure-activity analysis of synthetic autoinducing thiolactone peptides from staphylococcus aureus responsible for virulence*, Proc Natl Acad Sci U S A **96** (1999), no. 4, 1218-1223.

46. G. Y. C. Cheung, H.-S. Joo, S. S. Chatterjee and M. Otto, *Phenol-soluble modulins – critical determinants of staphylococcal virulence*, FEMS Microbiology Reviews **38** (2014), no. 4, 698-719.

47. P. M. Dunman, E. Murphy, S. Haney, D. Palacios, G. Tucker-Kellogg, S. Wu, E. L. Brown, R. J. Zagursky, D. Shlaes and S. J. Projan, *Transcription profiling-based identification of staphylococcus aureus genes regulated by the agr and/or sara loci*, Journal of bacteriology **183** (2001), no. 24, 7341-7353.

48. J. L. Taylor, R. S. De Silva, G. Kovacikova, W. Lin, R. K. Taylor, K. Skorupski and F. J. Kull, *The crystal structure of aphb, a virulence gene activator from vibrio cholerae, reveals residues that influence its response to oxygen and ph*, Molecular microbiology **83** (2012), no. 3, 457-470.

49. Z. Liu, H. Wang, Z. Zhou, Y. Sheng, N. Naseer, B. Kan and J. Zhu, *Thiol-based switch mechanism of virulence regulator aphb modulates oxidative stress response in vibrio cholerae*, Molecular microbiology **102** (2016), no. 5, 939-949.

50. C. Pardo-Esté, A. A. Hidalgo, C. Aguirre, A. C. Briones, C. E. Cabezas, J. Castro-Severyn, J. A. Fuentes, C. M. Opazo, C. A. Riedel, C. Otero, R. Pacheco, M. A. Valvano and C. P. Saavedra, *The arcab two-component regulatory system promotes resistance to reactive oxygen species and systemic infection by salmonella typhimurium*, PLoS One **13** (2018), no. 9, e0203497.

51. J. W. Hall, J. Yang, H. Guo and Y. Ji, *The staphylococcus aureus airsr two-component system mediates reactive oxygen species resistance via transcriptional regulation of staphyloxanthin production*, Infection and immunity **85** (2017), no. 2, e00838-00816.

52. M. E. McBee, Y. H. Chionh, M. L. Sharaf, P. Ho, M. W. L. Cai and P. C. Dedon, *Production of superoxide in bacteria is stress- and cell state-dependent: A gating-optimized flow cytometry method that minimizes ros measurement artifacts with fluorescent dyes*, Frontiers in microbiology **8** (2017), 459-459.

53. H. S. Marinho, C. Real, L. Cyrne, H. Soares and F. Antunes, *Hydrogen peroxide sensing, signaling and regulation of transcription factors*, Redox biology **2** (2014), 535-562.

Chapter 7: Conclusion

Based on their ability to target and degrade collagen both Col G and Col H are potential medical tools of interest. Biomolecules fused to non-catalytic segments of Col G (PKD, CBD1 and CBD2) and Col H (PKD1, PKD2 and CBD) are non-immunogenic and their ability to heal bone fractures and promote hair regrowth in mice is well known [**1**] [**2**]. Col G CBD1CBD2s ability to wedge between parallel oriented tropocollagens makes it ideal for the localized delivery of bFGF while Col H's lower binding affinity makes it better for promoting systemic effects of PTH [**1**] [**3**] [**4**]. Understanding the role of PKD domains in Col H binding may help in linker design for fusion proteins [**5**]. Full-length Col G and Col H have been used to isolate pancreatic islets and stem cells for purposes of auto cell transplant [**6**] [**7**]. In order to further improve Col G for cell isolation, future experiments should focus on strengthening the linker region between CBD1 and CBD2. Site directed mutagenesis of Lys residues to Asn or Asp in linker region may make the region resistant to trypsin and clostripain. Full-length SAXS derived structures for Col G and Col H may help to identify regions in the enzymes that can be chemically modified to reduce immune response in patients. Further understanding on how Col G and Col H enzymes work together to degrade collagen would help to further their development as medical tools.

In order to better understand collagenase expression the VarS/VarA two-component system from *V. alginolyticus* was studied. This systems or closely related systems are responsible for the expression of virulence factors for all *Vibrio* species [**8**] [**9**] [**10**] [**11**]. In two-component systems a signal is received which causes the protein to dimerize, following this a phosphate is donated from ATP to a conserved His residue in the dHp domain, next the phosphate is passed (either in a *cis* or *trans* method and through a normal or phosphorelay pathway) to a conserved

Asp residue on a response regulator, responsible for transcription and translation of specific genes and proteins. In a phosphorelay pathway phosphate passed to His residue in the dHp domain is passed through internal REC and HPT domains within the sensor histidine kinase. Once phosphate is passed to HPT it is then passed to response regulator like in the normal histidine kinase. Response regulator then modulates gene transcription. Solving the crystal structure of REC domain from VarS would give a new target for the design of anti-biotics. While investigating the potential redox sensing Cys residues in HPT may help to understand how some bacteria are able to survive contact with host immune systems.

Works Cited

1. H. Sekiguchi, K. Uchida, O. Matsushita, G. Inoue, N. Nishi, R. Masuda, N. Hamamoto, T. Koide, S. Shoji and M. Takaso, *Basic fibroblast growth factor fused with tandem collagen-binding domains from clostridium histolyticum collagenase colg increases bone formation*, Biomed Res Int **2018** (2018), 8393194.

2. R. Katikaneni, T. Ponnapakkam, O. Matsushita, J. Sakon and R. Gensure, *Treatment and prevention of chemotherapy-induced alopecia with pth-cbd, a collagen-targeted parathyroid hormone analog, in a non-depilated mouse model*, Anti-Cancer Drugs **25** (2014), no. 1, 30-38.

3. M. F. Holick, S. Ray, T. C. Chen, X. Tian and K. S. Persons, *A parathyroid hormone antagonist stimulates epidermal proliferation and hair growth in mice*, Proc Natl Acad Sci U S A **91** (1994), no. 17, 8014-8016.

4. E. M. Peters, K. Foitzik, R. Paus, S. Ray and M. F. Holick, *A new strategy for modulating chemotherapy-induced alopecia, using pth/pthrp receptor agonist and antagonist*, J Invest Dermatol **117** (2001), no. 2, 173-178.

5. X. Chen, *Fusion protein linkers: Property, design and functionality*, **65** (2013), no. 10, 1357-1369.

6. M. L. Green, A. G. Breite, C. A. Beechler, F. E. Dwulet and R. C. McCarthy, *Effectiveness of different molecular forms of c. Histolyticum class i collagenase to recover islets*, Islets **9** (2017), no. 6, 177-181.

7. R. C. McCarthy, A. G. Breite, M. L. Green and F. E. Dwulet, *Tissue dissociation enzymes for isolating human islets for transplantation: Factors to consider in setting enzyme acceptance criteria*, Transplantation **91** (2011), no. 2, 137-145.

8. T. Tanabe, A. Kato, K. Shiuchi, K. Miyamoto, H. Tsujibo, J. Maki, S. Yamamoto and T. Funahashi, *Regulation of the expression of the vibrio parahaemolyticus peua gene encoding an alternative ferric enterobactin receptor*, PLoS One **9** (2014), no. 8, e105749.

9. A. R. Mey, H. A. Butz and S. M. Payne, *Vibrio cholerae csra regulates toxr levels in response to amino acids and is essential for virulence*, Mbio **6** (2015), no. 4, 11.

10. A. T. Cheng, K. M. Ottemann and F. H. Yildiz, *Vibrio cholerae response regulator vxrb controls colonization and regulates the type vi secretion system*, Plos Pathogens **11** (2015), no. 5, 21.

11. M. J. Kim, J. Kim, H. Y. Lee, H. J. Noh, K.-H. Lee and S.-J. Park, *Role of acsr in expression of the acetyl-coa synthetase gene in vibrio vulnificus*, BMC Microbiology **15** (2015), no. 1, 86.

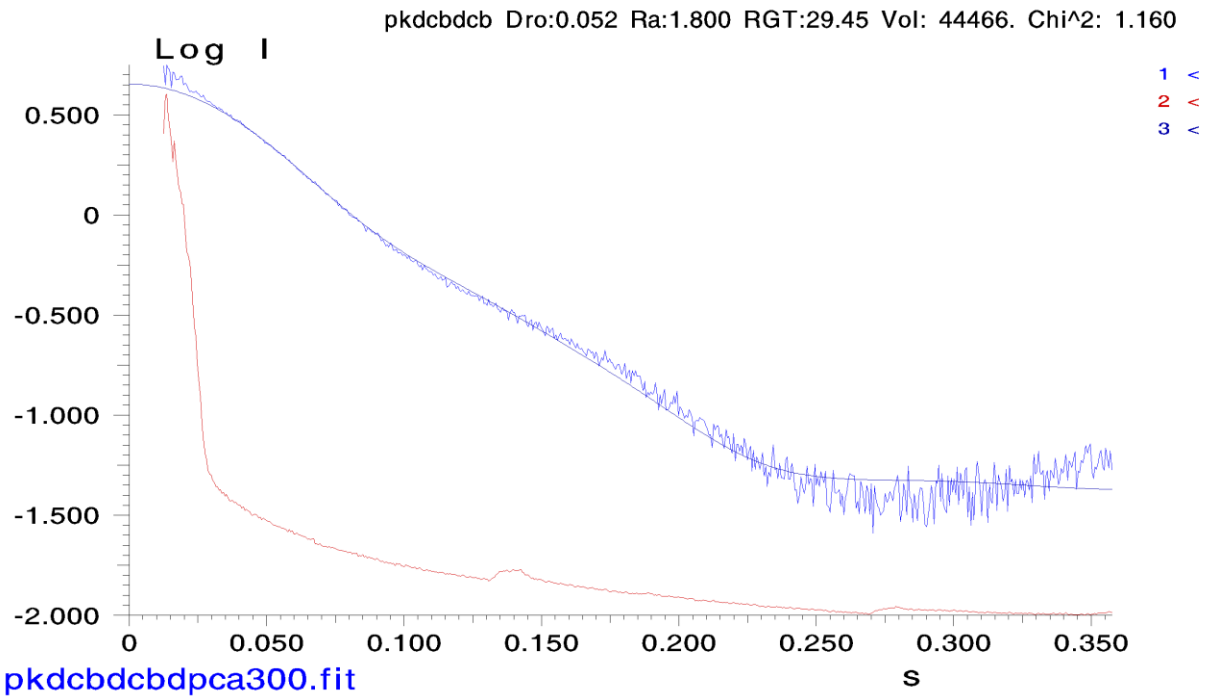


Figure 1: Crysol fit file for PKDCBD1CBD2 pCa 3. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. χ^2 value of 1.16 indicates the fit between the two is good.

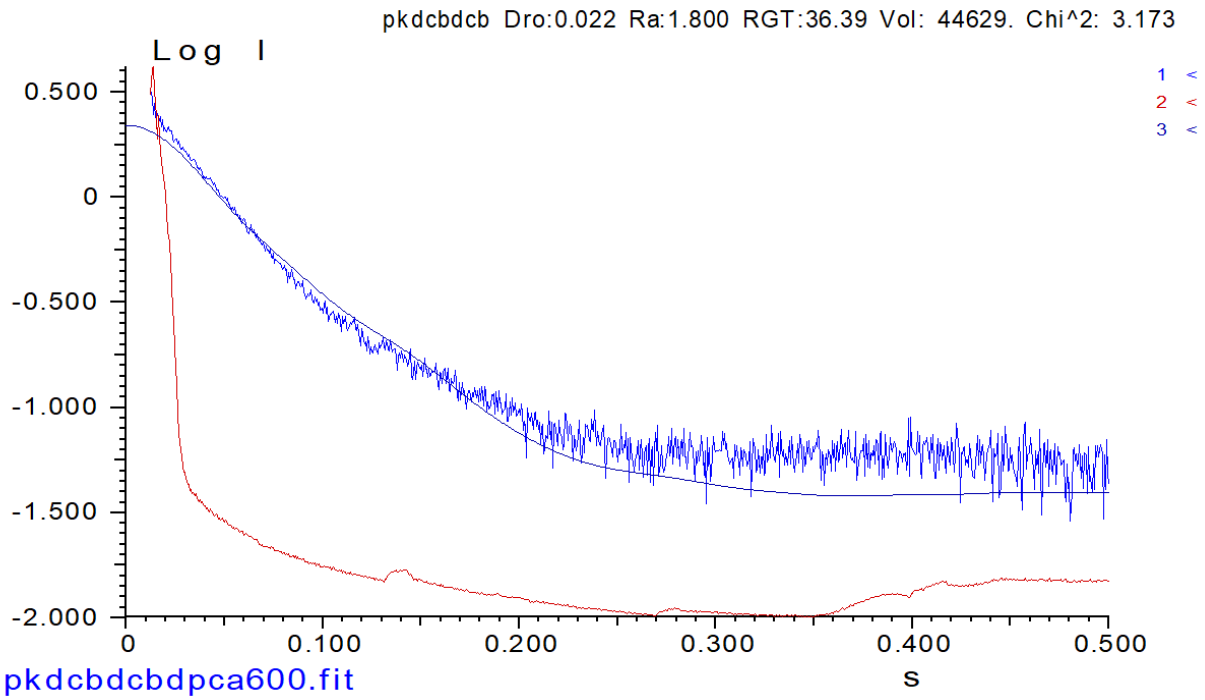


Figure 2: Crysol fit file for PKDCBD1CBD2 pCa 6. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. χ^2 value of 3.173 indicates the fit between the two is good.

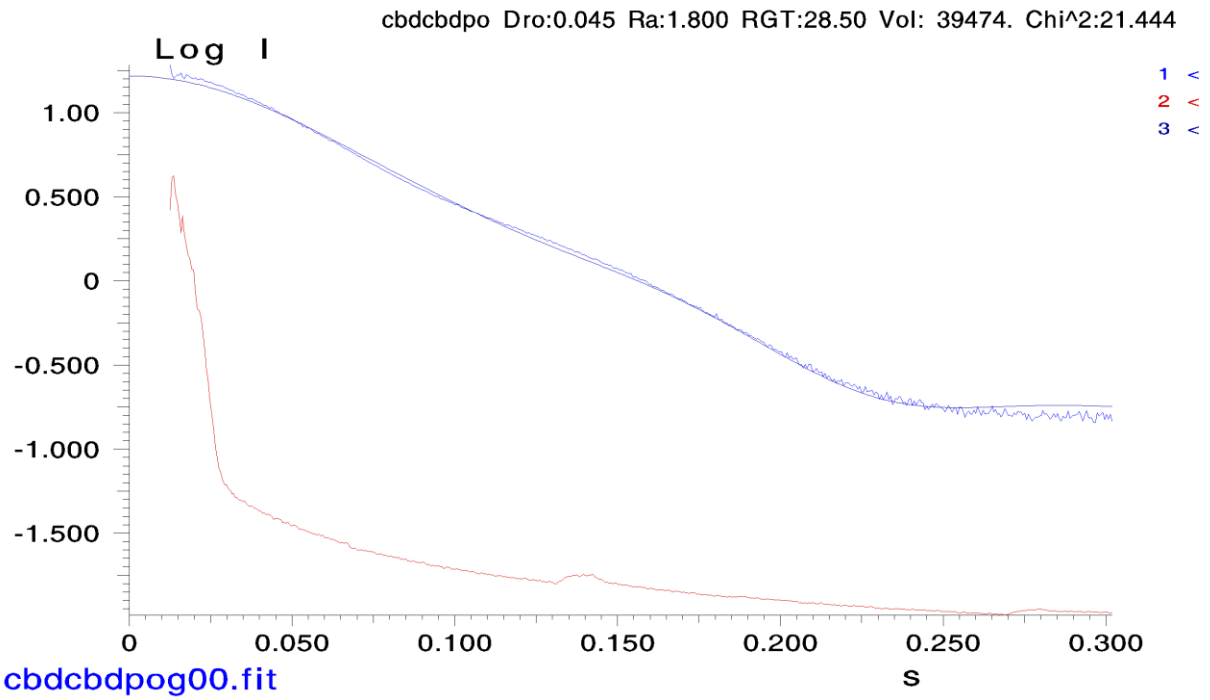


Figure 3: Crysol fit file for CBD1CBD2:$[(POG)_{10}]_3$. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. χ^2 value of 21.44 indicates the fit between the two could be better (ideally $\chi^2 = 1$). Repositioning domains in envelope may give a better fit.

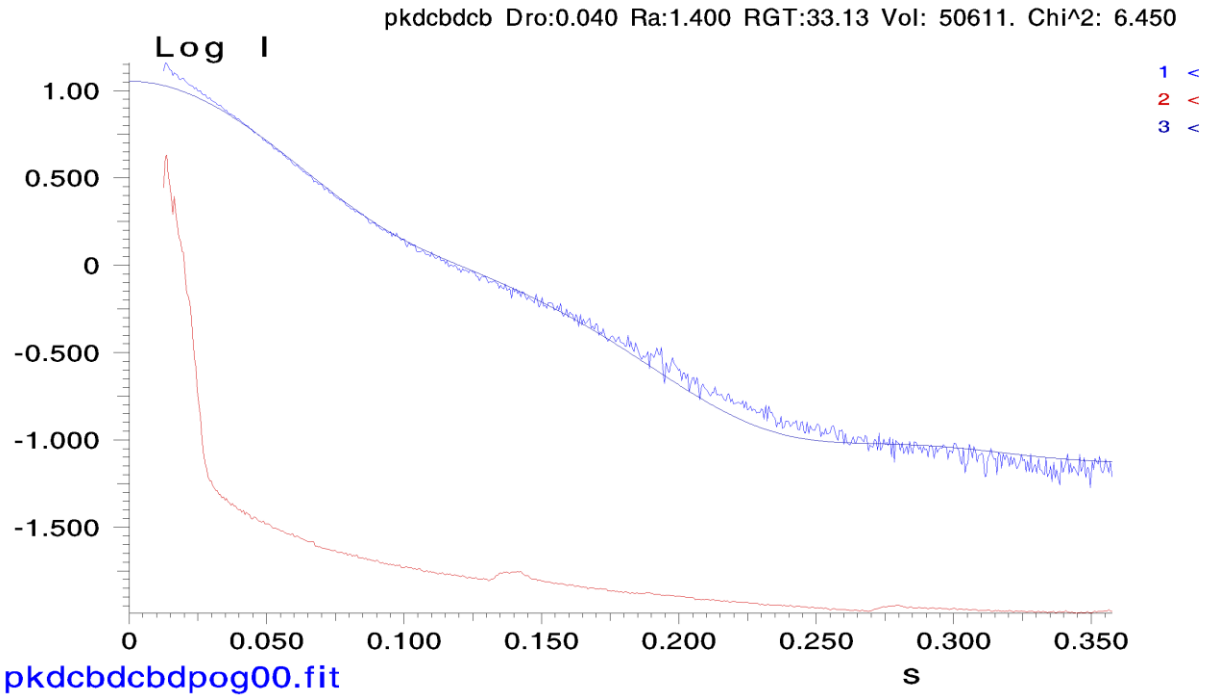


Figure 4: Crysol fit file for PKDCBD1CBD2:$[(POG)_{10}]_3$. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. χ^2 value of 6.45 indicates the fit between the two is good.

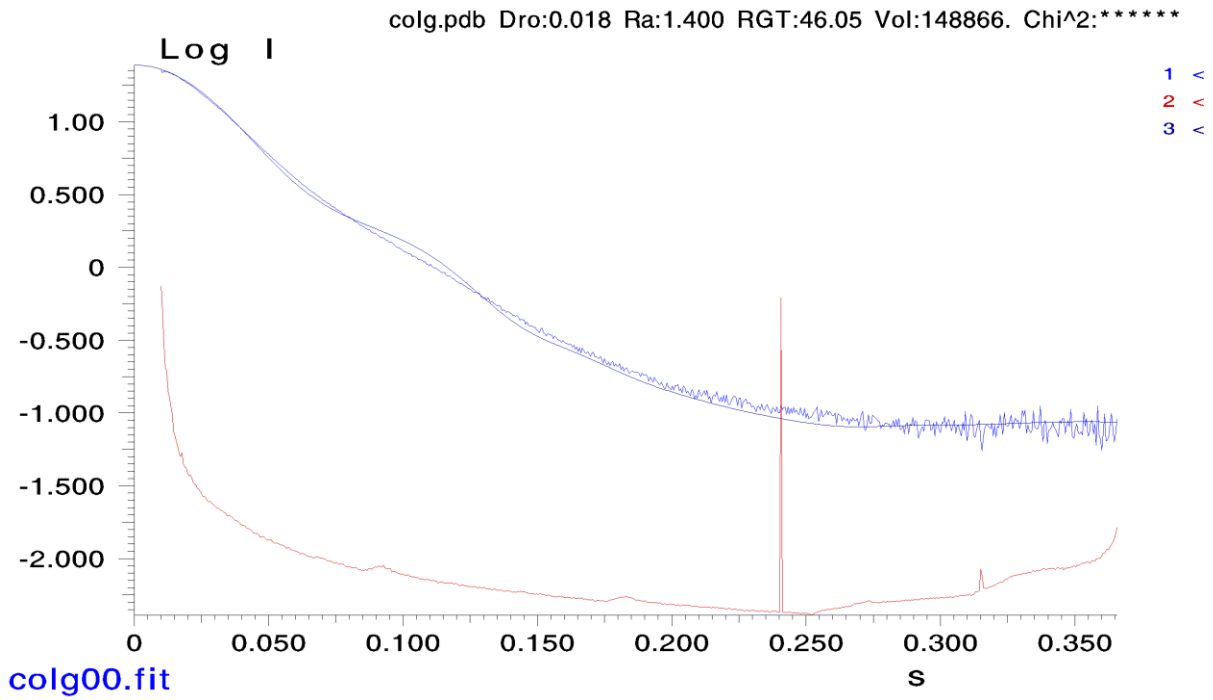


Figure 5: Crysol fit file for Col G pCa 3. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. χ^2 value of 111.15 indicates the fit between the two could be better. Averaging crystal structures for open and closed catalytic domain confirmations may give a better fit.

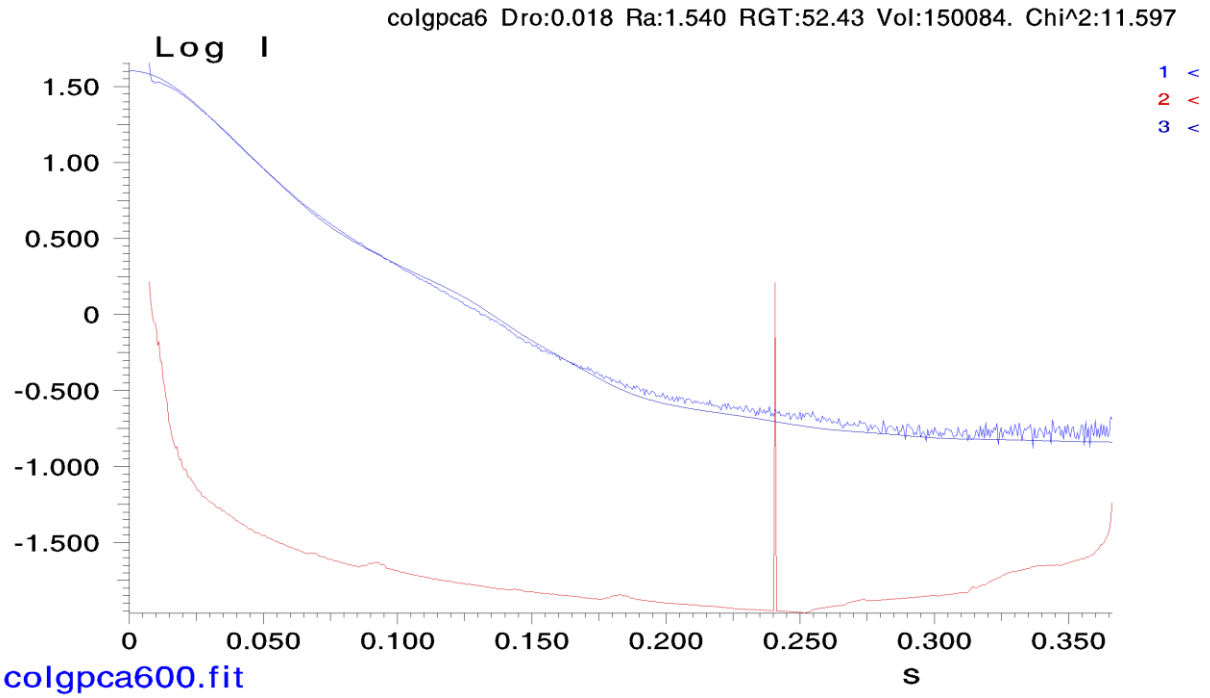


Figure 6: Crysol fit file for Col G pCa 6. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. χ^2 value of 11.597 indicates the fit between the two is good.

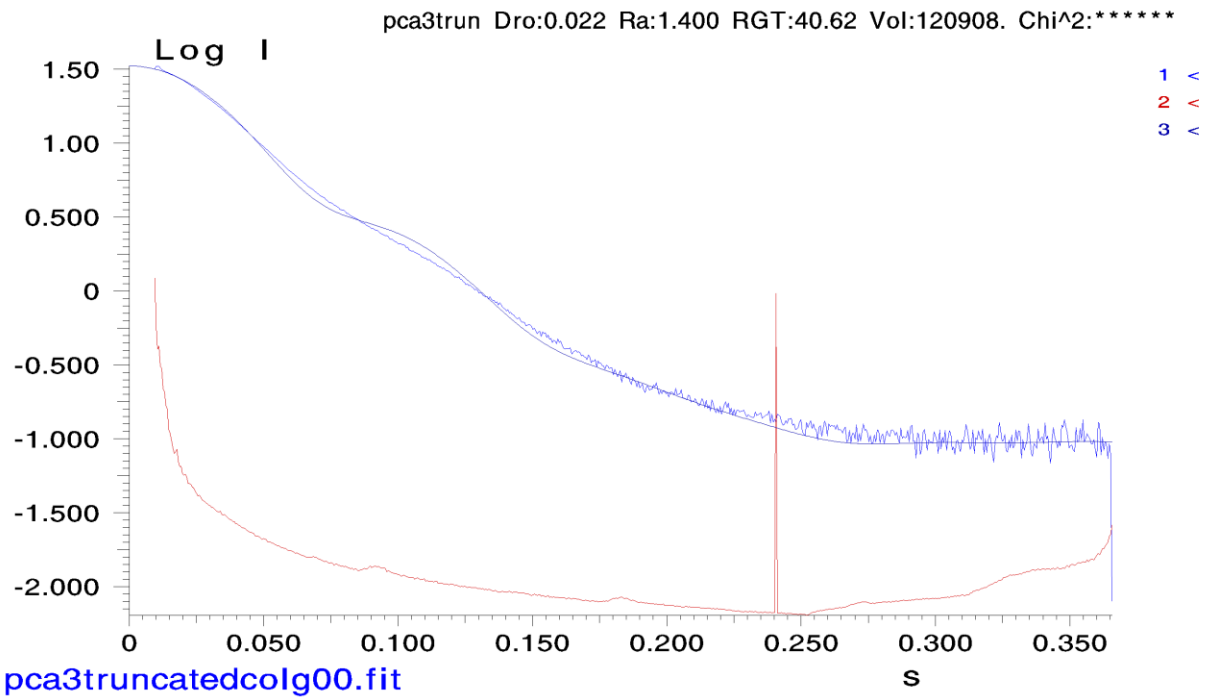


Figure 7: Crysol fit file for Col G -ΔCBD2 pCa 3. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. χ^2 value of 163.21 indicates the fit between the two could be better. Averaging crystal structures for open and closed catalytic domain confirmations may give a better fit.

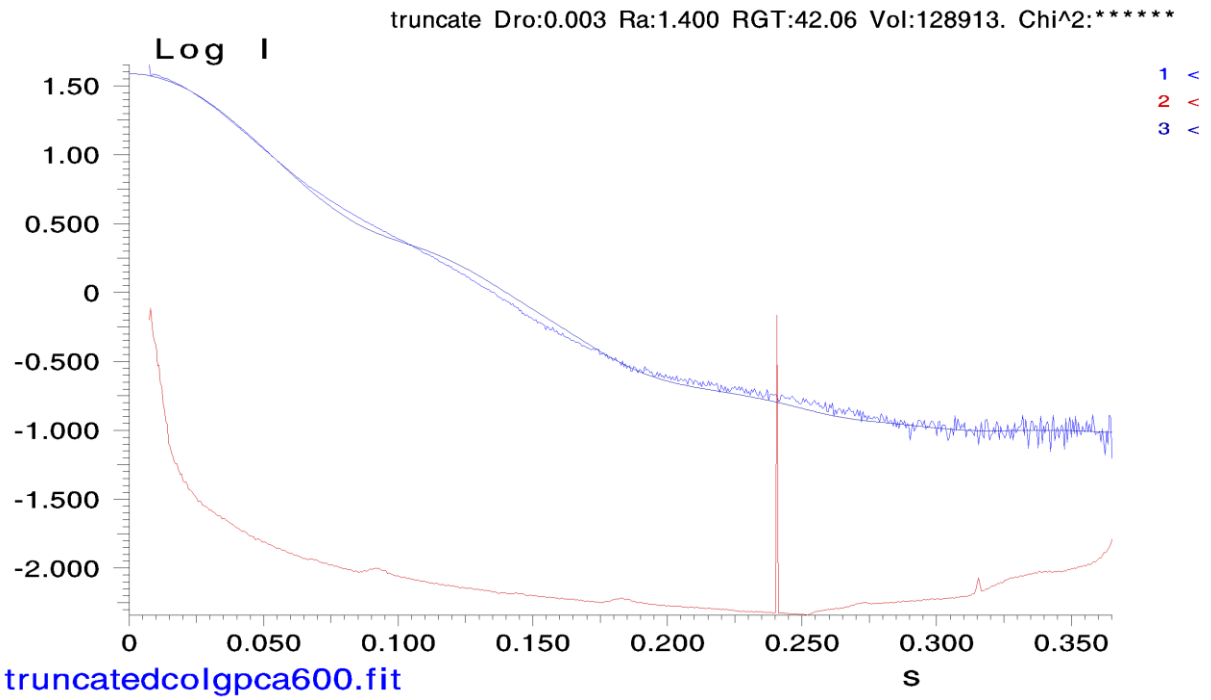


Figure 8: Crysol fit file for Col G -ΔCBD2 pCa 6. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. χ^2 value of 210.62 indicates the fit between the two could be better. Averaging crystal structures for open and closed catalytic domain may give a better fit.

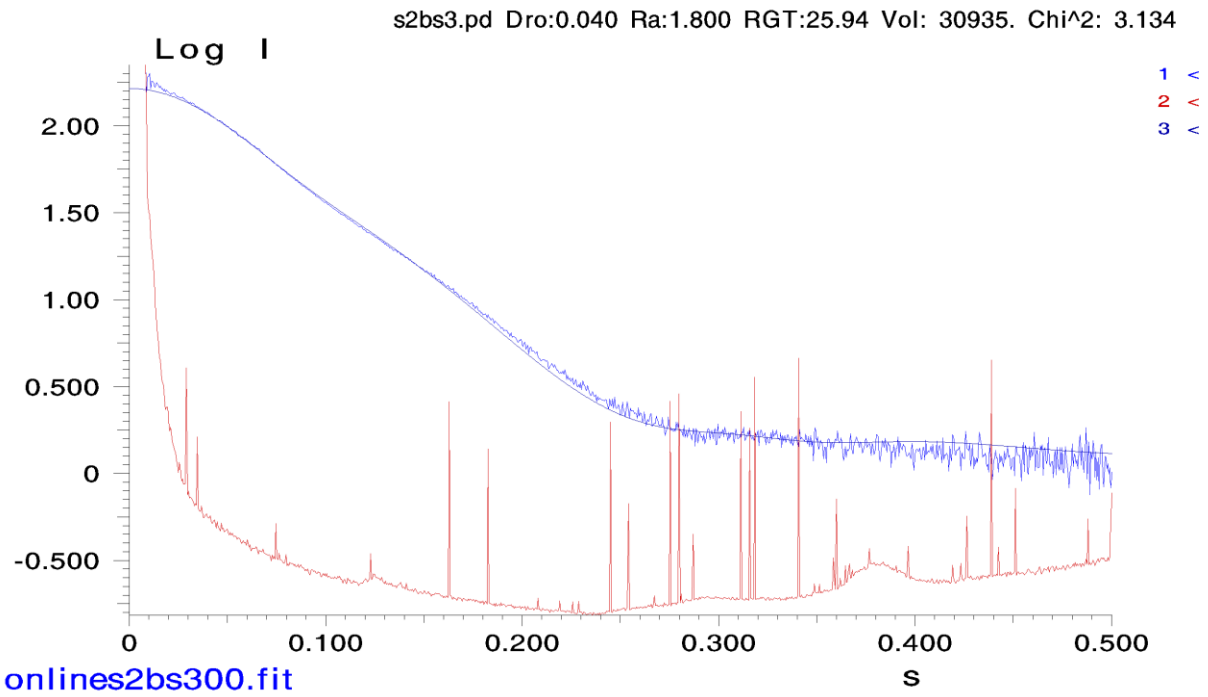


Figure 9: Crysol fit file for PKD2CBD. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. χ^2 value of 3.134 indicates the fit between the two is good.

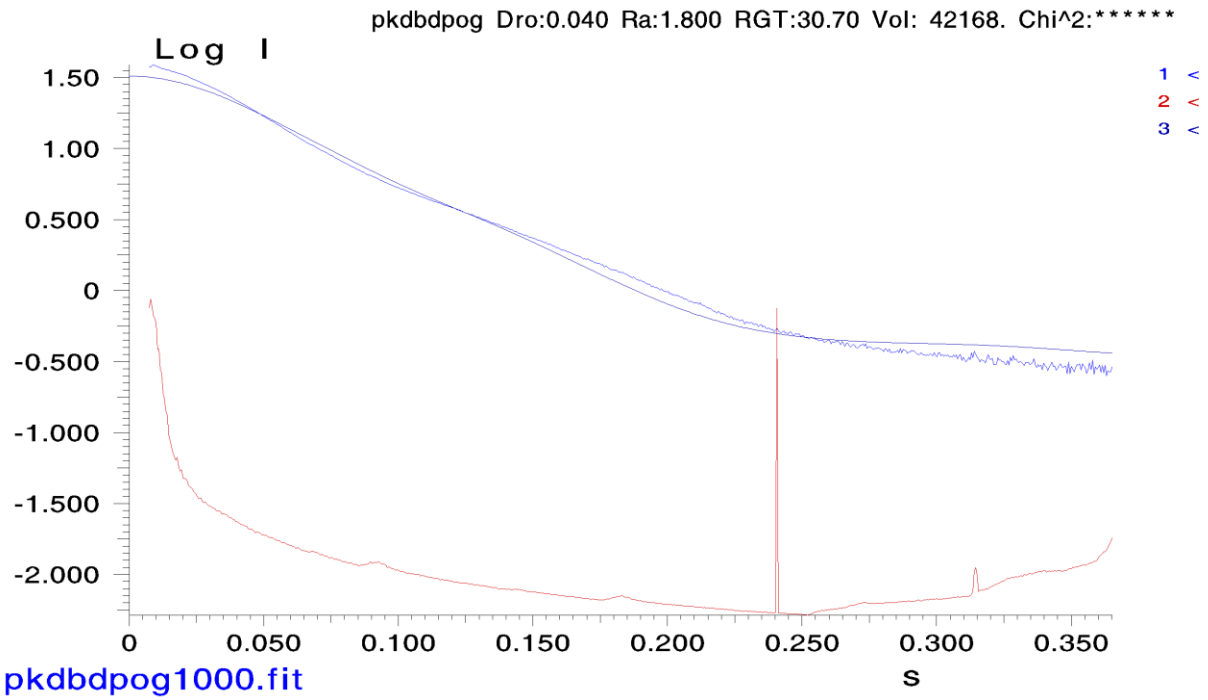


Figure 10: Crysol fit file for PKD2CBD:$[(POG)_{10}]_3$. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. Fit between experimental and theoretical scattering profile appears good however, χ^2 value of 972.67 indicates that fit could be better.

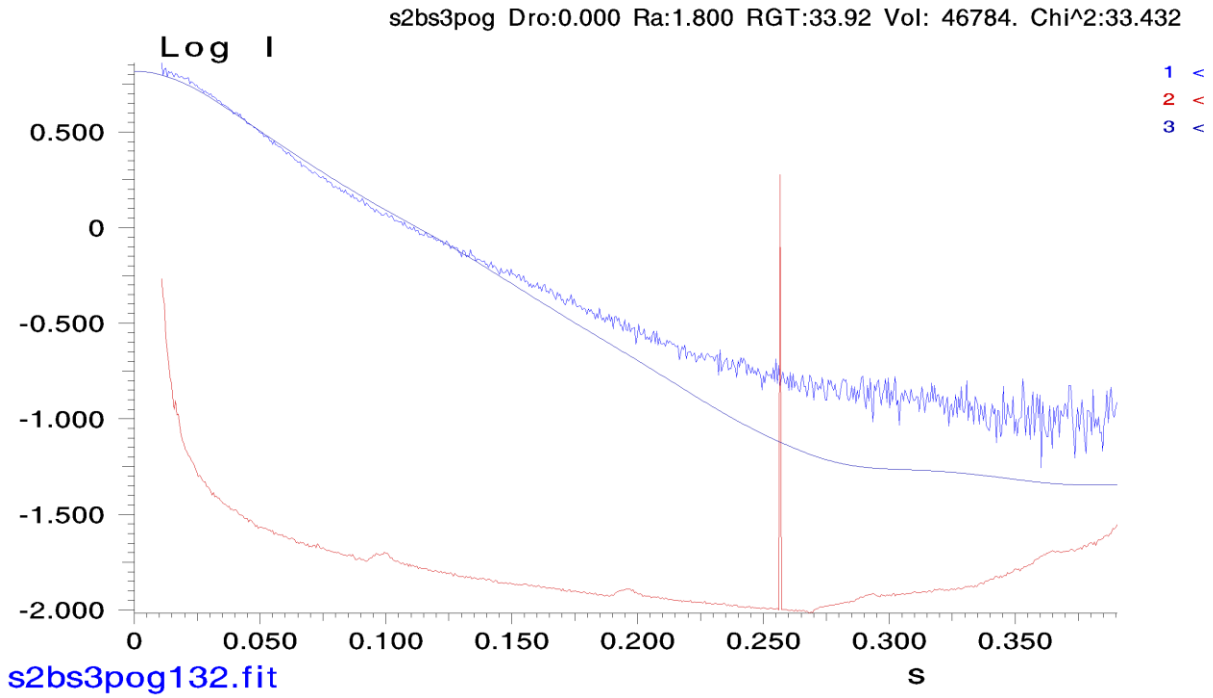


Figure 11: Crysol fit file for PKD2CBD:$[(POG)_{13}]_3$. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. χ^2 value of 33.432.67 indicates that fit could be better. Removing bumps in the high s region and recalculating may give a better fit.

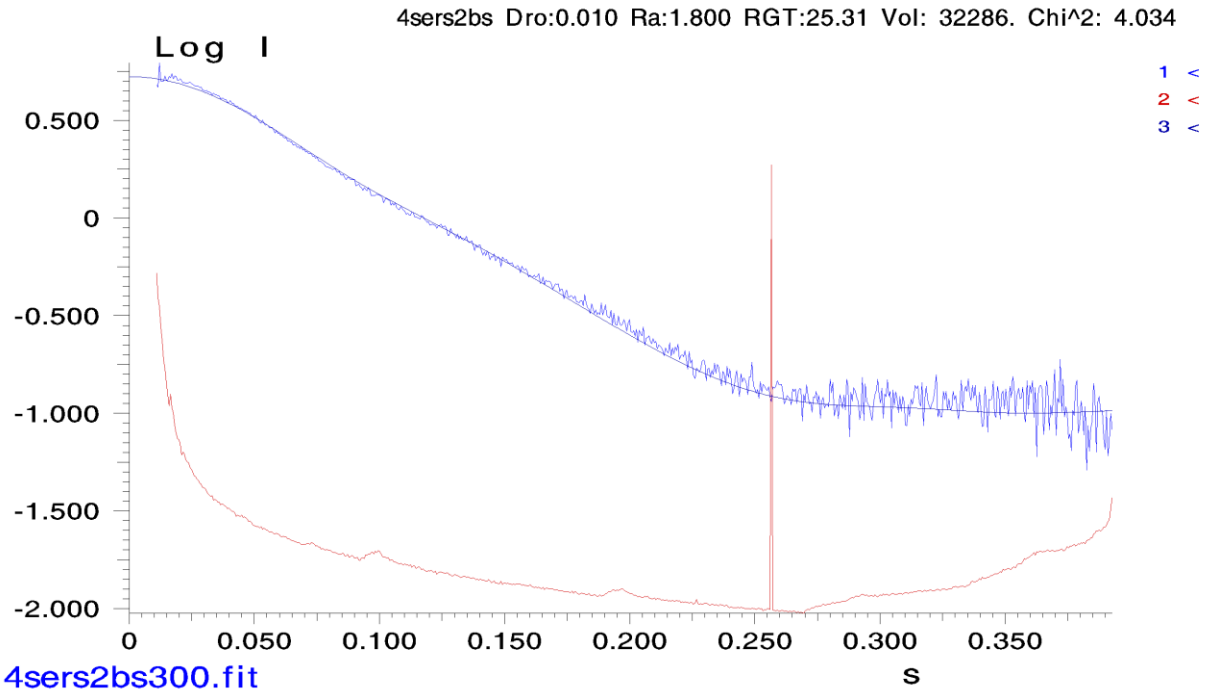


Figure 11: Crysol fit file for 4 Ser PKD2CBD. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. χ^2 value of 4.034 indicates that fit is good.

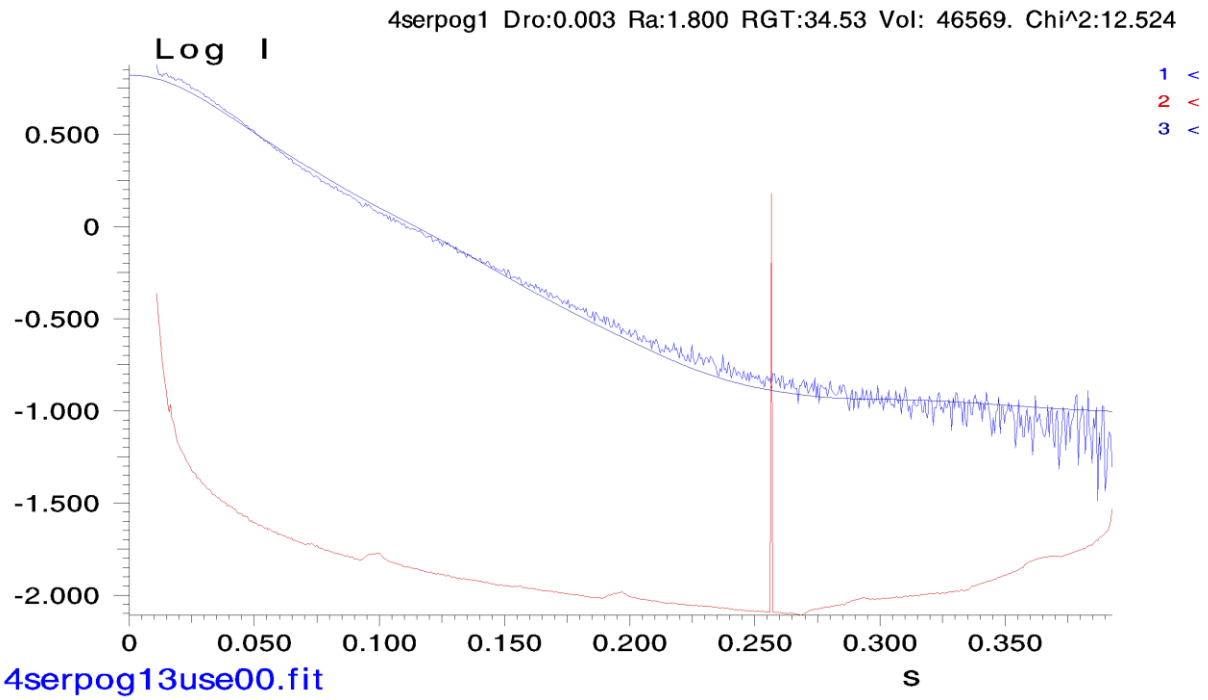


Figure 12: Crysol fit file for 4 Ser PKD2CBD:[GPRG(POG)$_{13}$]$_3$. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. χ^2 value of 12.524 indicates that fit is good.

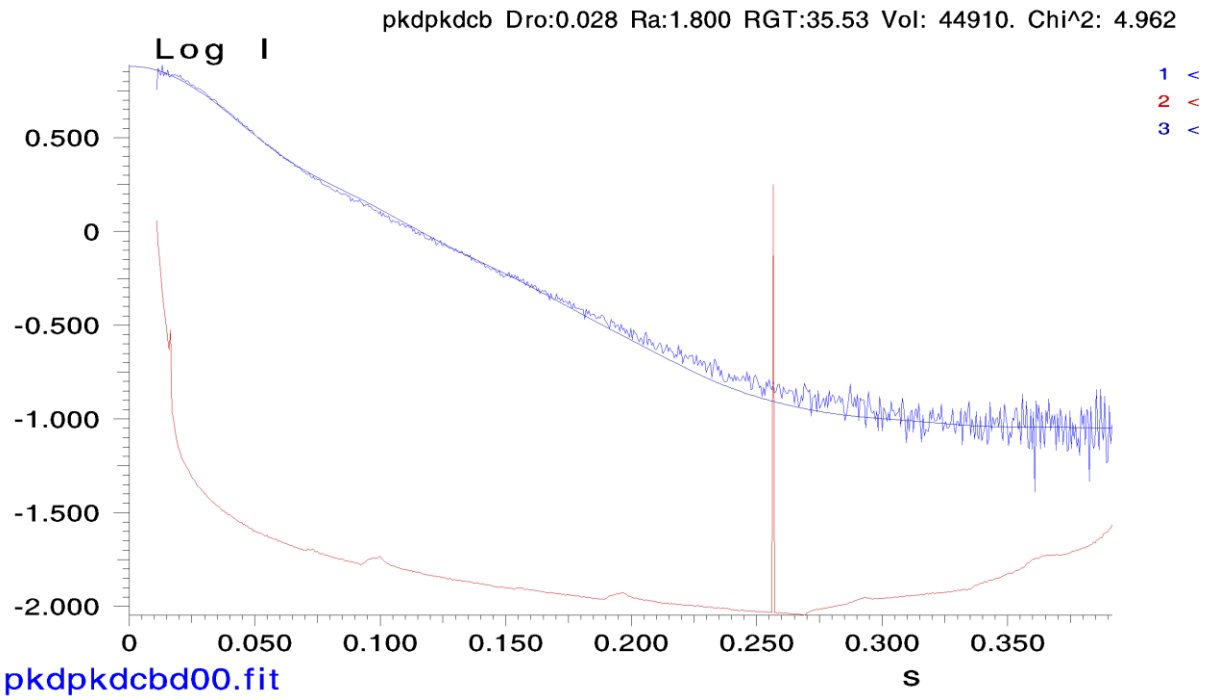


Figure 13: Crysol fit file for PKD1PKD2CBD. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. χ^2 value of 4.962 indicates that fit is good.

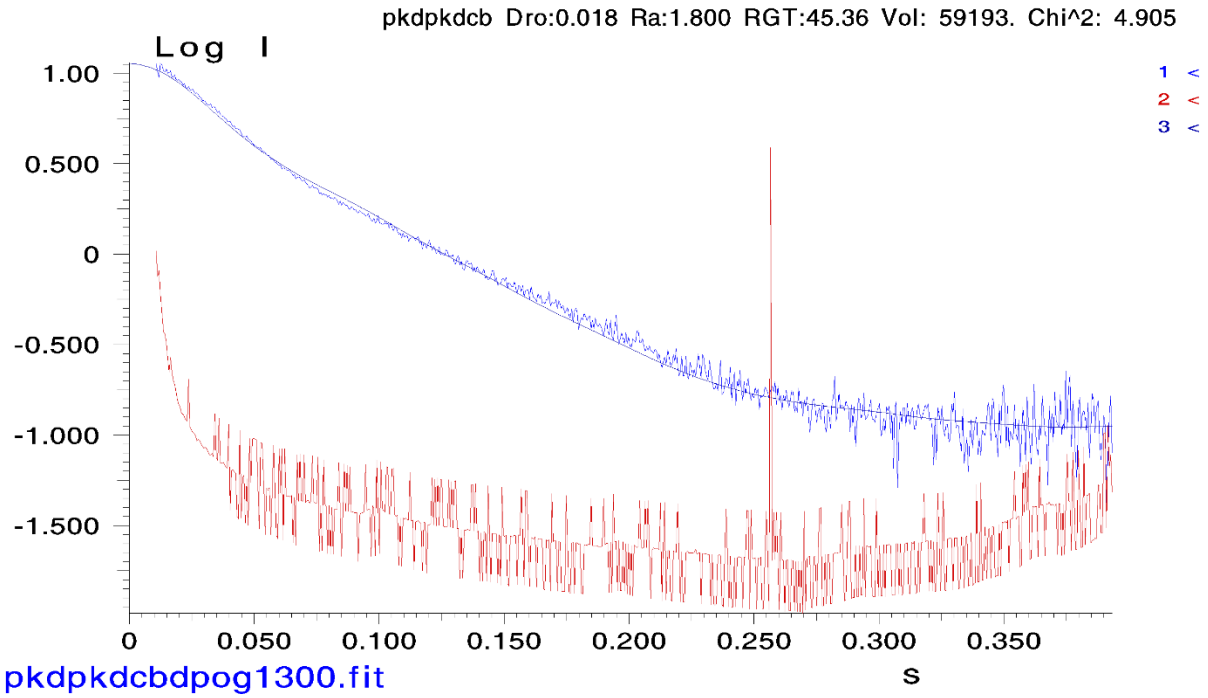


Figure 14: Crysol fit file for PKD1PKD2CBD:[GPRG$(POG)_{13}]_3$. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. χ^2 value of 4.905 indicates that fit is good.

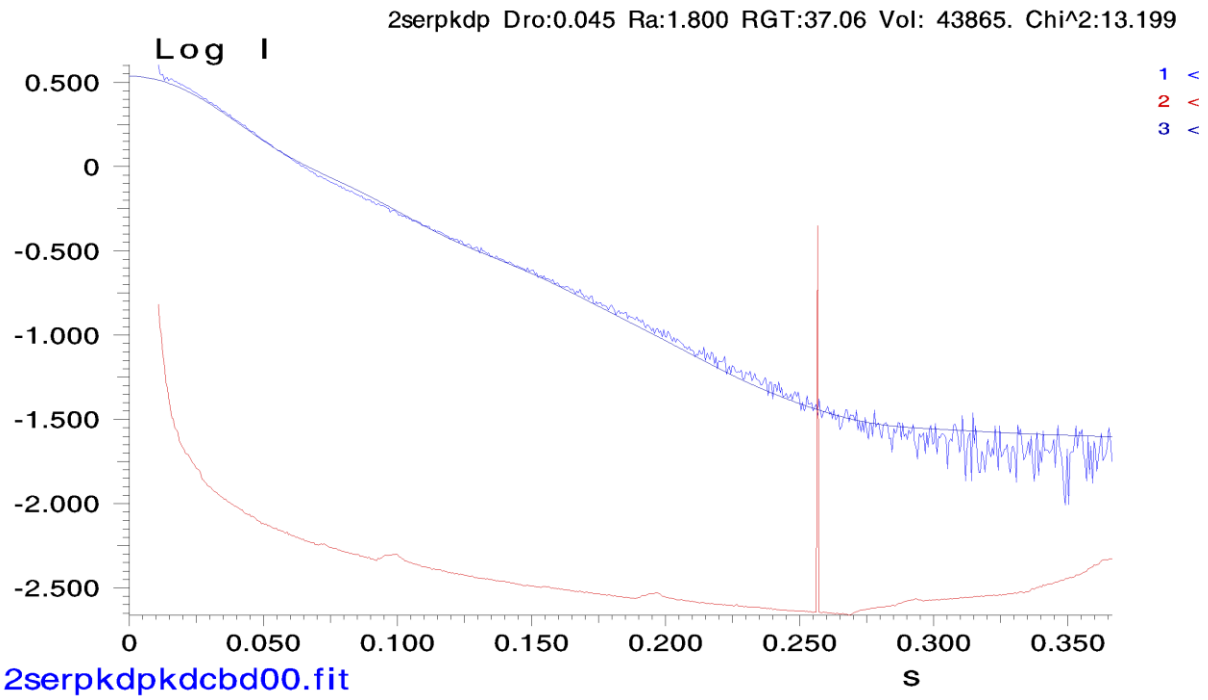


Figure 15: Crysol fit file for 2 Ser PKD1PKD2CBD. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. χ^2 value of 13.199 indicates that fit is good.

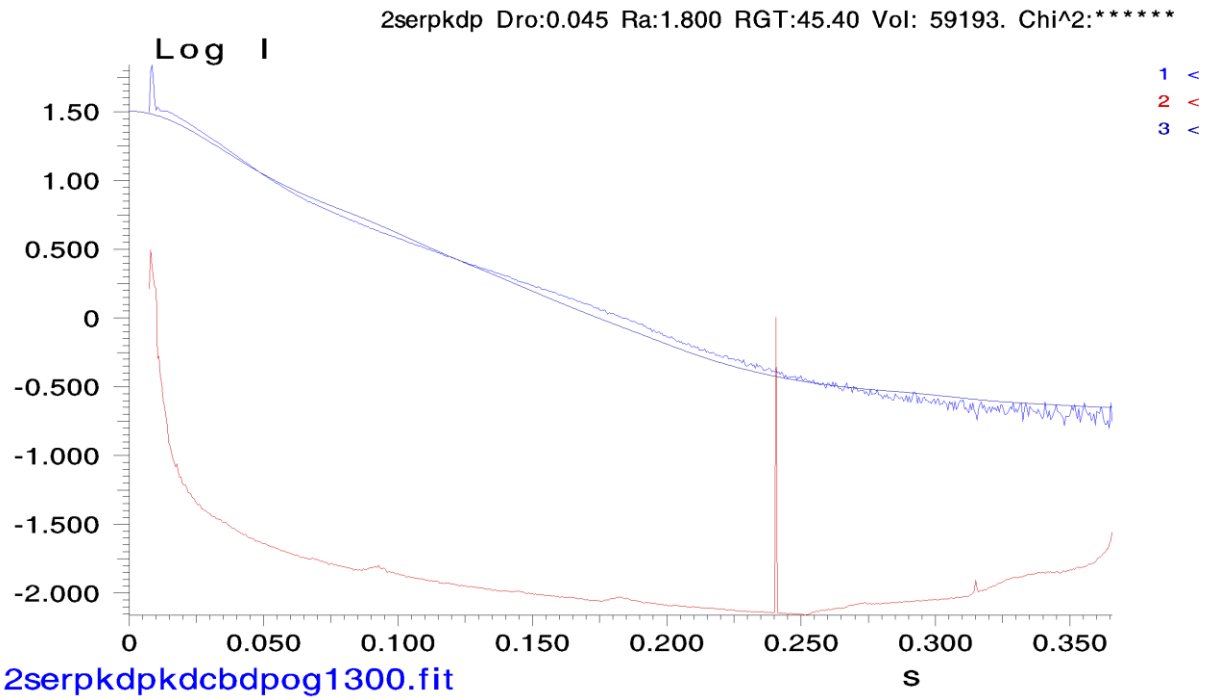


Figure 16: Crysol fit file for 2 Ser PKD1PKD2CBD:[GPRG(POG)$_{13}$]$_3$. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. χ^2 value of 323.06 indicates that fit between the two could be better. Removing low s region may improve fit.

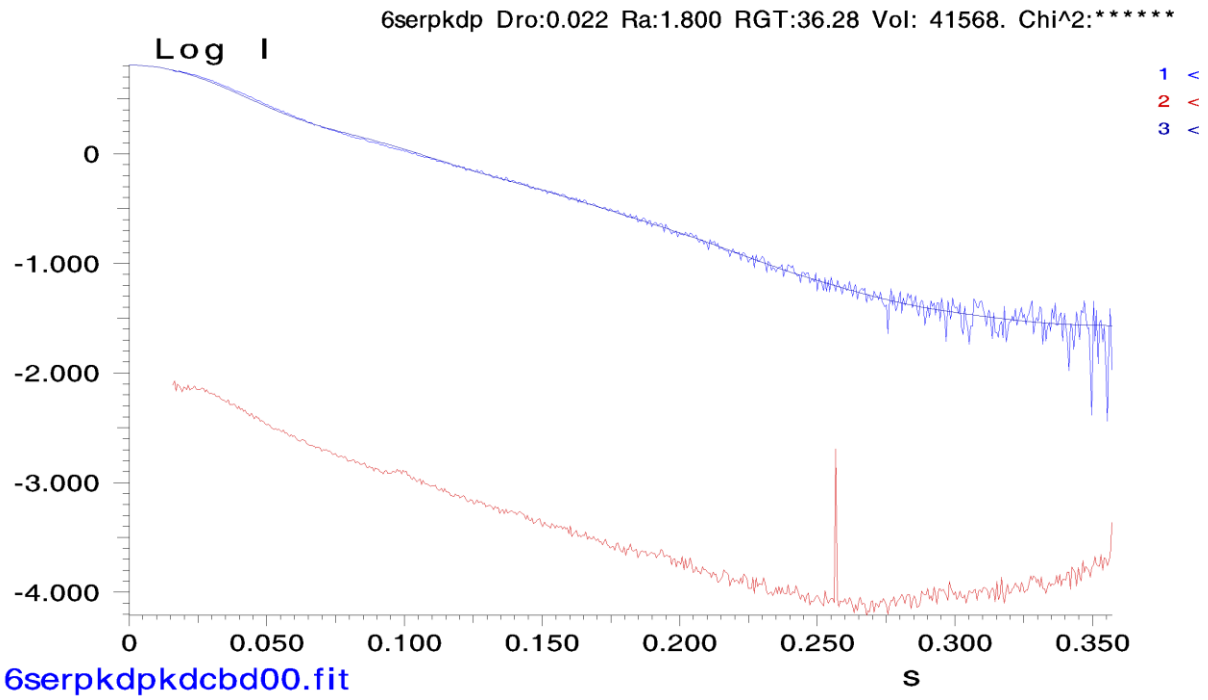


Figure 17: Crysol fit file for 6 Ser PKD1PKD2CBD. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. While the fit between the two appears good χ^2 value of 3085.88 indicates that 6 Ser PKD1PKD2CBD data may need to be reprocessed

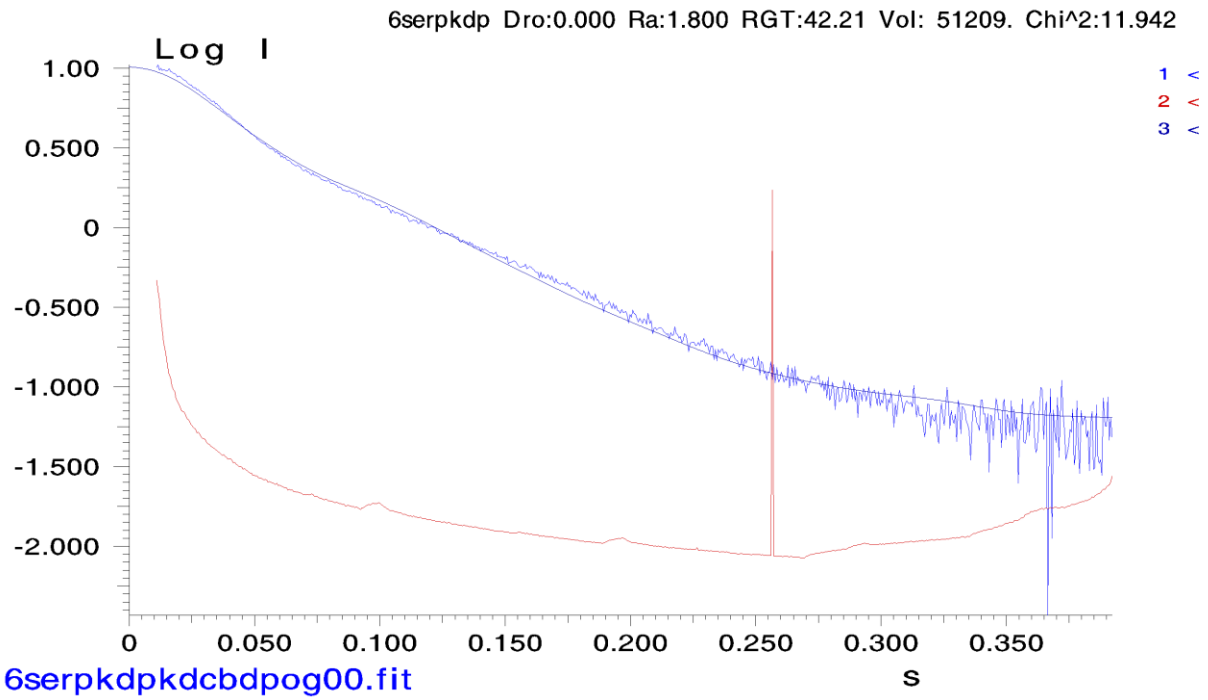


Figure 18: Crysol fit file for 6 Ser PKD1PKD2CBD:[GPRG(POG)$_{13}$]$_3$. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. χ^2 value of 11.942 indicates that fit between the two is good.

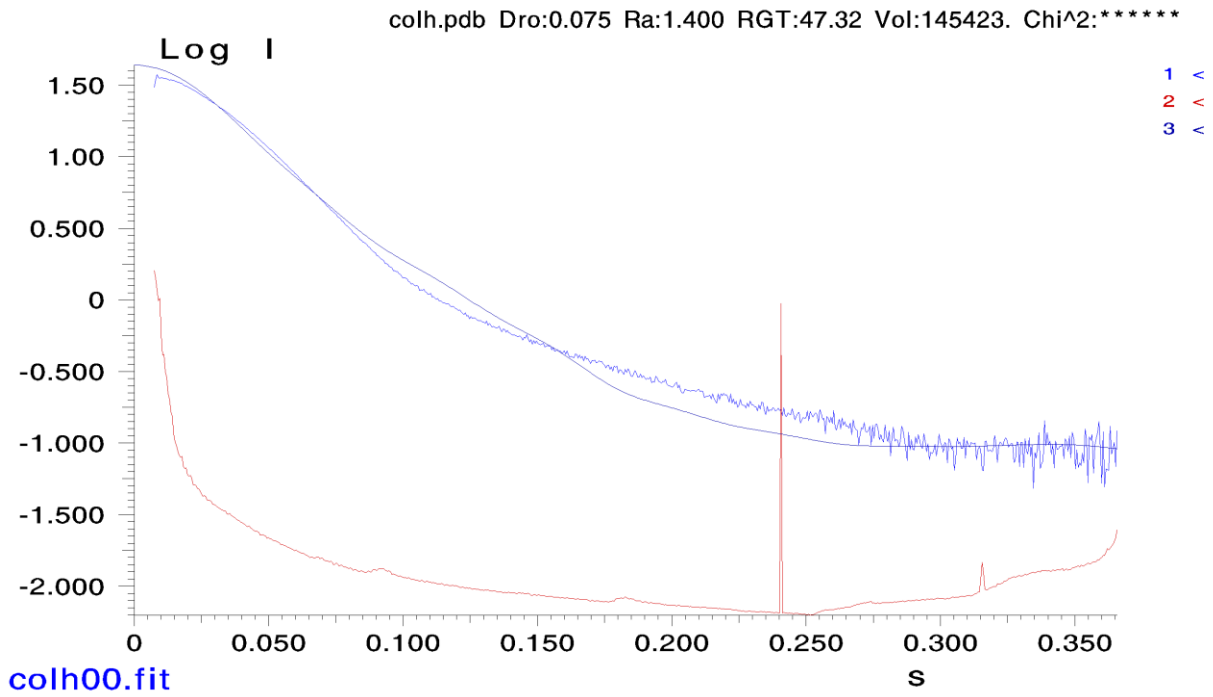
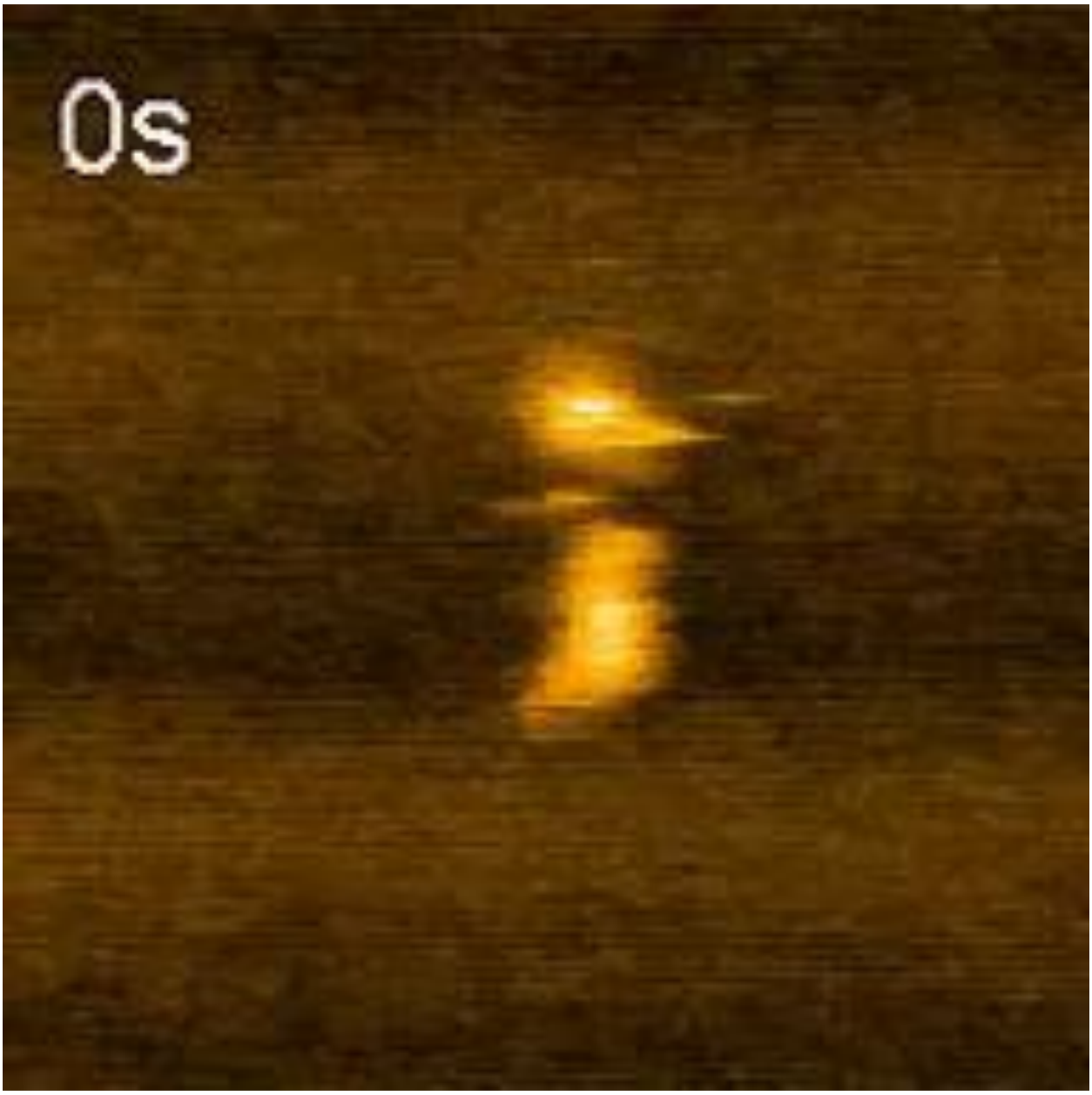
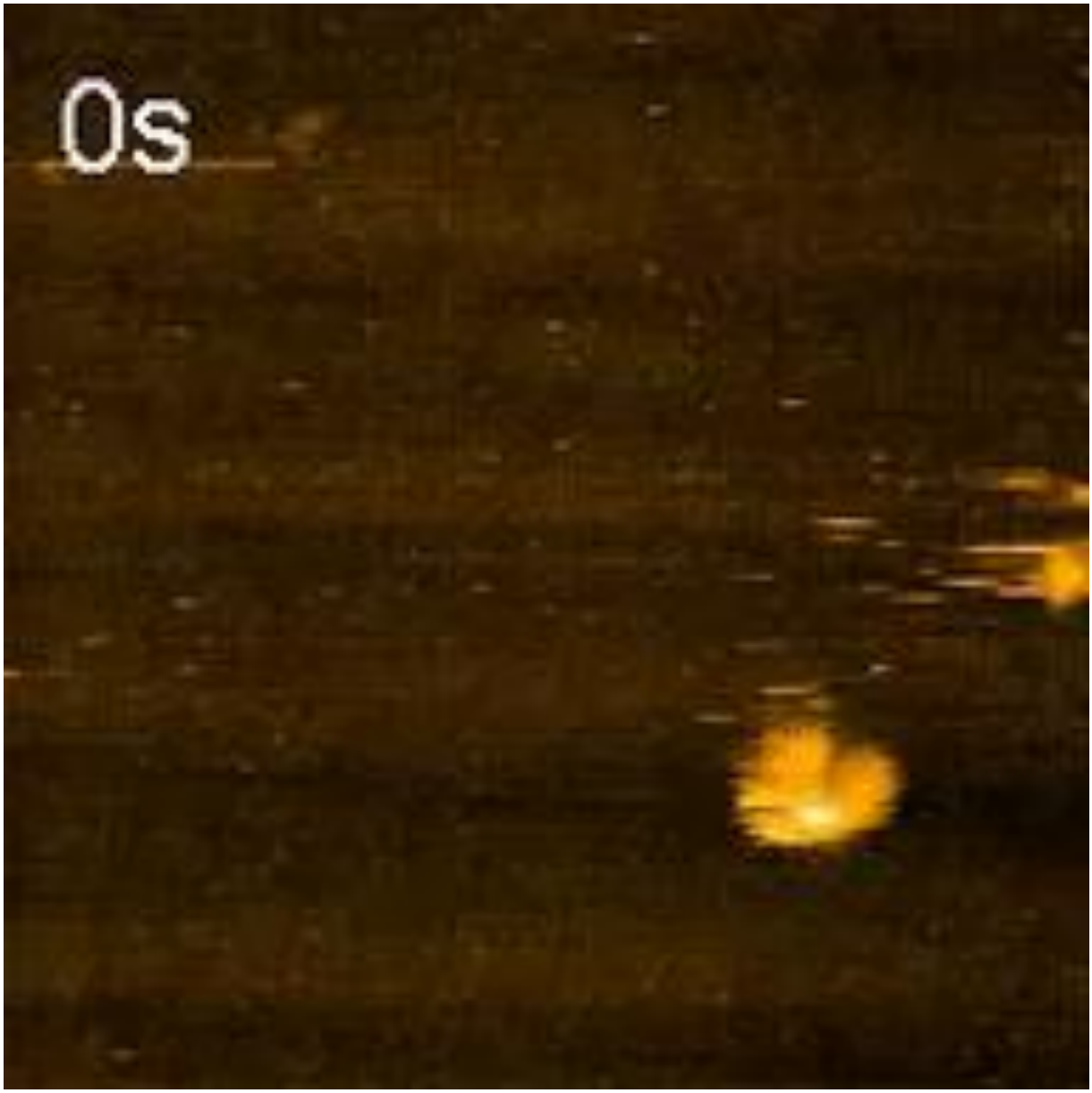


Figure 19: Crysol fit file for full-length Col H. Experimentally derived scattering profile is in blue and scattering profile for crystal structure is in purple. The red graph is the experimentally derived scattering profile in absolute scale. χ^2 value of 321.77 indicates that fit between the two could be better.



Video 1: High speed AFM video for full-length Col G at physiological Ca^{2+} concentration. Full-length protein is highly dynamic



Video 2: High speed AFM video for full-length Col H at physiological Ca^{2+} concentration. Full-length protein is highly dynamic