Structural and Functional Characterization of a Modular Immunoglobulin A Protease from *Thomasclavelia ramosa*

by

Norman Tran

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Biology

Waterloo, Ontario, Canada, 2024

© Norman Tran 2024

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Dr. Joanne Lemieux Professor – University of Alberta (Biochemistry)
Supervisor	Dr. Todd Holyoak Associate Professor – University of Waterloo (Biology)
Internal Member	Dr. David Rose Professor Emeritus – University of Waterloo (Biology)
Internal Member (2)	Dr. Brian Dixon Professor – University of Waterloo (Biology)
Internal-External Member	Dr. David Sychantha Assistant Professor – University of Waterloo (Chemistry)

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Norman Tran was the sole author for Chapters 1 and 5, which were written under the supervision of Dr. Todd Holyoak and were not written for publication.

This thesis consists of three Chapters (2, 3, and 4) written for publication. As the lead author of these chapters, I was responsible for overseeing the projects, conceptualizing and designing experiments, collecting and analyzing data, and drafting and submitting manuscripts.

Exceptions to sole authorship of material are as follows:

Dr. Aaron Frenette assisted in the conceptualization and design of all experiments in Chapters 2, 3, and 4, with emphasis on Chapters 3 and 4. Dr. Aaron Frenette also helped collect and analyze all experimental data in all Chapters. Dr. Aaron Frenette and Dr. Todd Holyoak provided intellectual input on manuscript drafts.

Abstract

Immunoglobulin A proteases (IgAPs) are a diverse group of enzymes secreted from a wide range of mucosal bacteria. These enzymes have convergently evolved to cleave immunoglobulin A (IgA), the main antibody found on the mucosa, as a means of modulating the bacterium's relationship with their host tissues. Due to the various biological functions and biochemical properties of these enzymes, the study of IgAPs has provided multifaceted insight into aspects of mucosal immunity, enzyme structure and function, and the structural basis for substrate specificity.

Only two of three known IgAP enzyme families have been investigated using an in-depth structural and functional approach. This thesis thus aimed to carry out these analyses on the IgAP from *Thomasclavelia ramosa* as a representative member of this last poorly characterized family. X-ray crystallographic, small-angle X-ray scattering, and gel-based kinetic techniques were used to reveal that, unlike the other two IgAP families, the *T. ramosa* IgAP has a truly modular protein architecture that can be split into and produced as distinct minimized domains that retain function. The crystal and solution-scattering structures of various domain constructs were also used to generate a working model for how the *T. ramosa* IgAP recognizes and has high specificity for IgA. This thesis provides the first in-depth biochemical account of this IgAP family and pave the way for advances in clinically relevant IgAP-related research and our understanding of IgAPs as a whole.

Acknowledgments

I give big thanks to Jameson Ngo, who thinks he deserves this PhD. For his endless support and encouragement, I dedicate my PhD to him.

I would like to thank Dr. Aaron Frenette for being a phenomenal friend, inspiration to work with, and someone I can truly rely on. Although we've gone through some tough times together, it has made us stronger, and we've come to realize what we value and want from life. I know we will be able to achieve whatever we set our minds to. I wouldn't have been able to do this PhD without you – thank you very much.

Thank you to Dr. Matthew McLeod for being my first scientific mentor, all the way back when I was still a naïve undergraduate student. You've never stopped mentoring me and I am still very grateful to have you as a friend.

I would like to thank Dr. Todd Holyoak for taking me on as a graduate student, giving me the right opportunities to grow and mature as a researcher, and providing me with the right amount of flexibility to carve my own path forward.

I would also like to thank Dr. Nicole Fraser, Dr. Sarah Barwell, Amy Xu, and Julia Solonenka for their support, friendship, and for listening to my ramblings. Julia has been a constant source of moral support and I am lucky to have gotten to know her.

And finally, thank you to Dr. David Rose and Dr. Brian Dixon for being a part of my committee. I would like to thank Dr. Rose in particular as he's supported me as a scientist from my undergraduate studies all the way until now.

Table of Contents

Examining Committee Membership ii
Author's Declaration iii
Statement of Contributions iv
Abstract v
Acknowledgments vi
List of Figures x
List of Tables xiii
List of Abbreviations xiv

Chapter 1 – Introduction to Immunoglobulin A Proteases and Its Biological Context 1
1.1 – Host-Pathogen Relationships of Mucosal Tissues 1
1.2 – Immunoglobulin A Proteases and Their Role in Immune Evasion 3
1.3 – Immunoglobulin A and Mucosal Immunity 4
1.4 - A Brief History and Description of Immunoglobulin A Proteases 7
1.5 – Immunoglobulin A Proteases as a Therapeutic Target 9
1.6 – Immunoglobulin A Proteases as a Protein Therapeutic for Immunoglobulin A Nephropathy 10
1.7 – Thesis Goals and Approaches 10
1.8 – Chapter References 13
Chapter 2 – Structurally Characterizing Domains of the *Thomasclavelia ramosa* Immunoglobulin A Protease with X-ray Crystallography 19
2.1 – Chapter Introduction 19
2.1.1 – Importance of X-ray Crystallography in Understanding Immunoglobulin A Proteases 19
2.1.2 – Immunoglobulin A Proteases are a Case of Convergent Evolution 20
2.1.3 – The S6 Family of Immunoglobulin A1 Proteases 22
2.1.4 – The M26 Family of Immunoglobulin A1 Proteases 26
2.1.5 – Introduction to Thomasclavelia ramosa and Its Immunoglobulin A Protease 29
2.1.6 – Chapter Objectives 31
2.2 – Materials and Methods 32
2.2.1 – Expression Vector Cloning 32
2.2.2 – Protein Expression and Purification 33
2.2.3 – Protein Crystallization 35
2.2.4 – Metal-Chelation and Zinc-Supplementation Crystallographic Experiments 36
2.2.5 – Crystallographic Data Collection and Processing 37

2.3 – Results and Discussion 38
2.3.1 – Domain Prediction and Construct Design 38
2.3.2 – Protein Expression, Purification, and Crystallization 41
2.3.3 – Crystal Structure of the N-terminal Domain 49
2.3.4 – Crystal Structure of the Protease Domain 52
2.3.5 – The Role of Zinc in the Protease Domain 54
2.3.6 – The E540A Mutation Quenches Disorder in the Protease Domain Active Site 61
2.3.7 – Hypothetical Mechanisms of Substrate Recognition in the Protease Domain 64
2.3.8 – The Protease Domain and Its Relationship to the M64 Protease Family 72
2.4 – Chapter Conclusions and Future Directions 74
2.5 – Chapter References 75
Chapter 3 – Structurally Characterizing the *Thomasclavelia ramosa* Immunoglobulin A Protease with Small-Angle X-ray Scattering Techniques 84
3.1 – Chapter Introduction 84
3.1.1 – Importance of Solution Scattering in Understanding Immunoglobulin A Proteases 84
3.1.2 – Introduction to Small-Angle X-ray Scattering and Its Limitations 85
3.1.3 – The Small-Angle X-ray Scattering Curve and Its Transformations 86
3.1.4 – Structural Validation with Small-Angle X-ray Scattering Data 88
3.1.5 – Chapter Objectives 88
3.2 – Materials and Methods 89
3.2.1 – Small-Angle X-ray Scattering Sample Preparation and Data Collection 89
3.2.2 – Small-Angle X-ray Scattering Data Processing 89
3.3 – Results and Discussion 91
3.3.1 – Summary of Small-Angle X-ray Scattering Data 91
3.3.2 – Solution Scattering Data Supports a Modular Domain Architecture for the Protease 93
3.3.3 – The Full-Length Protease Has an Extended Conformation in Solution 98
3.3.4 – The Full-Length Protease in the Context of Immunoglobulin A 102
3.4 – Chapter Summary and Next Steps 103
3.5 – Chapter References 105
Chapter 4 – Enzymologically Characterizing the *Thomasclavelia ramosa* Immunoglobulin A Protease 109
4.1 – Chapter Introduction 109
4.1.1 – Importance of Kinetic Characterization in Understanding Immunoglobulin A Proteases 109
4.1.2 – Chapter Objectives 109
4.1.3 – Limitations of Existing Immunoglobulin A Protease Kinetic Assays 110

4.1.4 – A Brief Synopsis of Michaelis-Menten Kinetics 112
4.2 – Materials and Methods 115
4.2.1 – Preparation of the Monomeric Immunoglobulin A1 Substrate 115
4.2.2 – Preparation of the Thomasclavelia ramosa Immunoglobulin A Protease Enzymes 116
4.2.3 – Preparation of the Fully Cleaved Immunoglobulin A1 Internal Standards 116
4.2.4 – Standardized Protocol for the Gel-Based Kinetic Assay 116
4.2.5 – Gel Image Processing and Michaelis-Menten Curve Fitting 117
4.2.6 –Peptide Cleavage Analysis by MALDI-ToF Mass Spectrometry 118
4.3 – Results and Discussion 119
4.3.1 – Developing a Versatile Gel-Based Kinetic Assay for Immunoglobulin A Proteases 119
4.3.2 – Limitations to the Gel-Based Immunoglobulin A Protease Kinetic Assay 128
4.3.3 – Kinetic Comparisons of Thomasclavelia ramosa Immunoglobulin A Protease Constructs . 130
4.3.4 – Verifying Substrate Specificity for the Reduced Thomasclavelia ramosa Construct 133
4.4 – Chapter Summary and Next Steps 134
4.5 – Chapter References 136
Chapter 5 – Concluding Remarks and Future Directions 139
5.1 – Role of Structural Biology and Enzymology in Understanding Immunoglobulin A Proteases 139
5.2 – Summary of Experimental Findings 140
5.3 – Impact of Work and Next Steps 142
5.4 – Chapter References 144
Appendix A 147

List of Figures

Figure 1.1.1 – Host-Commensal Relationships. 2

Figure 1.2.1 – Role of Virulence Factors in Subverting Host Immunity. 3

Figure 1.3.1 – Quaternary Structure of Monomeric Immunoglobulin A and Its Subclasses. 4

Figure 1.3.2 – Oligomeric States of Immunoglobulin A. 5

Figure 1.4.1 – Immunoglobulin A and Its Protease Cleavage Sites. 7

Figure 2.1.2.1 – Immunoglobulin A Protease Families and Their Representative Members. 21

Figure 2.1.3.1 – Secretion Pathway of S6 Immunoglobulin A1 Proteases. 23

Figure 2.1.3.2 – Annotated Tertiary Structure of the Type 1 *Haemophilus influenzae* Immunoglobulin A1 Protease. 24

Figure 2.1.4.1 – Putative Secretion Pathway of M26 Immunoglobulin A1 Proteases. 26

Figure 2.1.4.2 – Annotated Tertiary Structure of the *Streptococcus pneumoniae* Immunoglobulin A1 Protease. 27

Figure 2.1.5.1 – Secretion Pathway of the *Thomasclavelia ramosa* Immunoglobulin A Protease. 30

Figure 2.3.1.1 – Putative Domains of the *Thomasclavelia ramosa* Immunoglobulin A Protease. 38

Figure 2.3.1.2 – RosettaFold-Predicted Model for the Protease Domain and Its Adjacent C-terminal Domain. 39

Figure 2.3.1.3 – AlphaFold2-Predicted Model for the C-terminal Putative Zinc-Binding Domains. 39

Figure 2.3.2.1 – Purity of Recombinantly Expressed *Thomasclavelia ramosa* Immunoglobulin A Protease Constructs. 41

Figure 2.3.2.2 –Morphologies of the NTD and MD+CTD1 Crystals. 42

Figure 2.3.3.1 – Crystal Structure of the *Thomasclavelia ramosa* Immunoglobulin A Protease N-terminal Domain. 49

Figure 2.3.3.2 – The N-terminal Domain and Its Relationship to Bacteriophage Proteins and C-type Lectins. 51

Figure 2.3.4.1 – Crystal Structure of the *Thomasclavelia ramosa* Immunoglobulin A Protease Middle Domain. 52

Figure 2.3.4.2 – The Non-Catalytic Zinc-Binding Site in the Protease Domain. 54

Figure 2.3.5.1 – Catalytic Mechanism of Zinc Metalloproteases. 56

Figure 2.3.5.2 – Conformational Flexibility in the Active Site of the Wild Type Protease Domain. 57

Figure 2.3.5.3 – Anomalous Density Analysis for the Wild Type Protease Domain. 59

Figure 2.3.5.4 – Inherent Disorder in the Active-Site Zinc-Binding Pocket of the Protease Domain. 60
Figure 2.3.6.1 – Anomalous Density Analysis for the Protease Domain E540A Mutant. 61
Figure 2.3.6.2 – The E540A Mutation Orders the Active Site of the Protease Domain. 63
Figure 2.3.7.1 – Normalized B-factor Comparisons Between the Wild Type and E540A Protease Domain Structures. 64
Figure 2.3.7.2 – Protease and Peptide Substrate Nomenclature. 66
Figure 2.3.7.3 – Hypothesized Subsites in the Active-Site Groove of the Protease Domain. 67
Figure 2.3.7.4 – Working Model of the Protease-Peptide Complex. 68
Figure 2.3.7.5 – Working Model of the IgA C_H1-Protease Domain Complex. 70
Figure 2.3.7.6 – Working Model of the IgA-Protease Domain Complex. 71
Figure 2.3.8.1 – Structural Comparisons Between Members of the M64 Protease Family. 73
Figure 3.1.3.1 – Experimental Setup for Small-Angle X-ray Scattering. 86
Figure 3.3.2.1 – SAXS Profiles and Transformations of the Globular N-terminal and Protease Domains. 93
Figure 3.3.2.2 – SAXS Profiles and Transformations of the Rod-Like C-terminal Domain 2 and Human α-Actinin. 95
Figure 3.3.2.3 – Simulated Scatter of Crystal Structures and AlphaFold2 Models Fit Raw Scattering Data. 97
Figure 3.3.3.1 – SAXS Profiles of Multidomain *Thomasclavelia ramosa* Immunoglobulin A Protease Constructs. 98
Figure 3.3.3.2 – Simulated Scatter of CORAL-Relaxed and AlphaFold2 Models for the NTD+MD Construct. 99
Figure 3.3.3.3 – Simulated Scatter of CORAL-Relaxed and AlphaFold2 Models for Constructs with the C-terminal Domains. 101
Figure 3.3.4.1 – Working Model of the Full-Length Enzyme-Substrate Complex. 102
Figure 4.3.1.1 – Cleavage Pattern of IgA1 on an SDS-PAGE Gel. 120
Figure 4.3.1.2 – Testing Linearity of the Gel-Based Kinetic Assay Using a Substrate Standard Curve. 121
Figure 4.3.1.3 – Testing Linearity of the Gel-Based Kinetic Assay Using a Product Standard Curve. 122
Figure 4.3.1.4 – The Cleaved Heavy-Chain Fragment Associated with the Fc Provided the Most Reliable Readout for the Gel-Based Assay. 123
Figure 4.3.1.5 – The Gel-Based Assay is Capable of Determining Initial Rates Under Michaelis-Menten Conditions. 124

Figure 4.3.1.6 – The *Thomasclavelia ramosa* Immunoglobulin A Protease is Robust Against Product Inhibition. 126

Figure 4.3.3.1 – Michaelis-Menten Curves for Various Domain Constructs. 130

Figure 4.3.4.1 – The *Thomasclavelia ramosa* Immunoglobulin A Protease Middle Domain Does Not Nonspecifically Cleave the IgA1 Hinge Peptide. 133

List of Tables

Table 2.3.2.1 – Data Collection and Refinement Statistics. 43
Table 3.3.1.1 – Small-Angle X-ray Scattering Data Details and Statistics. 91
Table 4.3.3.1 – Michaelis-Menten Parameters for Various Domain Constructs. 131

List of Abbreviations

General

Immunoglobulin A protease	IgAP
Immunoglobulin A1 protease	IgA1P
Immunoglobulin	Ig
Immunoglobulin A	IgA
Immunoglobulin A1	IgA1
Immunoglobulin A2, allotype A2m(1)	IgA2m(1)
Immunoglobulin A2, allotype A2m(2)	IgA2m(2)
Immunoglobulin A2, allotype A2n	IgA2n
Antibody's antigen-binding domain	Fab
Antibody's constant/crystallizable domain	Fc
Thomasclavelia ramosa IgAP N-terminal domain	NTD
Thomasclavelia ramosa IgAP protease/middle domain	MD
Thomasclavelia ramosa IgAP C-terminal domain #1	CTD1
Thomasclavelia ramosa IgAP C-terminal domain #2	CTD2
Thomasclavelia ramosa IgAP C-terminal domain #3	CTD3
Thomasclavelia ramosa IgAP C-terminal domain #4	CTD4
Full-length *Thomasclavelia ramosa* IgAP	FL
Sodium dodecylsulfate polyacrylamide gel electrophoresis	SDS-PAGE
Molecular-weight cut-off	MWCO
2-[4-(2-Hydroxyethyl)piperazin-1-yl]ethane-1-sulfonic acid	HEPES
Tris(2-carboxyethyl)phosphine	TCEP

Chapter 1

Joining chain	J-chain
Secretory chain	SC
Polymeric immunoglobulin receptor	pIgR
Immunoglobulin M	IgM
Lower respiratory tract infection	LRTI
Immunoglobulin A nephropathy	IgAN

Chapter 2

Ethylenediaminetetracetic acid	EDTA
Cryoelectron microscopy	cryoEM
Signal peptide	SP
Zinc metalloprotease	Zmp
Streptococcus pneumoniae IgA1P N-terminal domain	NTD (introduction)
Streptococcus pneumoniae IgA1P middle domain	MD (introduction)
Streptococcus pneumoniae IgA1P C-terminal domain	CTD (introduction)
Inflammatory bowel disease	IBD
Maltose-binding protein resin	MBP
Hexahistidine tag	His-tag

Lysogeny broth	LB
Nickel nitrilotriacetic acid resin	NiNTA
Column volume	CV
Phenylmethylsulfonyl fluoride	PMSF
Tris(hydroxymethyl)aminomethane	Tris
Polyethylene glycol	PEG
N,N-Bis(2-hydroxyethyl)glycine	BICINE
Bis-(2-hydroxy-ethyl)-amino-tris(hydroxymethyl)-methane	Bis-Tris
Predicted aligned error	PAE
Domain of unknown function	DUF
Immunoglobulin fold	Ig fold
Alpha-carbon root mean square deviation	C_α-RMSD
Nuclear magnetic resonance	NMR
Matrix metalloprotease-8	MMP-8
Constant heavy chain immunoglobulin domain #1	C_H1
Constant heavy chain immunoglobulin domain #2	C_H2

Chapter 3

Small-angle X-ray scattering	SAXS
Complexes with random loops program	CORAL
Molecular weight	MW

Chapter 4

Surface plasmon resonance	SPR
Iodine-125-labelled myeloma IgA	^{125}I-IgA
Enzyme-linked immunosorbent assay	ELISA
Intact IgA1 heavy chain	HC
IgA1 light chain	LC
IgA1 heavy chain product fragment associated with Fc	H1
IgA1 heavy chain product fragment associated with Fab	H2
Trifluoroacetic acid	TFA
α-cyano-4-hydroxycinnamic acid	HCCA
Matrix-assisted laser desorption/ionization time-of-flight	MALDI-ToF

Chapter 1 – Introduction to Immunoglobulin A Proteases and Its Biological Context

1.1 – Host-Pathogen Relationships of Mucosal Tissues

Mucosal tissues are surfaces that are essential in facilitating the exchange of material in and out of the body.[1] These include the exchange of gases in the respiratory system,[2] nutrients and waste products in the gastrointestinal tract,[3] and the secretion and storage of fluids in the urogenital system.[4] Although they have structural commonalities, each type of mucosal tissue has evolved its own set of biological mechanisms to cater to its niche as a result of different physiological requirements and biological pressures.[2–5]

Mucosal tissues are the primary entryway into the body as they account for the largest surface area that is exposed to the external environment.[2–4] These vulnerable regions require versatile mechanisms of defense, which differ from tissue to tissue depending on the needs of each location.[1] For example, because the gastrointestinal tract encounters a plethora of microorganisms and food antigens, it has evolved a unique immune strategy that strikes a balance between pathogen response and maintaining the integrity of the mucosal surface by modulating the tissue's level of inflammation.[3,5,6] In contrast, the urogenital tract is normally considered a sterile environment and has a tightly sealed epithelia to prevent the re-entry of toxic compounds in the urine.[7] However in the bladder mucosa, inflammatory responses are dampened to prioritize the integrity of the epithelial layer, even in the presence of pathogens.[8]

There are, however, mechanisms of defense that are common to all mucosal tissues. Mucosal tissues are lined with a single layer of epithelial cells that form a tightly sealed barrier that physically blocks material and microorganisms from exchanging with the internal environment.[1] This mucosal epithelium is also covered by a thick heterogeneous web of glycoproteins collectively known as mucins, which act as a physical spacer between the external environment and the epithelium.[9] The mucin layer is also semi-permeable, allowing for host-produced antimicrobial agents and other immune factors to help with host defenses.[10,11] In addition to these innate defenses, the adaptive immune system can be trained to recognize specific foreign entities to mediate a stronger innate response through the use of antibodies.[12]

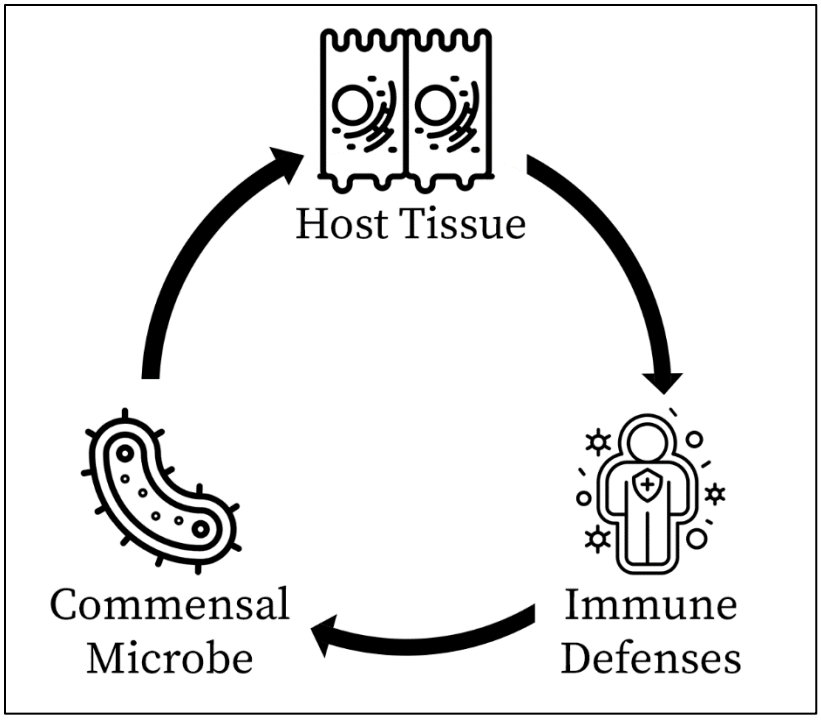


Figure 1.1.1 – Host-Commensal Relationships. A simplistic view of the host-commensal relationship as a cycle of interactions between the host, innate and adaptive immune defenses, and commensal microorganisms.

Creative Commons images were made by dDara, Uswa KDT, and Ricardo Job-Reese from Noun Project.

The study of host-pathogen relationships questions the interactions between microorganisms and host defense systems, and how these interactions lead to their resulting outcome - disease or lack thereof.[13] The goal is to understand why certain microorganisms are pathogenic, how they become pathogenic, and what mechanisms facilitate this interaction.[14] The same approach is also used to understand why commensal microorganisms remain non-pathogenic or why certain species are considered to be opportunistic pathogens.[13,14] This host-commensal relationship can be summarized simplistically with the process described in *Fig. 1.1.1*, whereby a commensal microorganism first encounters an innate or adaptive immune defense, the host reacts, and the microorganism adapts to establish a homeostatic niche with the host.

1.2 – Immunoglobulin A Proteases and Their Role in Immune Evasion

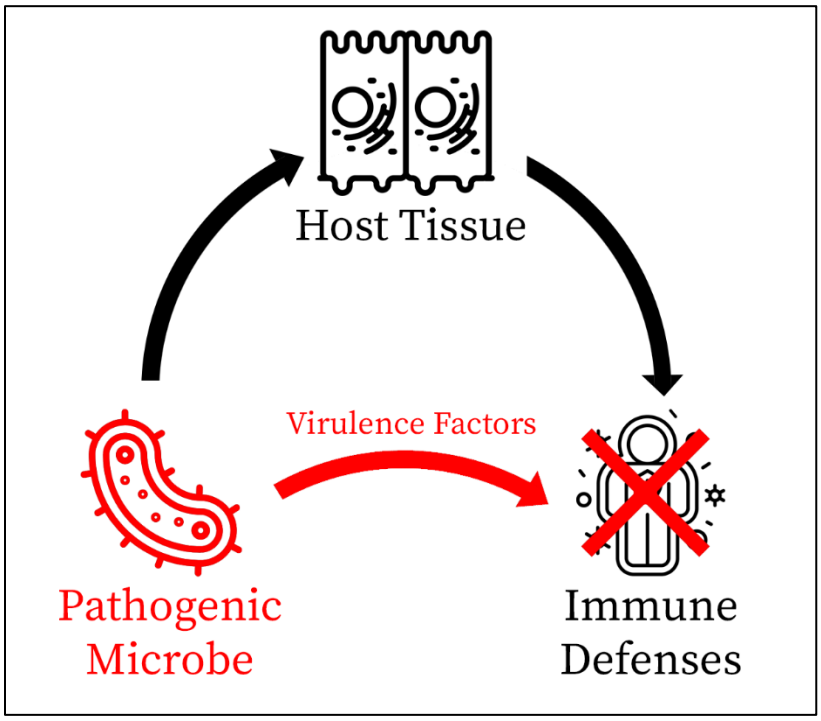


Figure 1.2.1 – Role of Virulence Factors in Subverting Host Immunity. A simplistic view of the host-pathogen relationship whereby pathogens are able to subvert the immune system via the production of virulence factors.

Creative Commons images were made by dDara, Uswa KDT, and Ricardo Job-Reese from Noun Project.

As is often the case, the relationship between host and microorganism is not always commensal (*Fig. 1.1.1*); the cycle can easily turn into a host-pathogen relationship if the microorganism finds efficient ways to subvert host defense systems (*Fig. 1.2.1*).[14,15] In the context of mucosal tissues, microorganisms have evolved virulence factors to chemically alter the structure of the mucus layer, adhere tightly to mucosal tissues to avoid being shed, and dismantle epithelial tight junctions to invade deeper into the tissue.[9–11,14] These virulence factors are thus necessary for these microorganisms to establish their pathogenic niche.[14,15]

The adaptive mucosal immune system is also not safe from the threat of virulence factors (*Fig. 1.2.1*). The main immunoglobulin (Ig) type found on mucosal surfaces is of class A – immunoglobulin A (IgA).[16] Because of IgA's role in adaptive mucosal immunity (*Section 1.3*), it is not surprising that some microorganisms have evolved to secrete a group of enzymes known as immunoglobulin A proteases (IgAPs) to selectively proteolyze and inactivate IgA (*Section 1.4*).[17,18] These proteases contribute significantly to the pathogenicities of mucosal pathogens and thus play an important role in their host-pathogen relationships.[17] To better understand these interactions, this thesis aims to shed light on the relationship between these protease-producing microorganisms and the host through a *comprehensive biochemical understanding of the structure*

and function of these proteases, with a focus on the IgAP from Thomasclavelia ramosa. The following section will detail the role of IgA in mucosal immunity to provide some biological context and reiterate the importance of IgAPs in mucosal host-microorganism relationships.

1.3 – Immunoglobulin A and Mucosal Immunity

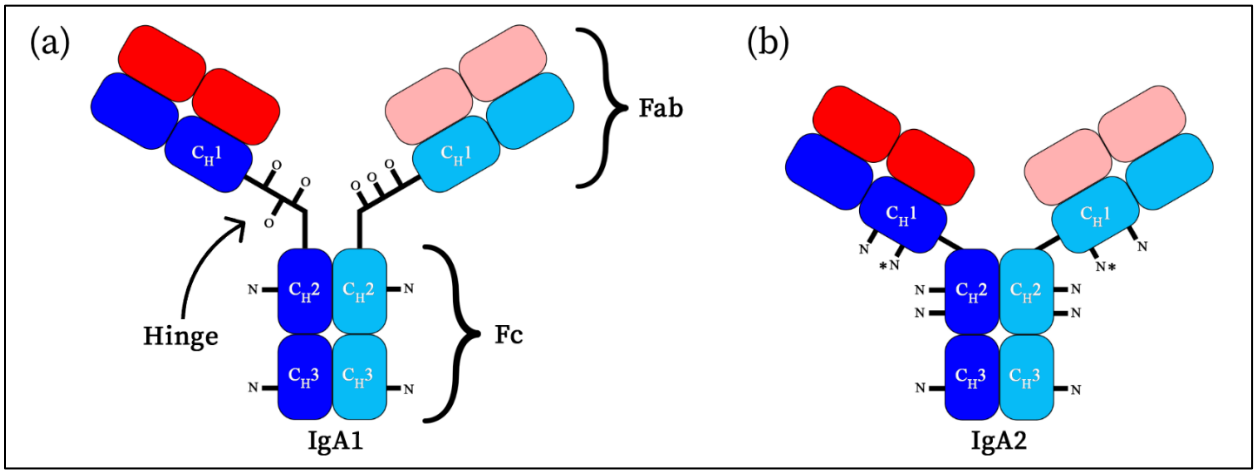


Figure 1.3.1 – Quaternary Structure of Monomeric Immunoglobulin A and Its Subclasses. The structures of **(a)** IgA1 and **(b)** IgA2 are shown based on the arrangement of their immunoglobulin domains (rounded rectangles). Each IgA monomer is composed of two copies of the IgA heavy (light and dark blue) and light (faded and dark red) chains. These chains fold into antigen-binding (Fab) and constant (Fc) domains which are linked by a flexible hinge region. The IgA heavy chains are heavily N-linked glycosylated (sites denoted with 'N'), with the IgA2 allotypes A2m(2) and A2n containing an extra N-linked glycosylation on its first constant heavy chain domain (C_H1; denoted with '*'). The IgA1 hinge is elongated relative to IgA2 and contains several serine and threonine resides that are O-linked glycosylated (sites denoted with 'O'). Although the figure depicts three O-linked glycosylation sites per hinge, the number and location of glycosylated sites is highly variable.

IgA has a canonical T- or Y-shaped immunoglobulin structure (*Fig. 1.3.1*), a quaternary structure consisting of two heavy and two light chains associated via non-covalent interactions and covalent disulfide bonds.[19] These chains are heavily N-linked glycosylated and fold into a string of modular beta-sandwich immunoglobulin domains.[16] The "wings" of the antibody are made up of both heavy and light chains and function to bind foreign antigens (antigen-binding fragment; Fab).[19] The "stem" of the Y shape is made up of two heavy chains and binds to cell receptors to illicit immune responses (crystallizable or constant fragment; Fc).[19] The region that links the IgA Fabs and Fc is known as the hinge region.[19]

Human IgA is produced as two subclasses – subclass 1 (IgA1) and 2 (IgA2).[19] The major structural difference between these subclasses is the extended hinge region found in IgA1 (*Fig.*

1.3.1a), which is thought to aid in the crosslinking of more distant antigens.[19] This feature is not found in IgA2 nor in the IgA of Old-World monkeys, who only have one IgA heavy-chain gene.[20] The IgA1 hinge consists of an octapeptide repeat of a proline-, serine-, and threonine-rich sequence (PSTPPTPS) that is subjected to heavy O-linked glycosylation.[19] Although this region is generally considered proteolytically resistant, it is the target of specific proteolysis by IgAPs as a mechanism of immune evasion (*Section 1.2*).

IgA2 can also be produced as several allotypes – A2m(1), A2m(2), and A2n – which have distinct structural differences in their disulfide-bond crosslinking and glycosylation profiles (*Fig. 1.3.1b*).[19,21] IgA2m(1) has been hypothesized to be more closely related to IgA1 in that it might have been a result of a recombination event between an ancestral IgA2 and IgA1 heavy-chain genes.[21,22] The function and prevalence of IgA2n in the population has yet to be characterized.[23]

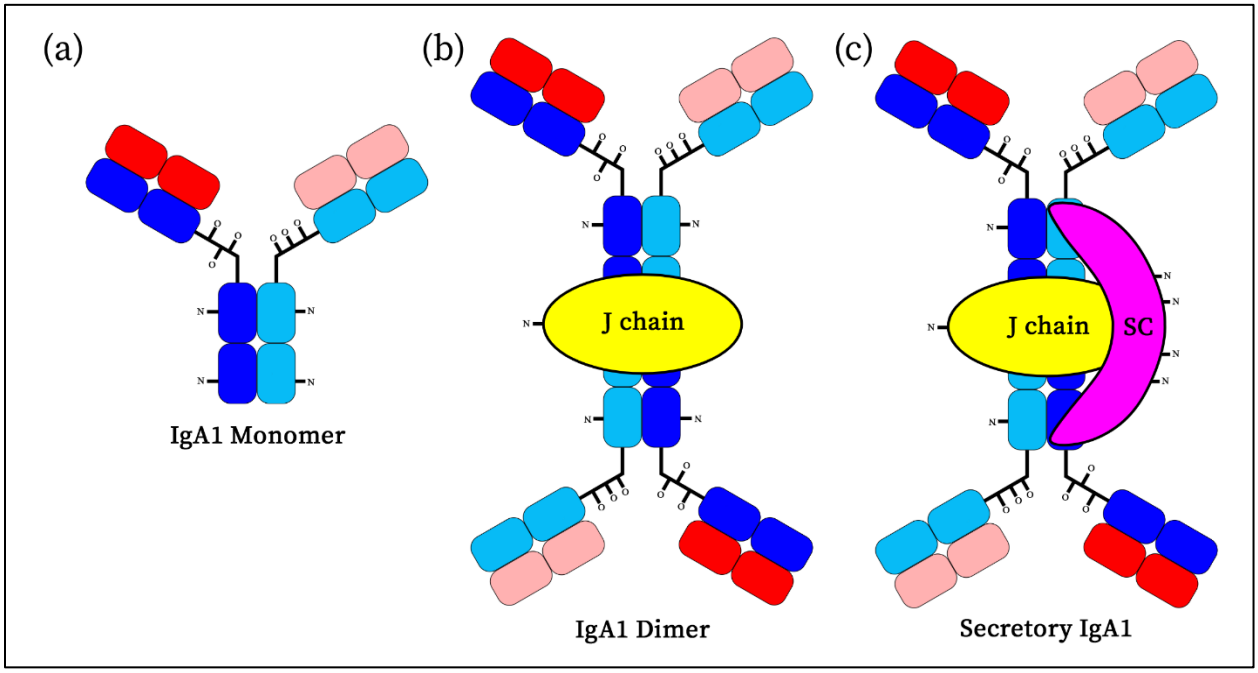


Figure 1.3.2 – Oligomeric States of Immunoglobulin A. IgA can form higher oligomeric states depending on its function and location. **(a)** An IgA monomer is comprised of two heavy and light chains. **(b)** Two monomers can be made into a dimeric IgA molecule via covalent coupling with the joining (J) chain end-to-end at their Fc domains. **(c)** Through the process of IgA transcytosis, this IgA dimer is covalently attached to the ectodomain of the polymeric Ig receptor (also known as the secretory chain; SC) to form secretory IgA. Only IgA1 is shown in this figure for simplicity, however any form of IgA1 or IgA2 and its allotypes can form these oligomers. These illustrations follow the same style and annotations as *Fig. 1.3.1*.

The textbook Y-shaped structure of an antibody is the lowest oligomeric state of IgA and is known as monomeric IgA (*Fig. 1.3.2a*).[23] This is the predominant form of IgA in sera.[24] IgA

molecules can also be synthesized as dimers, which are joined end-to-end at the Fc via covalent attachment to a joining chain (J-chain; *Fig. 1.3.2b*).[23] These dimers are secreted by B cells into the lamina propria, the layer of connective tissue underlying mucosal surfaces, before they are bound by the polymeric immunoglobulin receptor (pIgR) on the basolateral surface of the mucosal epithelium and undergo transcytosis towards the mucosal surface.[23] During this process, the ectodomain of the pIgR is covalently coupled to the dimeric IgA at its Fcs and proteolyzed from the vesicle membrane to form the secretory component (SC).[23] This dimeric IgA-SC complex is known as secretory IgA (*Fig. 1.3.2c*) and is the predominant oligomeric state of IgA found in mucosal secretions.[23] Like in IgM, the SC can form even higher oligomeric complexes with IgA, with the structures of IgA trimers, tetramers, and pentamers in complex with SC having been determined with cryo-electron microscopy.[25,26]

An additional level of biological complexity is seen when comparing the IgA profiles between mucosal surfaces as they can contain different proportions of IgA1 and IgA2, each in a mixture of oligomeric states.[26] Serum IgA is predominantly of the IgA1 subclass and monomeric,[23,27] while secretory IgA1 dominates the upper respiratory tract, nasal cavity, and oral cavity.[28,29] The female reproductive tract contains primarily secretory IgA, with an approximately equal proportion of both subclasses.[4,30] On the other hand, the gastrointestinal tract switches the dominant secretory IgA subclass over the length of the intestine: the primary subclass in the stomach, duodenum, and jejunum is IgA1, which transitions to IgA2 in the ileum and the rest of the large intestine.[28,31]

Secretory IgA primarily acts through a non-inflammatory pathway, binding to and neutralizing antigens via opsonization and agglutination to allow them to pass through the mucosal tissue without eliciting an inflammatory response.[32] This process of immune exclusion functions to maintain the integrity of the mucosal epithelia (*Section 1.1*).[3,5,6] This is primarily carried out in the gut by a subset of innate-like secretory IgA molecules that recognize common structural elements of bacteria and viruses similar to how the innate immune system recognizes pattern-associated molecular patterns.[33] Glycans on secretory IgA also act as competitive inhibitors for virulence factors and as agglutination agents to bolster the effects of immune exclusion.[34] In contrast to secretory IgA, monomeric IgA can bind to the FcαRI receptor on immune cells to elicit

a direct inflammatory response.[35] However, monomeric IgA can also elicit an anti-inflammatory response via FcαRI by inducing IL-10 expression, a general anti-inflammatory cytokine.[36]

Despite the complexities of how IgA contributes to innate and adaptive immunity, there is no doubt that IgA plays a critical role in the homeostasis and defense of mucosal surfaces. As such, there is a clear evolutionary advantage for microorganisms to produce proteases to inactivate IgA. The following sections will provide some historical context for IgAPs and emphasize the applications of basic research on these proteases.

1.4 - A Brief History and Description of Immunoglobulin A Proteases

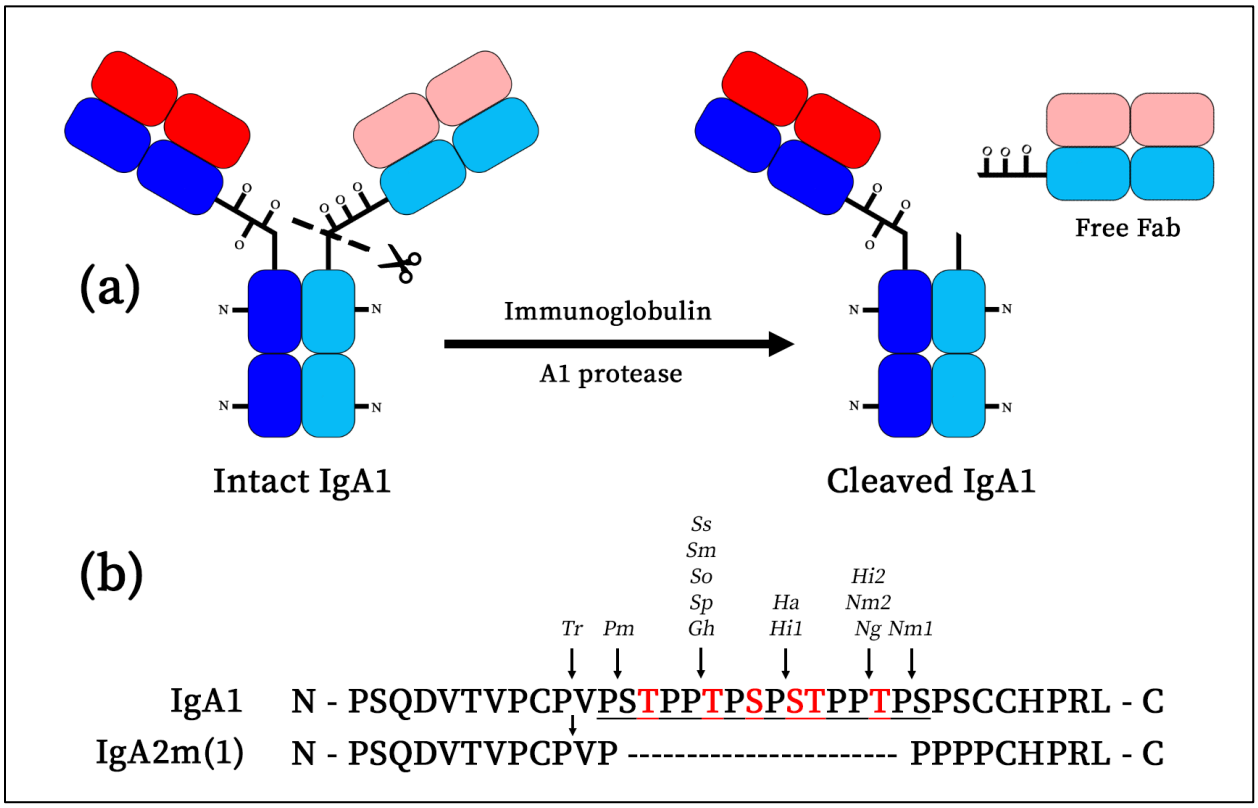


Figure 1.4.1 – Immunoglobulin A and Its Protease Cleavage Sites. (a) IgA is cleaved at its hinge by immunoglobulin A proteases (IgAPs) to separate the Fab and Fc domains. **(b)** Although most IgAPs cleave IgA1 in its hinge (underlined sequence; red residues are sites of O-linked glycosylation), the protease from *Thomasclavelia ramosa* (*Tr*) is the only known IgAP to cleave both IgA1 and IgA2 allotype A2m(1). The IgA1Ps from *Prevotella melaninogenicus* (*Pm*), *Streptococcus sanguinis* (*Ss*), *S. mitis* (*Sm*), *S. oralis* (*So*), *S. pneumoniae* (*Sp*), *Gemella haemolyans* (*Gh*), *Haemophilus influenzae* (*Hi*), *H. aegyptius* (*Ha*), *Neisseria meningitidis* (*Nm*), and *N. gonorrhoeae* (*Ng*) all cleave within the IgA1 hinge. There are also two IgA1P isozymes that can be found in certain serotypes of *H. influenzae* and *N. meningitidis* that cleave at different locations along the hinge. Protein sequences for IgA1 were taken from UniProt Accession P01876 and IgA2m(1) from NCBI GenBank Gene 3494.

IgAPs were first described in 1973 as an enzyme present in fecal matter capable of cleaving human IgA into two specific fragments.[37] These fragments were identified as the antibody's Fab and Fc domain, which are generated when IgA is cleaved at its hinge region (*Fig. 1.4.1a*).[17,37] Soon after, the IgAP from the oral bacterial pathogen *Streptococcus sanguinis* was purified directly from growth media supernatant and subjected to various protease inhibitors to determine what type of chemical mechanism the protease used.[38] This kind of general biochemical characterization would become routine for subsequently discovered IgAPs,[17,37,38] giving insight into what we now know about the different structural and enzymatic families of the proteases (*Section 2.1.2*). The high substrate specificity characteristic of IgAPs was also initially seen in the *S. sanguinis* protease, where it was shown to be unable to cleave common protease substrates and other Ig classes.[38] Since then, several IgAPs have been identified from bacteria that inhabit the upper respiratory tract,[39] gastrointestinal system,[40] urogenital tract,[41] and oral cavity.[42] These proteases can be categorized into several evolutionally distinct families based on their structure and enzymatic mechanism (*Section 2.1.2*). A cysteine protease secreted from the parasite protozoan *Entamoeba histolytica* has also been shown to cut IgA into fragments containing Fab and Fc domains.[43] Although there is some evidence to suggest that this protease may be a true IgAP, in that it cleaves with high specificity at the IgA hinge, these studies used unpurified extracts for characterization and no definitive gene has yet to be found.[44,45] These initial studies paved the way to the eventual structural and functional characterization of the *Haemophilus influenzae*,[46] *Streptococcus pneumoniae*,[47] and *Gemella haemolyans* IgAPs (*Section 2.1.3* and *2.1.4*),[48] the only IgAPs with experimentally determined structures to date.

By physically decoupling the domains responsible for detecting and responding to antigens, IgAPs are able to reduce the effectiveness of adaptive immune responses in order to facilitate infection (*Section 1.3*).[17] Once IgA has been proteolyzed, free Fab domains can also bind to bacterial epitopes to shield the pathogen from further opsonization and detection.[49] All IgAPs cleave the IgA hinge C-terminal to a proline residue (*Fig. 1.4.1b*), although the specific proline varies between proteases from different species as well as isozymes within the same species (type 1 vs. type 2 IgAPs).[42,50–53] The majority of IgAPs only cleave the IgA1 subclass and are hence often referred to as IgA1 proteases (IgA1Ps).[17] The protease from *Thomasclavelia ramosa* is the only known IgAP that can cleave IgA2, specifically IgA2 allotype A2m(1).[18] It is also able to cleave IgA1 at the primary sequence of the cleavage site shared between the two subclasses (*Fig.*

1.4.1b).[52] As will be described in *Chapter 2*, its unique ability to specifically cleave both IgA subtypes and its distinct modular structure make it an important protease to study to better understand IgAPs as a whole.

1.5 – Immunoglobulin A Proteases as a Therapeutic Target

Many of the bacterial species that produce IgA1Ps (particularly *Haemophilus influenzae* and *Streptococcus pneumoniae*) are heavily associated with lower respiratory tract infections (LRTIs),[39,54,55] one of the leading causes of death for all ages worldwide.[56] Patients who display severe LRTI symptoms are often prescribed antibiotics as a precautionary measure in the event that the infection is of bacterial origin, as bacterial LRTIs have a worse disease prognosis compared to viral LRTIs.[57] As such, the heavy use of antibiotics in LRTI treatment has been criticized to contribute to the global issue of antibiotic resistance.[57,58] With three IgA1P-producing pathogens (*H. influenzae, S. pneumoniae*, and *Neisseria gonorrhoeae*) already identified by the World Health Organization as being the highest concern for antibiotic resistance, there has been a need to find new strategies to combat these pathogens without resorting to traditional antibiotics.[59]

Because of IgA1P's role as an essential virulence factor in the pathogenesis of certain mucosal pathogens,[17] there have been efforts to develop IgA1P inhibitors to treat LRTIs in lieu of general antibiotics. IgA1Ps seem to be virulence factors in the traditional sense,[15] meaning that bacterial strains and species that produce the protease are strongly correlated with being pathogenic while those that do not are non-pathogenic.[60–62] As such, inhibiting IgA1P activity has been postulated to be an alternative way to stop the progression of LRTIs. Despite the identification of many small-molecule inhibitors for the *H. influenzae*, *N. gonorrhoeae*, and *S. pneumoniae* IgA1Ps,[63–68] none of these molecules have been further developed or characterized since their discovery, likely owing to the lack of structural data on any protease-inhibitor complexes and the biochemical difficulties in obtaining these structures (*Section 2.1.3*). On the other hand, the structure of a neutralizing monoclonal antibody in complex with the *S. pneumoniae* IgA1P has been determined.[47,69] This lends credence to the idea of using an inactive enzyme as a *S. pneumoniae* immunogen to vaccinate against strains not covered under the current polysaccharide vaccines.[70]

1.6 – Immunoglobulin A Proteases as a Protein Therapeutic for Immunoglobulin A Nephropathy

The ability of IgAPs to discern between IgA and other structurally similar substrates makes them good protein-therapeutic candidates to treat diseases associated with aberrant IgA.[71,72] This is because, in theory, their high substrate specificity lowers the risk of off-target effects relative to small-molecule drugs.[73] One of these diseases is immunoglobulin A nephropathy (IgAN), a chronic and debilitating autoimmune disease characterized by the buildup of IgA aggregates in the kidney.[74] As issues with hampered kidney filtration can easily impact and propagate to other organ systems, IgAN is often found comorbid with other diseases.[75] Although there have been some prevailing hypotheses to why and how IgAN manifests,[76–78] the precise molecular nature of these IgA aggregates still remains unclear.[79] Current treatment options can therefore only provide temporary relief by slowing down disease progression and alleviating symptoms of inflammation.[75,79]

The potential of IgAPs to be promising protein therapeutic candidates for IgAN has been demonstrated in the kidneys of a humanized IgAN *in vivo* mouse model, in which IgA1 aggregates were introduced into mouse glomeruli and successfully cleared by intravenous injection of IgAPs.[71,80] It has also been shown *ex vivo* with the clearance of IgA aggregates in histological samples from IgAN patients[81] and *in vitro* with human serum as the matrix for aggregate proteolysis.[80] Unfortunately, one of the challenges with using IgAPs as a protein therapeutic is the strong possibility of the enzyme being highly immunogenic, precisely because of their biological role in pathogenesis.[72]

1.7 – Thesis Goals and Approaches

Regardless of whether IgAPs are to be studied for their role in pathogenicity and host-microorganism relationships or as a protein-based therapeutic for IgAN, an in-depth biochemical analysis of IgAP structure and function is needed to lay the foundation for a deeper understanding of these biological and clinical systems. As will be described in *Chapter 2*, the path towards in-depth structural analysis has already been set in motion for the *H. influenzae*, *S. pneumoniae*, and *G. haemolyans* IgA1Ps as their structures have been determined using structural-biology techniques.[46–48] These IgA1Ps, however, only represent two of three known IgAP enzyme families; the third IgAP enzyme family has yet to be structurally or functionally characterized.

This third IgAP family is represented by the protease from *Thomasclavelia ramosa*, a commensal bacterial species that inhabits the human gastrointestinal tract.[82] The *T. ramosa* IgAP has a predicted domain architecture analogous to beads on a string (*Section 2.3.1*), where domains are strung together by protein linkers and act independent of each other. This is unlike the globular folds of the IgA1Ps from *H. influenzae* and *S. pneumoniae* which cannot be substantially reduced in size without abolishing function.[46–48]

The *T. ramosa* IgAP is therefore a key missing piece in the IgAP puzzle. Its distinct protein architecture coupled with the fact that it is produced from a commensal, a non-pathogenic bacterial species makes it a unique system to study amongst all IgAPs. As such, there are many open questions about how the *T. ramosa* IgAP functions and how it relates to IgAPs as a whole. Is the enzyme truly as modular as predicted, meaning that the proteolytic function remains intact when the enzyme is reduced in size, or is there some cooperativity between domains that is necessary for function? What are the structures and functions of the proteolytic and non-proteolytic domains and how do these compare to other IgAPs? And, akin to the role of IgAPs in the virulence of pathogenic bacteria, does the *T. ramosa* IgAP play a role in maintaining a commensal relationship with the gut ecosystem?

My thesis will attempt to address these questions through an in-depth biochemical characterization of the *T. ramosa* IgAP. This will be accomplished through three complimentary and intertwined experimental approaches, each detailed in their own individual chapter:

(1) *Chapter 2 - Structurally Characterizing Domains of the Thomasclavelia ramosa Immunoglobulin A Protease with X-ray Crystallography*. The predicted domain architecture of the *T. ramosa* IgAP suggests that the enzyme contains modular domains. To confirm if these domains are truly modular, I recombinantly expressed and purified these proteins in *E. coli* and experimentally determined their structures with X-ray crystallography. These crystal structures revealed critical information on how the protease functions, potentially why the enzyme is so specific, and what structural elements may be important for binding IgA.

(2) *Chapter 3 - Structurally Characterizing the Thomasclavelia ramosa Immunoglobulin A Protease with Small-Angle X-ray Scattering Techniques*. The domain architecture of the *T. ramosa* IgAP suggests that the protein linkers between each well-folded domain are conformationally flexible. The structures of multidomain *T. ramosa* IgAP constructs must

therefore be investigated using an alternative structural approach as their inherent flexibility makes them generally unsuitable to analysis via X-ray crystallography. Small-angle X-ray scattering was used to determine the solution structures of domains on their own and in the context of the full-length enzyme. These data revealed information about the spatial arrangement of the domains relative to each other and reinforced the distinct modular nature of this family of IgAPs.

(3) *Chapter 4 - Enzymologically Characterizing the Thomasclavelia ramosa Immunoglobulin A Protease*. The crystal and solution structures of the enzymatically active domain suggest that it can function independent of the rest of the enzyme. Given the apparent modular nature of the rest of the protein, is it possible that some of the other domains have non-catalytic functions by aiding in the enzymatic activity of the active domain? As enzymatic comparisons between different protein constructs should only be done under well-defined experimental conditions, a gel-based kinetic assay was developed and validated to obtain Michaelis-Menten parameters for kinetic comparisons. This assay is capable of testing structure-based hypotheses against functional kinetic data to gain insight into what structural elements are required for IgAP activity.

And finally, *Chapter 5* will summarize the main experimental findings of each chapter, propose future experiments, and bring these findings full circle to address some of the questions posed in this introduction.

1.8 – Chapter References

1. McGhee, J. R. & Fujihashi, K. Inside the Mucosal Immune System. *PLoS Biol* **10**, (2012).
2. Mettelman, R. C., Allen, E. K. & Thomas, P. G. Mucosal immune responses to infection and vaccination in the respiratory tract. *Immunity* **55**, 749–780 (2022).
3. Perez-Lopez, A., Behnsen, J., Nuccio, S. & Raffatellu, M. Mucosal immunity to pathogenic intestinal bacteria. *Nat Rev Immunol* **16**, 135–148 (2016).
4. Mestecky, J. & Fultz, P. N. Mucosal Immune System of the Human Genital Tract. *J Infect Dis* **179**, S470-474 (1999).
5. McDermott, A. J. & Huffnagle, G. B. The microbiome and regulation of mucosal immunity. *Immunology* **142**, 24–31 (2014).
6. Dillon, A. & Lo, D. D. M cells: Intelligent engineering of mucosal immune surveillance. *Front Immunol* **10**, (2019).
7. Abraham, S. N. & Miao, Y. The nature of immune responses to urinary tract infections. *Nat Rev Immunol* **15**, 655–663 (2015).
8. Chieng, C. C. Y., Kong, Q., Liou, N. S. Y., Khasriya, R. & Horsley, H. The clinical implications of bacterial pathogenesis and mucosal immunity in chronic urinary tract infection. *Mucosal Immunol* **16**, 61–71 (2023).
9. Linden, S. K., Sutton, P., Karlsson, N. G., Korolik, V. & McGuckin, M. A. Mucins in the mucosal barrier to infection. *Mucosal Immunol* **1**, 183–197 (2008).
10. Bakshani, C. R., Morales-Garcia, A. L., Althaus, M., Wilcox, M. D., Pearson, J. P., Bythell, J. C., & Burgess, J. G. Evolutionary conservation of the antimicrobial function of mucus: a first defence against infection. *NPJ Biofilms Microbiomes* **4**, (2018).
11. Sheng, Y. H. & Hasnain, S. Z. Mucus and Mucins: The Underappreciated Host Defence System. *Front Cell Infect Microbiol* **12**, (2022).
12. Chaplin, D. D. Overview of the Immune Response. *J Allergy Clin Immunol* **125**, S3-23 (2010).
13. Casadevall, A. & Pirofski, L.-A. Host-Pathogen Interactions: Basic Concepts of Microbial Commensalism, Colonization, Infection, and Disease. *Infect Immun* **68**, 6511–6518 (2000).
14. Ribet, D. & Cossart, P. How bacterial pathogens colonize their hosts and invade deeper tissues. *Microbes Infect* **17**, 173–183 (2015).
15. Casadevall, A. & Pirofski, L.-A. Host-Pathogen Interactions: The Attributes of Virulence. *J Infect Dis* **184**, 337–344 (2001).
16. Yoo, E. M. & Morrison, S. L. IgA: An immune glycoprotein. *Clinical Immunology* **116**, 3–10 (2005).

17. Mistry, D. & Stockley, R. A. IgA1 protease. *Int J Biochem Cell Biol* **38**, 1244–1248 (2006).

18. Potempa, J. & Poulsen, K. IgA protease of Clostridium ramosum. in *Handbook of Proteolytic Enzymes* (eds. Barrett, A. J., Rawlings, N. D. & Woessner, J. F.) 759–762 (Academic Press, 2004).

19. Woof, J. M. & Ken, M. A. The function of immunoglobulin A in immunity. *J Pathol* **208**, 270–282 (2006).

20. Sumiyama, K., Saitou, N. & Ueda, S. Adaptive Evolution of the IgA Hinge Region in Primates. *Mol Biol Evol* **19**, 1093–1099 (2002).

21. Chintalacharuvu, K. R., Raines, M. & Morrison, S. L. Divergence of Human α-Chain Constant Region Gene Sequences: A Novel Recombinant α2 Gene. *The Journal of Immunology* **152**, 5299–5304 (1994).

22. Tsuzukida, Y., Wang, C.-C. & Putnam, F. W. Structure of the A2m(1) allotype of human IgA - a recombinant molecule. *Proc Natl Acad Sci U S A* **76**, 1104–1108 (1979).

23. de Sousa-Pereira, P. & Woof, J. M. IgA: Structure, Function, and Developability. *Antibodies* **8**, (2019).

24. Woof, J. M. & Kerr, M. A. IgA function - variations on a theme. *Immunology* **113**, 175–177 (2004).

25. Kumar, N., Arthur, C. P., Ciferri, C. & Matsumoto, M. L. Structure of the secretory immunoglobulin A core. *Science* **367**, 1008–1014 (2020).

26. Lombana, T. N., Rajan, S., Zorn, J. A., Mandikian, D., Chen, E. C., Estevez, A., Yip, V., Bravo, D. D., Phung, W., Farahi, F. *et al.* Production, characterization, and in vivo half-life extension of polymeric IgA molecules in mice. *MAbs* **11**, 1122–1138 (2019).

27. Delacroix, D. L., Divet, C., Rambaudt, J. C. & Vaerman, J. P. IgA subclasses in various secretions and in serum. *Immunology* **47**, 383 (1982).

28. Woof, J. M. & Mestecky, J. Mucosal immunoglobulins. *Immunol Rev* **206**, 64–82 (2005).

29. Brandtzaeg, P. Humoral Immune Response Patterns of Human Mucosae: Induction and Relation to Bacterial Respiratory Tract Infections. *J Infect Dis* **165**, S167-176 (1992).

30. Kutteh, W. H., Princez, S. J., Hammondy, K. R., Kutteh, C. C. & Mesteckyz, J. Variations in immunoglobulins and IgA subclasses of human uterine cervical secretions around the time of ovulation. *Clin Exp Immunol* **104**, 538–542 (1996).

31. Brandtzaeg, P., Kett, K. & Rognum, T. O. Subclass Distribution of IgG- and IgA-Producing Cells in Secretory Tissues and Alterations Related to Gut Diseases. in *Recent Advances in Mucosal Immunology* (eds. Mestecky, J., McGhee, J. R., Bienenstock, J. & Ogra, P. L.) 321–333 (1987).

32. Li, Y., Jin, L. & Chen, T. The Effects of Secretory IgA in the Mucosal Immune System. *Biomed Res Int* **2020**, (2020).

33. Bunker, J. J. & Bendelac, A. IgA Responses to Microbiota. *Immunity* **49**, 211–224 (2018).

34. Mantis, N. J., Rol, N. & Corthésy, B. Secretory IgA's Complex Roles in Immunity and Mucosal Homeostasis in the Gut. *Mucosal Immunol* **4**, 603–611 (2011).

35. Monteiro, R. C. & van de Winkel, J. G. J. IgA Fc receptors. *Annu Rev Immunol* **21**, 177–204 (2003).

36. Lecocq, M., Detry, B., Guisset, A. & Pilette, C. FcαRI-Mediated Inhibition of IL-12 Production and Priming by IFN-γ of Human Monocytes and Dendritic Cells. *J Immunol* **190**, 2362–2371 (2013).

37. Mehta, S. K., Plaut, A. G., Calvanico, N. J. & Tomasi, T. B. Human Immunoglobulin A: Production of an Fc Fragment by an Enteric Microbial Proteolytic Enzyme. *J Immunol* **111**, 1274–1276 (1973).

38. Plaut, A. G., Genco, R. J. & Tomasi, T. B. Isolation of an Enzyme from Streptococcus sanguis Which Specifically Cleaves IgA. *J Immunol* **113**, 289–291 (1974).

39. Male, C. J. Immunoglobulin A1 Protease Production by Haemophilus influenzae and Streptococcus pneumoniae. *Infect Immun* **26**, 254–261 (1979).

40. Kosowska, K., Reinholdt, J., Rasmussen, L. K., Sabat, A., Potempa, J., Kilian, M. & Poulsen, K. The Clostridium ramosum IgA proteinase represents a novel type of metalloendopeptidase. *Journal of Biological Chemistry* **277**, 11987–11994 (2002).

41. Plaut, A. G., Gilbert, J. V, Artenstein, M. S. & Capra, J. D. Neisseria gonorrhoeae and Neisseria meningitidis: Extracellular Enzyme Cleaves Human Immunoglobulin A. *Science* **190**, 1103–1105 (1975).

42. Mortensen, S. B. & Kilian, M. Purification and Characterization of an Immunoglobulin A1 Protease from Bacteroides melaninogenicus. *Infect Immun* **45**, 550–557 (1984).

43. Quezada-Calvillo, R. & López-Revilla, R. IgA protease in Entamoeba histolytica trophozoites. *Adv Exp Med Biol* **216B**, 1283–1288 (1987).

44. Kelsall, B. L. & Ravdin, J. I. Degradation of Human IgA by Entamoeba histolytica. *J Infect Dis* **168**, 1319–1322 (1993).

45. Garcia-Nieto, R. M., Rico-Mata, R., Arias-Negrete, S. & Avila, E. E. Degradation of human secretory IgA1 and IgA2 by Entamoeba histolytica surface-associated proteolytic activity. *Parasitol Int* **57**, 417–423 (2008).

46. Johnson, T. A., Qiu, J., Plaut, A. G. & Holyoak, T. Active-Site Gating Regulates Substrate Selectivity in a Chymotrypsin-Like Serine Protease. The Structure of Haemophilus influenzae Immunoglobulin A1 Protease. *J Mol Biol* **389**, 559–574 (2009).

47. Wang, Z., Rahkola, J., Redzic, J. S., Chi, Y. C., Tran, N., Holyoak, T., Zheng, H., Janoff, E. & Eisenmesser, E. Mechanism and inhibition of Streptococcus pneumoniae IgA1 protease. *Nat Commun* **11**, (2020).

48. Redzic, J. S., Rahkola, J., Tran, N., Holyoak, T., Lee, E., Martín-Galiano, A. J., Meyer, N., Zheng, H. & Eisenmesser, E. A substrate-induced gating mechanism is conserved among Gram-positive IgA1 metalloproteases. *Commun Biol* **5**, (2022).

49. Mansa, B. & Kilian, M. Retained Antigen-Binding Activity of Fab(α) Fragments of Human Monoclonal Immunoglobulin A1 (IgA1) Cleaved by IgA1 Protease. *Infect Immun* **52**, 171–174 (1986).

50. Kilian, M., Thomsen, B., Petersen, T. E. & Bleeg, H. Molecular Biology of Haemophilus influenzae IgA1 Proteases. *Mol Immunol* **20**, 1051–1058 (1983).

51. Spooner, R. K., Russell, W. C. & Thirkell, D. Characterization of the Immunoglobulin A Protease of Ureaplasma urealyticum. *Infect Immun* **60**, 2544–2546 (1992).

52. Fujiyama, Y., Masayoshi, I., Hodohara, K., Hosoda, S. & Kobayashi, K. The Site of Cleavage in Human Alpha Chains of IgA1 and IgA2:A2m(1) Allotype Paraproteins by the Clostridial IgA Protease. *Mol Immunol* **23**, 147–150 (1986).

53. Qiu, J., Brackee, G. P. & Plaut, A. G. Analysis of the Specificity of Bacterial Immunoglobulin A (IgA) Proteases by a Comparative Study of Ape Serum IgAs as Substrates. *Infect Immun* **64**, 933–937 (1996).

54. Troeger, C., Blacker, B., Khalil, I. A., Rao, P. C., Cao, J., Zimsen, S. R. M., Albertson, S. B., Deshpande, A., Farag, T., Abebe, Z. *et al.* Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Infect Dis* **18**, 1191–1210 (2018).

55. Brooks, L. R. K. & Mias, G. I. Streptococcus pneumoniae's Virulence and Host Immunity: Aging, Diagnostics, and Prevention. *Front Immunol* **9**, (2018).

56. Liu, L., Oza, S., Hogan, D., Chu, Y., Perin, J., Zhu, J., Lawn, J. E., Cousens, S., Mathers, C. & Black, R. E. Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. *The Lancet* **388**, 3027–3035 (2016).

57. Örtqvist, Å. Treatment of community-acquired lower respiratory tract infections in adults. *European Respiratory Journal* **20**, S40-53 (2002).

58. World Health Organization. *Global Antimicrobial Resistance Surveillance System (GLASS) Report Early Implementation*.

59. World Health Organization. *Prioritization of Pathogens to Guide Discovery, Research and Development of New Antibiotics for Drug-Resistant Bacterial Infections, Including Tuberculosis*.

60. Mulks, M. H., Kornfeld, S. J., Frangione, B. & Plaut, A. G. Relationship Between the Specificity of IgA Proteases and Serotypes in Haemophilus influenzae. *J Infect Dis* **146**, 266–274 (1982).

61. Kilian, M., Mestecky, J. & Schrohenloher, R. E. Pathogenic Species of the Genus Haemophilus and Streptococcus pneumoniae Produce Immunoglobulin A1 Protease. *Infect Immun* **26**, 143–149 (1979).

62. Mulks, M. H. & Plaut, A. G. IgA Protease Production as a Characteristic Distinguishing Pathogenic from Harmless Neisseriaceae. *N Engl J Med* **299**, 973–976 (1978).

63. Bachovchin, W. W., Plaut, A. G., Flentke, G. R., Lynch, M. & Kettner, C. A. Inhibition of IgA1 Proteinases from Neisseria-Gonorrhoeae and Haemophilus-Influenzae by Peptide Prolyl Boronic Acids. *Journal of Biological Chemistry* **265**, 3738–3743 (1990).

64. Burton, J., Wood, S. G., Lynch, M. & Plaut, A. G. Substrate Analogue Inhibitors of the IgA1 Proteinases from Neisseria Gonorrhoeae. *J Med Chem* **31**, 1647–1651 (1988).

65. Wood, S. G., Lynch, M., Plaut, A. G. & Burton, J. Tetrapeptide Inhibitors of the IgA1 Proteinases from Type I Neisseria gonorrhoeae. *J Med Chem* **32**, 2407–2411 (1989).

66. Garner, A. L., Fullagar, J. L., Day, J. A., Cohen, S. M. & Janda, K. D. Development of a high-throughput screen and its use in the discovery of Streptococcus pneumoniae immunoglobulin A1 protease inhibitors. *J Am Chem Soc* **135**, 10014–10017 (2013).

67. Choudary, S. K., Qiu, J., Plaut, A. G. & Kritzer, J. A. Versatile Substrates and Probes for IgA1 Protease Activity. *ChemBioChem* **14**, 2007–2012 (2013).

68. Shehaj, L., Choudary, S. K., Makwana, K. M., Gallo, M. C., Murphy, T. F. & Kritzer, J. A. Small-Molecule Inhibitors of Haemophilus influenzae IgA1 Protease. *ACS Infect Dis* **5**, 1129–1138 (2019).

69. Chi, Y. C., Rahkola, J. T., Kendrick, A. A., Holliday, M. J., Paukovich, N., Roberts, T. S., Janoff, E. N. & Eisenmesser, E. Z. Streptococcus pneumoniae IgA1 protease: A metalloprotease that can catalyze in a split manner in vitro. *Protein Science* **26**, 600–610 (2017).

70. Mehr, S. & Wood, N. Streptococcus pneumoniae - a review of carriage, infection, serotype replacement and vaccination. *Paediatr Respir Rev* **13**, 258–264 (2012).

71. Lamm, M. E., Emancipator, S. N., Robinson, J. K., Yamashita, M., Fujioka, H., Qiu, J. & Plaut, A. G. Microbial IgA protease removes IgA immune complexes from mouse glomeruli in vivo: Potential therapy for IgA nephropathy. *Am J Pathol* **172**, 31–36 (2008).

72. Xie, L.-S., Huang, J., Qin, W. & Fan, J.-M. Immunoglobulin A1 protease: A new therapeutic candidate for immunoglobulin A nephropathy. *Nephrology* **15**, 584–586 (2010).

73. Kintzing, J. R., Filsinger Interrante, M. V. & Cochran, J. R. Emerging Strategies for Developing Next-Generation Protein Therapeutics for Cancer Treatment. *Trends Pharmacol Sci* **37**, 993–1008 (2016).

74. Julian, B. A., Waldo, F. B., Rifai, A. & Mestecky, J. IgA Nephropathy, the Most Common Glomerulonephritis Worldwide - A Neglected Disease in the United States? *Am J Med* **84**, 129–132 (1988).

75. Lafayette, R. A. & Kelepouris, E. Immunoglobulin A Nephropathy: Advances in Understanding of Pathogenesis and Treatment. *Am J Nephrol* **47**, 43–52 (2018).

76. Novak, J., Barratt, J., Julian, B. A. & Renfrow, M. B. Aberrant Glycosylation of the IgA1 Molecule in IgA Nephropathy. *Semin Nephrol* **38**, 461–476.

77. Matsumoto, Y., Aryal, R. P., Heimburg-Molinaro, J., Park, S. S., Wever, W. J., Lehoux, S., Stavenhagen, K., E van Wijk, J. A., Van Die, I., Chapman, A. B. *et al.* Identification and characterization of circulating immune complexes in IgA nephropathy. *Sci. Adv* **8**, (2022).

78. Coppo, R., Basolo, B., Piccoli, G., Mazzucco, G., Rosa Bulzomi, M., Roccatello, D., de Marchi, M., Carbonara, A. O. & di Belgiojoso, B. IgA1 and IgA2 immune complexes in primary IgA nephropathy and Henoch-Schönlein nephritis. *Clinical Experimental Immunology* **57**, 583–590 (1984).

79. Habas, E., Ali, E., Farfar, K., Errayes, M., Alfitori, J., Habas, E., Ghazouani, H., Akbar, R., Khan, F., Al Dab, A. *et al.* IgA nephropathy pathogenesis and therapy: review & updates. *Medicine* **101**, (2022).

80. Wang, L., Li, X., Shen, H., Mao, N., Wang, H., Cui, L., Cheng, Y. & Fan, J. Bacterial IgA protease-mediated degradation of agIgA1 and agIgA1 immune complexes as a potential therapy for IgA Nephropathy. *Sci Rep* **6**, (2016).

81. Lechner, S. M., Abbad, L., Boedec, E., Papista, C., Le Stang, M. B., Moal, C., Maillard, J., Jamin, A., Bex-Coudrat, J., Wang, Y. *et al.* IgA1 Protease Treatment Reverses Mesangial Deposits and Hematuria in a Model of IgA Nephropathy. *Journal of the American Society of Nephrology* **27**, 2622–2629 (2016).

82. Senda, S., Fujiyama, Y., Ushijima, T., Hodohara, K., Bamba, T., Hosoda, S. & Kobayashi, K. Clostridium ramosum, an IgA Protease-Producing Species and Its Ecology in the Human Intestinal Tract. *Microbiology and Immunity* **29**, 1019–1028 (1985).

Chapter 2 – Structurally Characterizing Domains of the *Thomasclavelia ramosa* Immunoglobulin A Protease with X-ray Crystallography

2.1 – Chapter Introduction

2.1.1 – Importance of X-ray Crystallography in Understanding Immunoglobulin A Proteases

Immunoglobulin A proteases (IgAPs) are a group of enzymes produced from mucosal bacteria that function to cleave and inactivate human immunoglobulin A (IgA).[1] Many IgAPs have been heavily implicated in the virulence of certain mucosal bacterial pathogens and hence have been the focus of therapeutic research in an effort to develop treatments for these pathogens.[2–4] Unfortunately, many of the molecules that were discovered to inhibit IgAPs have yet to be characterized past their initial discovery.[5–10] The main barrier to many of these therapeutic studies is the lack of structural data that can validate and improve on these findings. As will be discussed later in this chapter, the structures of the *Haemophilus influenzae*, *Streptococcus pneumoniae*, and *Gemella haemolyans* IgA1Ps have already been determined[11–13] and therefore have sufficient structural data to form the basis to develop these clinically relevant ventures.

These IgA1Ps, however, only represent two of three known IgAP enzyme families; the third family is represented by the IgAP from *Thomasclavelia ramosa*, which is poorly characterized and has yet to be understood beyond its initial biochemical characterization.[14] Despite these three IgAP families having divergent biochemical and structural features, they have all convergently evolved to cleave the same substrate (*Section 2.1.2*); much of what can be learned from studying one IgAP family can therefore be applied to these other families. The key to deepening our understanding of IgAPs as a whole and progressing applied IgAP research may lie in the structures of this third family of IgAPs.

Recent advances in protein structural modelling have allowed for fairly accurate predictions of previously uncharacterized protein structures.[15] Although the *T. ramosa* IgAP has yet to be studied structurally, the predicted structure of *T. ramosa* IgAP can be modelled using AlphaFold2 (*Section 2.3.1*).[15,16] The ease of generating a semi-reliable predictive protein model has significantly lowered the barrier of entry for initial structural studies. However, now more than ever, these computational models must be accompanied with experimentally validated structural data because important but subtle nuances in protein structure and dynamics cannot be predicted using language models. These computational predictions are also not a rigorous method of testing

hypotheses. The ability to make hypothesis-driven mutations and changes to experimental conditions is also essential to fully understand the structure-function relationships of any enzyme.

It is for these reasons that X-ray crystallography was used to determine the high-resolution structures of various domains of the *T. ramosa* IgAP. The modular nature of the *T. ramosa* enzyme allowed it to be split and produced as individual well-folded domains (*Section 2.3.1* and *2.3.2*), which worked in favor of the structural technique. Not only was X-ray crystallography important for validating the computationally predicted models, but it also provided functional insights into the metal-binding capabilities (*Section 2.3.5*) and functional conformations (*Section 2.3.6*) of the *T. ramosa* IgAP protease domain. X-ray crystallography is a thus powerful technique that can reveal the inner workings of enzymes and is the most suitable technique to use for structural analysis of the *T. ramosa* IgAP due to its modular protein architecture.

It is important to briefly summarize what is known about the biochemistry of the two other IgAP families to give additional context prior to discussing the *T. ramosa* IgAP crystal structures. This is because, as previously stated, the three IgAP families should not be viewed in isolation and characteristics from the other two IgAP families can be used to frame and aid our understanding of the *T. ramosa* IgAP.

2.1.2 – Immunoglobulin A Proteases are a Case of Convergent Evolution

As previously described in *Section 1.4*, IgAPs were initially biochemically categorized based on their chemical mechanism. This was done empirically by incubating crude or purified bacterial extract with a variety of compounds that selectively inhibited certain proteases based on known chemistries and observing the inhibition or lack thereof of IgA cleavage. These included the use of prolyl boronic acids to identify serine proteases,[10] iodoacetamide for cysteine proteases,[17] and ethylenediaminetetraacetic acid (EDTA) for metal-dependent proteases.[18]

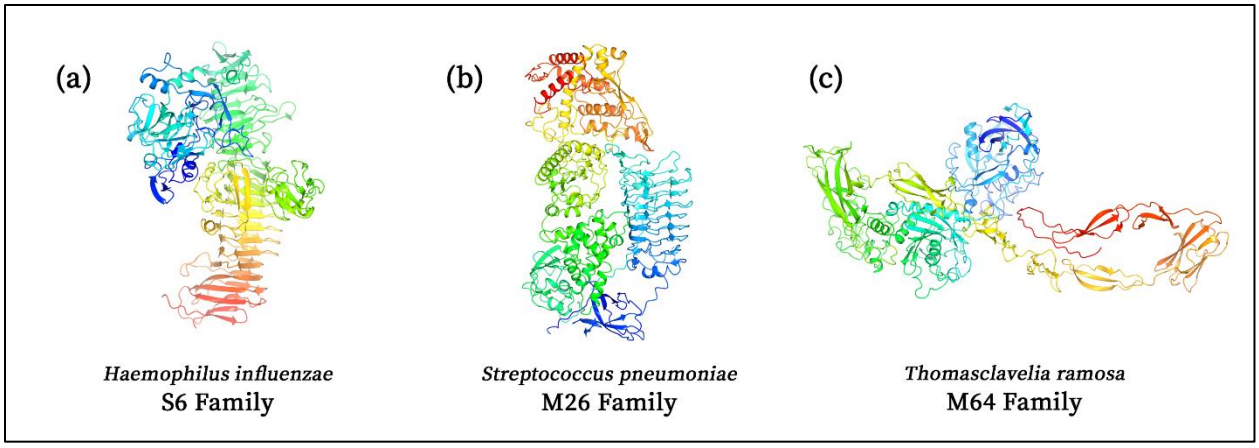


Figure 2.1.2.1 – Immunoglobulin A Protease Families and Their Representative Members. Characterized and annotated IgAPs fall under three enzyme families based on their MEROPS database assignment. One representative member from each family is shown as a ribbon structure and colored as a rainbow from N- (blue) to C-terminus (red). **(a)** The crystal structure (PDB 3H09) of the serine-type *H. influenzae* Type 1 protease depicts common structural characteristics of the S6 family of IgA1Ps. **(b)** The cryoEM structure (PDB 6XJB) of the *S. pneumoniae* metalloprotease depicts common structural characteristics of the M26 family of IgA1Ps. **(c)** The AlphaFold2-predicted model of the full-length *T. ramosa* metalloprotease shows several domains in tandem. The central and largest domain (green) is predicted to be a member of the M64 family of IgAPs. All molecular figures in this thesis were rendered with CCP4MG (version 2.10.11).[112]

With the advent of genome sequencing and, more recently, language-model-driven structural prediction, it's become apparent that these categories represent enzyme families – a set of homologous enzymes that have similar protein sequences, biochemical characteristics, and enzyme structures.[19,20] While IgAPs have been historically classified based on their function, not their structure, it is now clear that the wide structural and biochemical diversity between protease families suggest that they are a result of convergent evolution.[1] This is not surprising given the evolutionary advantage for pathogens to cleave IgA in establishing their pathogenic niche (*Sections 1.2* and *1.3*).[18,21] Despite IgAPs having been historically classified as serine, cysteine, and metal-dependent proteases, only serine and metal-dependent IgAPs have been annotated in bacterial genomes, assigned enzyme families, and structurally characterized;[11–13] cysteine IgAPs have yet to be further described beyond their initial discovery.[17,22]

IgAPs fall under three known enzyme families:

(1) The S6 family of serine IgA1Ps (*Fig. 2.1.2.1a*). This family is represented by the type 1 *Haemophilus influenzae* IgA1P, whose crystal structure revealed that the family has a chymotrypsin-like protease fold.[13] IgA1Ps from *H. aegyptius*, *Neisseria meningitidis*, and *N. gonorrhoeae* are also part of the S6 family.[3,23,24]

(2) The M26 family of metal-dependent IgA1Ps (*Fig. 2.1.2.1b*). This family is represented by the *Streptococcus pneumoniae* IgA1P, whose cryogenic electron microscopic (cryoEM) structure reveals a unique fold that surrounds an otherwise conserved zinc-metalloprotease active site.[11] This cryoEM structure was solved by the Eisenmesser Lab at the University of Colorado, Anschutz Medical Campus with assistance from my previous work in the Holyoak Lab at the University of Waterloo. The IgA1Ps from *Streptococcus sanguinis*, *S. mitis*, *S. oralis*, and *Gemella haemolyans* also fall under the M26 family.[12,25]

And (3), the M64 family of metal-dependent proteases (*Fig. 2.1.2.1c*). This family is represented by the *T. ramosa* IgAP, whose structure has only been computationally modelled and sparsely characterized. The *T. ramosa* IgAP is currently the only protease with an M64 fold that has an identified substrate. As a result, all other putative proteases with M64 folds have thus been bioinformatically identified to be IgAPs without experimental evidence for activity. Due to our lack of understanding of this family, this M64 protease will be the focus of the thesis and will be further described in *Section 2.1.5* and all subsections of *Section 2.3*.

The S6 and M26 families will now be described in more detail to highlight the similarities and differences between these two IgA1P families. These details will provide additional biological and biochemical context for understanding the structure and function of the *T. ramosa* IgAP.

2.1.3 – The S6 Family of Immunoglobulin A1 Proteases

All bacterial species that produce S6 IgA1Ps are Gram-negative.[26,27] This means that the protease must cross both the inner and outer cell membranes as part of its secretion pathway into the extracellular space.[28] Unlike many other secreted Gram-negative proteins, S6 IgA1Ps are able to carry out part of their own secretion independent of host machinery and are hence part of a group of proteins known as Type Va autotransporters.[28]

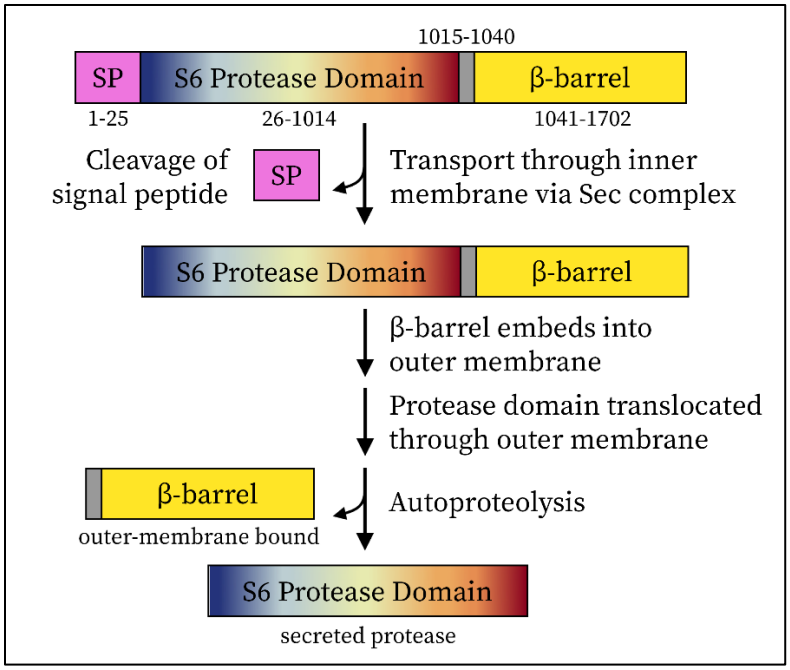


Figure 2.1.3.1 – Secretion Pathway of S6 Immunoglobulin A1 Proteases. The protease from *Haemophilus influenzae* is exemplified here as a representative for S6 IgA1Ps. Residues are numbered based on UniProt accession #P45384. The naïve protein contains a signal peptide (SP) and β-barrel domain that are cleaved off during secretion. The *H. influenzae* IgA1P contains a proline-rich region (grey) where autoproteolysis releases the protease from the outer membrane.

The naïve S6 IgA1P contains additional domains that are required for this secretion pathway (*Fig. 2.1.3.1*).[29,30] Following translation, the N-terminal signal peptide is recognized by the inner-membrane-bound Sec secretion complex and the protein is translocated across the inner membrane into the periplasm.[13,31] The signal peptide is removed during this process.[31] Similar to other serine proteases like trypsin and chymotrypsin, the mature hydrophobic N-terminus buries itself into the core of its chymotrypsin-like fold after signal peptide cleavage, and is essential for stabilizing the S6 protease fold.[13] When the enzyme is recombinantly expressed in the cytosol, the initiator methionine sterically interferes with the hydrophobic core and results in the formation of inclusion bodies.[32] These challenges have made it difficult to study their biochemistry due to the lack of a soluble and well-folded enzyme source that yields a large amount of protein.[32] A suitable recombinant source has only been recently developed through my previous work on the expression and purification of the *H. influenzae* IgA1P from inclusion bodies in *Escherichia coli*.[32]

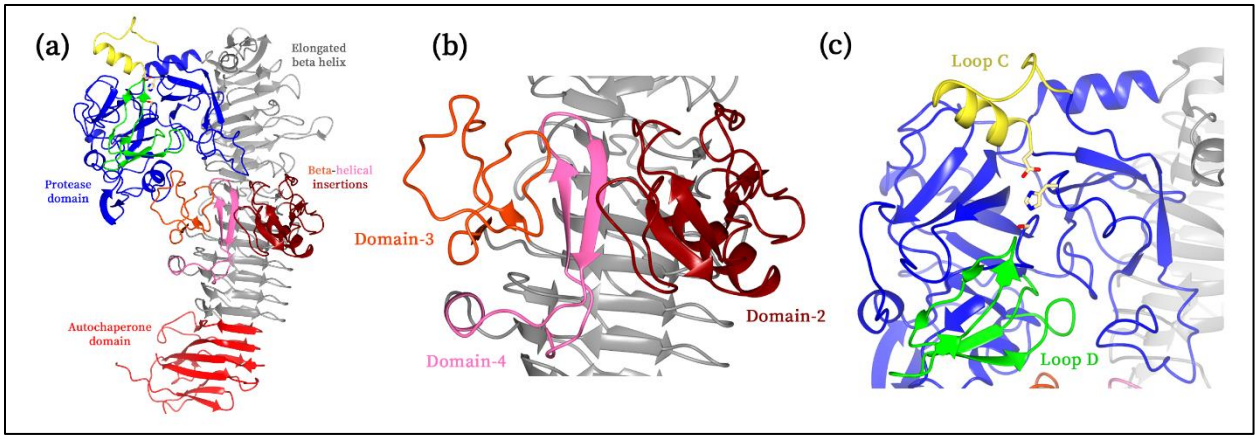


Figure 2.1.3.2 – Annotated Tertiary Structure of the Type 1 *Haemophilus influenzae* Immunoglobulin A1 Protease. (a) The *H. influenzae* IgA1P has features typical of Type Va autotransporter proteins. These include a C-terminal autochaperone domain (red) and a characteristic elongated beta helix (grey). The protease domain (blue) is appended on the N-terminus of the beta helix. **(b)** Three beta-helical insertions (Domain-2, maroon; Domain-3, orange; Domain-4, pink) are hypothesized to recognize the IgA1 Fc as part of its substrate-recognition mechanism. **(c)** Once the Fc is bound, two active-site loops (Loop C, yellow; Loop D, green) are hypothesized to open up the active site to allow the IgA1 hinge to enter. The catalytic triad is shown as cylinders.

The naïve S6 IgA1P also contains an extended C-terminal beta-barrel domain which is characteristic of members of the Type Va autotransporter group (*Fig. 2.1.3.1*).[13,28] Following translocation into the periplasm, this C-terminal domain embeds itself into the bacterial outer membrane and shuttles the rest of the protein through the membrane.[29,30] The molecular mechanism of how this occurs and whether these beta-barrel domains form larger oligomeric complexes as part of its secretion mechanism still remains unclear.[28,33] The protein is then able to fold into its mature and active conformation on the surface of the bacterium.[34] Its ability to fold spontaneously is attributed to the C-terminal region of the mature protease (*Fig. 2.1.3.2a*), which consists of a large beta helix that folds with the help of an autochaperone domain at the C-terminus.[13,34] The mature enzyme then remains bound to the membrane until it is shed through proteolysis, either by exogenous proteases or by autoproteolysis.[35,36] The autoproteolysis sites for S6 IgA1Ps are similar to the cleavage site of the IgA hinge in that they are both proline-rich.[37] As such, they have been successfully used as the basis for the design of certain IgA1P substrates and inhibitors.[8,9]

The overall architecture of the type 1 *H. influenzae* IgA1P (*Fig. 2.1.3.2*) is consistent with other proteins that follow the Type Va secretion pathway.[13,28] Due to the need for these proteins to fold spontaneously during secretion, the functional portions of the protein are designed to be built

around a central beta helix, as the helix acts as a structural and folding scaffold.[34,38] The linkers between the beta strands of the helix are often substituted with extended loop or domain insertions that can serve additional functional purposes.[28,38] The *H. influenzae* IgA1P has three large beta-strand insertions (Domains-2, -3, and -4; *Fig. 2.1.3.2b*) that have been hypothesized to form distal interactions with IgA1 as their conformation in the crystal structure can be rigid-body modelled to interact with the IgA1 Fc.[13] In addition to these insertions, two loops (loops C and D; *Fig. 2.1.3.2c*) adjacent to the active site are elongated relative to other structurally homologous serine proteases.[13] These loops have been hypothesized to sterically occlude the entrance to the active site when no IgA1 is bound, effectively acting as a gate that prevents non-specific substrates from entering.[13] This active-site gate would then be pinned back to open up the active site when distal interactions occur between IgA1 and the other beta-strand insertions.[13] Structural and sequence differences in these distal domains and active-site-gating loops may be the keys to understanding how the *H. influenzae* IgA1P functionally differs from other S6 proteases and proteases from the Type Va autotransporter group.[13] As will be seen for the M26 protease family, active-site gating loops seem to be a mechanism that has convergently evolved for IgAPs to achieve high substrate specificity.

2.1.4 – The M26 Family of Immunoglobulin A1 Proteases

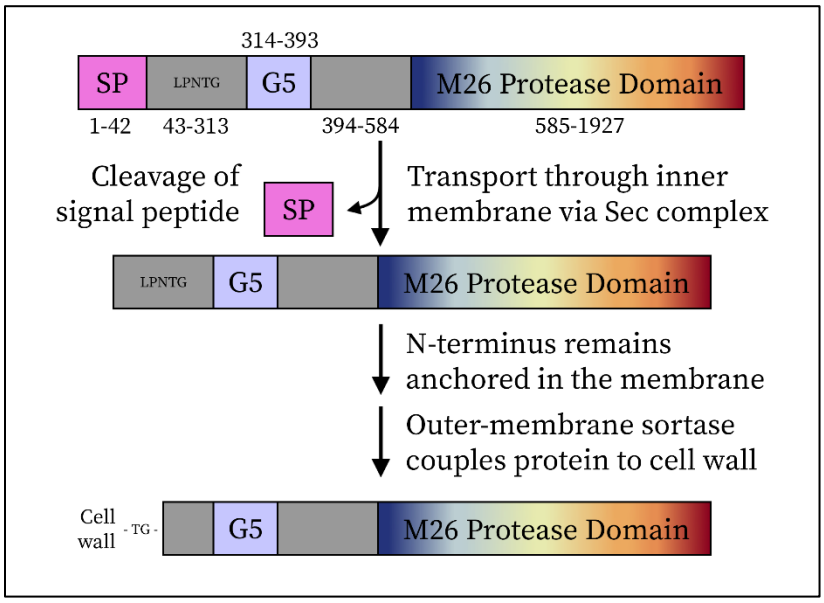


Figure 2.1.4.1 – Putative Secretion Pathway of M26 Immunoglobulin A1 Proteases. The protease from *Streptococcus pneumoniae* is exemplified as a representative for M26 IgA1Ps. Residues are numbered based on UniProt accession #Q54875. The naïve protein contains a signal peptide (SP) and LPNTG cell-anchoring motif that are involved in secretion. The mature protein includes an N-terminal G5 domain that functions separately to the C-terminal M26 IgA1P domain and is tethered together by linker regions (grey).

In contrast to S6 IgA1Ps, all M26 IgA1Ps are produced from Gram-positive bacteria.[39] The differences in membrane anatomy between Gram-positive and Gram-negative bacteria mean that the secretion mechanism and, by consequence, the protein sequences that are compatible with this mechanism must fundamentally be different.[28,40] Like S6 IgA1Ps, the *S. pneumoniae* M26 IgA1P contains additional domains not required for proteolytic activity (*Fig. 2.1.4.1*).[41] These domains, however, are not used in its secretion pathway and have separate functions from the IgA1P.[41] A G5 domain found N-terminal to the *S. pneumoniae* IgA1P is hypothesized to bind to certain small molecules to promote cell adhesion.[41–43] The *Gemella haemolyans* M26 IgA1P contains a trypsin-like and G5 domain N-terminal to its IgA1P.[12] The protein architecture of M26 IgA1Ps is thus reminiscent of beads on a string – with individually folded and functional domains joined together by protein linkers.

Following translation, the N-terminal signal peptide is processed in a similar fashion to the Gram-negative pathway and the protein is translocated across the bacterial membrane into the extracellular space (*Fig. 2.1.4.1*).[31,40] The translocated protein is then hypothesized to be covalently anchored onto the cell wall by bacterial sortases, through the recognition of several conserved anchoring elements in the primary sequence (one of which is often referred to as the cell-anchoring LPXTG motif).[1,40] Unlike most Gram-positive cell-wall-anchored proteins, the anchoring motif for the *S. pneumoniae* IgA1P is on the N-terminal rather than C-terminal end, a

feature which is shared with the structurally homologous *S. pneumoniae* zinc metalloproteases ZmpB, ZmpC, and ZmpD (*Fig. 2.1.4.1*).[25,44,45] The protease then remains tethered to the cell wall until it is released via proteolysis at its N-terminal proline-rich region.[25,46] Bacterial species that produce M26 IgA1Ps keep the majority of the enzyme tethered to the cell wall rather than fully secreting them into the extracellular space.[47] This is in stark contrast with S6 IgA1P-producing bacterial species, which have their proteases almost completely shed from the membrane.[47]

It is pertinent to note that because of the location of the cell-anchoring motif, sortases would only be able to couple the small region N-terminal to the cell-anchoring motif to the cell wall.[40] The G5 and IgA1P domains would hence be unsuccessfully secreted, suggesting that there may be alternative mechanisms for the secretion of some *S. pneumoniae* zinc metalloproteases.[48] As such, the mechanism of secretion for the M26 family of IgA1Ps still remains unclear.

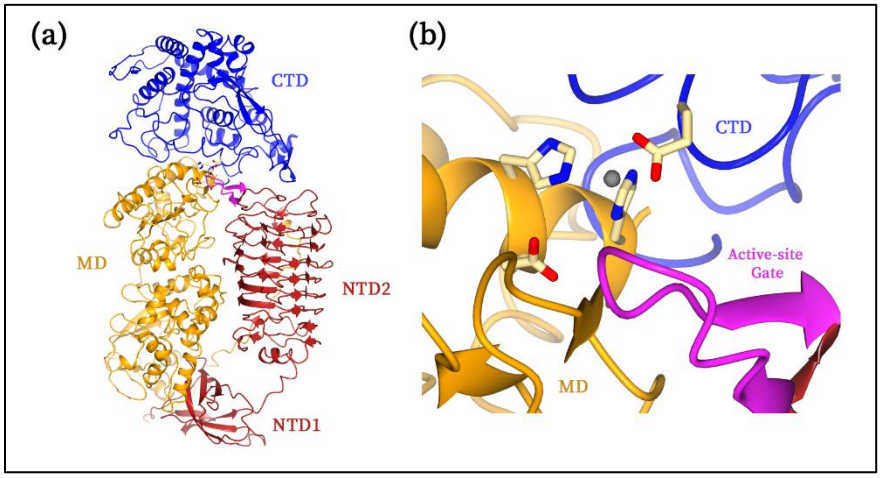


Figure 2.1.4.2 – Annotated Tertiary Structure of the *Streptococcus pneumoniae* Immunoglobulin A1 Protease. (a) The *S. pneumoniae* IgA1P can be split into three pseudo-domains. These include an N-terminal domain (NTD), middle domain (MD), and C-terminal domain (CTD). The NTD consists of a small globular region (NTD1) and a beta helix (NTD2), which is similar to the elongated beta helix in the *Haemophilus influenzae* IgA1P. **(b)** The NTD+MD can be combined with the CTD to reconstitute activity as each construct contains a portion of the active site. The MD contains the HEXXH motif which is involved in zinc binding (grey sphere) while the CTD contains the C-terminal glutamic acid used as a base to deprotonate and activate the nucleophilic water molecule. A beta-hairpin structure (magenta) is hypothesized to be the loop that gates the active site to facilitate substrate specificity.

The globular structure of M26 IgA1Ps can be divided into three pseudo-domains (*Fig. 2.1.4.2a*): an N-terminal cap and beta helical structure (N-terminal domain; NTD), a central middle domain (MD), and a C-terminal structure (C-terminal domain; CTD).[11,46] Although this pseudo-domain architecture is conserved amongst other metalloprotease paralogues from *Streptococcus spp.* (ZmpB, ZmpC, and ZmpD), only those more closely related to the ZmpA clade have IgA1P

activity.[49] The M26 N-terminal beta helix is structurally similar to the C-terminal beta helix in the S6 IgA1Ps and thus is thought to play a role in the enzyme folding and stabilization.[50] The *S. pneumoniae* IgA1P NTD likely is necessary for the proper folding of the MD as the NTD can be produced on its own but the MD can only be produced in the context of the NTD (data not shown; E. Z. Eisenmesser, personal communication). Initial structural characterization of the *S. pneumoniae* IgA1P revealed that the protease could be split into two parts: the NTD with MD (NTD+MD) and CTD alone.[46] These parts are inactive by themselves but regain activity when the separately purified domains were combined *in vitro*.[46] This was particularly interesting as each part contains a portion of the active site.[46]

M26 IgA1Ps share a conserved active site with many other zinc metallo-endopeptidases, with their active-site components being closely related to that of the zinc metalloprotease thermolysin.[46] Thermolysin-like proteases use a metal-coordinated water molecule as one of the substrates for catalysis.[51] This water molecule is polarized by a tightly bound zinc ion to form an activated nucleophilic Zn^{2+}-hydroxide complex.[51] The active site contains a metal-binding HEXXH motif and an additional C-terminal glutamic acid residue to carry out catalysis (*Fig. 2.1.4.2b*).[51] The side chains of the two histidines of the active-site **H**EXX**H** motif and the C-terminal glutamic acid are ligands that coordinate and stabilize the Zn^{2+} ion, while the glutamic acid in the H**E**XXH motif is the catalytic base that deprotonates the polarized water molecule.[51] The active site is found in the cleft between the MD and CTD, whereby the MD contributes the HEXXH motif and the CTD contains the C-terminal glutamic acid.[11,46]

Like S6 IgA1P, M26 IgA1Ps have also convergently evolved structural mechanisms to allow for high substrate specificity. The C-terminal end of the NTD beta helix contains an elongated beta-hairpin structure that occludes the active site when no IgA1 is bound (*Fig. 2.1.4.2b*).[11,12] When the IgA1 Fab interacts with the NTD, a large conformational change compresses the beta helix against the N-terminal cap.[11,12] This moves the beta-hairpin structure out of the way and allows the IgA1 hinge to enter the active site.[11,12] The cryoEM structure of the *Gemella haemolyans* IgA1P depicts a similar beta-hairpin structure and conformational change, suggesting that this active-site gating mechanism is conserved amongst the larger M26 IgA1P family.[12] This gating mechanism is hypothesized to be a consequence of how M26 proteases in the

ZmpA clade have evolved IgA1P activity while ZmpB, ZmpC, and ZmpD proteases have alternate functions.[11,12,49]

Despite the distinct structural and mechanistic differences between the S6 and M26 IgA1P families, IgA1Ps seem to have evolved similar functional characteristics: they remain bound to the cell following secretion, contain proline-rich autoproteolysis sites that allow for secretion into the extracellular space, and have active-site gating mechanisms that involve distal substrate interactions to retain high specificity towards IgA1. As will be discussed later, some of these features are analogously seen in the *T. ramosa* IgAP as they seem to be important for biological function.

2.1.5 – Introduction to Thomasclavelia ramosa and Its Immunoglobulin A Protease

Thomasclavelia ramosa (formerly named *Erysipelatoclostridium ramosum* and *Clostridium ramosum*) is a Gram-positive bacterial species that forms part of the gut microbiota.[52] It was initially identified from pulmonary gangrene and appendicitis patients and was historically called *Bacillus ramosus* and *Nocardia ramosa*.[52] This nomenclature later changed to its more well-known name, *C. ramosum*, when it was discovered that they could form spores.[52] More recently, *C. ramosum* was reclassified as *T. ramosa* when the *Clostridium* genus was revisited with more stringent taxonomic analysis.[53]

T. ramosa makes up a part of the normal human gut microbiome, as it is commonly detected and isolated from the feces of a large percentage of adults.[54] In rare cases, *T. ramosa* has been found to cause bacteremia and other tissue infections and hence is sometimes considered an opportunistic pathogen.[55–59] Despite this, *T. ramosa* is generally considered a commensal gut microorganism as these infrequent cases seem to only arise in immunocompromised patients and children with immune-system complications.[55,58]

After the discovery and initial characterization of IgAPs from common pathogens that infect the upper respiratory tract and oral cavity,[60,61] the search for more IgAP-producing bacteria continued through the investigation of other mucosal surfaces. It was hypothesized that IgAPs may act as virulence factors that aid pathogens in causing certain intestinal diseases as IgAPs were initially identified through protease activity in feces.[14,62] A strain of *T. ramosa* isolated from a

patient with ulcerative colitis was first discovered to harbor IgAP activity.[14] This IgAP was able to cleave both IgA1 and IgA2 allotype A2m(1), not allotype A2m(2), at a specific site in the hinge region shared between both subclasses (*Fig. 1.4.1b*) and the protease was determined to be metal-dependent based on the fact that it was inhibited by EDTA.[14] The ability of the *T. ramosa* IgAP to cleave both IgA1 and IgA2m(1) should not come as too much of a surprise as the function of the enzyme and its biological role likely evolved in the context of the intestine which is rich in both subclasses of IgA (*Section 1.3*).[14,63] Along the same vein, the IgA1-specific proteases from pathogens associated with lower-respiratory tract infections (*Section 1.5*) evolved in the context of the IgA1-rich respiratory tract (*Section 1.3*).[60,64]

Given what had been known about IgAPs and their role in virulence, it came as a surprise to discover that there didn't seem to be any correlation between the production of the *T. ramosa* IgAP and inflammatory bowel disease (IBD).[54] It was shown that patients with no IBD had IgAP-producing *T. ramosa* while none of the clinical isolates of *T. ramosa* from patients with IBD showed IgAP activity.[54] Furthermore, it was shown that that human sera and milk contain neutralizing antibodies against many IgA1Ps from pathogenic species but not against the *T. ramosa* IgAP.[65] This poses the question of what the biological role of *T. ramosa* IgAP is and if it is even related to virulence given the species' commensal nature.

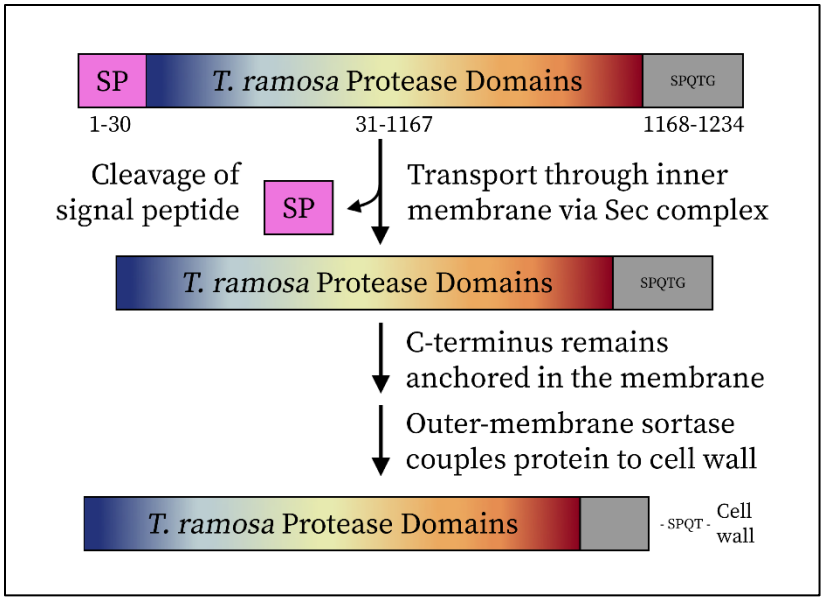


Figure 2.1.5.1 – Secretion Pathway of the *Thomasclavelia ramosa* Immunoglobulin A Protease. The protease from *T. ramosa* is exemplified as a representative for M64 proteases as it is the only member with a known substrate. Residues are numbered based on UniProt accession #Q9AES2. The naïve protein contains a signal peptide (SP) and SPQTG cell-anchoring motif (grey) that are involved in secretion. The M64 protease is one of many domains simplistically grouped together and colored in rainbow.

Since the initial discovery of the *T. ramosa* IgAP, the gene encoding the protease has been identified and the enzyme has been recombinantly expressed in *Escherichia coli*.[66,67] The open-reading frame encodes a ~130 kDa protein (*Fig. 2.1.5.1*) which, like S6 and M26 proteases (*Section 2.1.3* and *2.1.4*), contains additional sequences necessary for its secretion.[66,68] As *T. ramosa* is Gram-positive, the protease follows a very similar secretion path to that hypothesized for the M26 IgA1Ps:[68] the protein is first shuttled through the cell membrane via the Sec pathway, during which its N-terminal signal peptide is cleaved off, and the protease is then anchored to the peptidoglycan cell wall by a sortase enzyme where it remains until it is released via proteolysis or when the peptidoglycan is shed.[68] However unlike M26 IgA1Ps, the protease is anchored to the cell wall at its C-terminus[68] and is predominantly secreted into the media rather than being retained on the cell surface.[66]

2.1.6 – Chapter Objectives

Many questions remain regarding the function of the IgAP in *T. ramosa* and if it plays a role in maintaining a commensal relationship with the gut ecosystem. This is opposed to the S6 and M26 IgA1Ps which are strongly correlated with bacterial pathogenicity.[1] The *T. ramosa* IgAP is also unique amongst these other proteases in that it is predicted to have a modular domain architecture. This is unlike the globular pseudo-domain architecture of the S6 and M26 IgA1Ps, which cannot be reduced in size without abolishing activity.

To confirm if the *T. ramosa* IgAP protein architecture is truly modular, the full-length enzyme was divided into putative domains which were expressed and purified outside of the context of the other domains (*Section 2.3.1*). These proteins were subject to structural analysis via X-ray crystallography (this chapter), small-angle X-ray scattering (*Chapter 3*), and kinetic analysis (*Chapter 4*) to observe if their biophysical and biochemical behaviors changed when produced independent of the full-length enzyme.

The main goal of this chapter is to determine the high-resolution crystal structure of the *T. ramosa* IgAP protease domain. This structure will be used to reveal features in and adjacent to the active site that are potentially important for proteolysis, metal binding, and substrate interaction. These hypotheses will form the basis for a working model of how and why the *T. ramosa* IgAP has high substrate specificity for IgA and will go hand-in-hand with the working models proposed

for the S6 and M26 IgAPs to better understand IgAP substrate specificity as a convergently evolved phenomenon. Finally, the working model will be discussed in the context of the M64 protease family as a warning against using bioinformatic classifications as *a priori* knowledge to make biochemical conclusions.

2.2 – Materials and Methods

2.2.1 – Expression Vector Cloning

The *T. ramosa* IgAP sequence detailed in UniProt (accession #Q9AES2) and GenBank (accession #AY028440) was used and modified as the open reading frame for all protein expression vectors. Protein expression constructs for the N-terminal domain (NTD; residues 31-313), middle domain (MD; 329-807), MD with the first C-terminal domain (MD+CTD1; 329-876), second CTD (CTD2; 880-1008), and third CTD (CTD3; 1011-1096) were made. The fourth CTD (CTD4) was only made in the context of constructs with the longer C-terminal tail, as will be described below. A mutant of the active-site base (H**E**XXH; E540A) was also made in the background of MD+CTD1 to investigate structural changes via X-ray crystallography. All protein inserts were codon optimized for *Escherichia coli*, synthesized *de novo*, and cloned into their respective vectors with services provided by GenScript (Piscataway, NJ, USA).

Additional protein constructs were made to assess if the domains not directly involved in proteolysis would contribute to activity and substrate recognition (*Chapter 4*). These constructs included the full-length (FL) secreted enzyme (31-1195) and NTD with MD (NTD+MD; 31-807). A final set of constructs were designed to investigate the overall free conformation of the FL enzyme (*Chapter 3*), which included the constructs MD with the entire C-terminal tail (MD+CTD1-4; 329-1195) and just the C-terminal tail (CTD1-4; 806-1170).

To get around proteolytic issues in purification,[67] the FL and MD+CTD1-4 inserts were cloned into a dual-tagged vector (modified pJ401; DNA2.0; Menlo Park, CA, USA). These inserts were cloned into the pJ401 vector C-terminal to a maltose-binding protein (MBP) tag and N-terminal to a hexahistidine tag (His-tag) using dual *BamHI* cut sites (GenScript). The MBP tag was removed via treatment with thrombin. The rest of the constructs were subcloned from the FL pJ401 construct into a SUMOstar vector (LifeSensors Inc.; Malvern, PA, USA) C-terminal to a

small ubiquitin-related modifier (SUMO) tag using *BsaI* and *XhoI* cut sites (GenScript). Full vector and primary amino acid sequences for all constructs are listed in *Appendix A*.

2.2.2 – Protein Expression and Purification

Bacterial growth media components were purchased from Bioshop Canada (Burlington, ON, Canada). Chemicals used for protein purification were purchased from ChemImpex Inc. (Wood Dale, IL, USA) with the exception of IgePal-CA630, which was purchased from Thermo Fisher Scientific (Waltham, MA, USA).

Heterologous expression for all constructs was done in *E. coli* BL21(DE3) cells. A 16-to-18-hour LB overnight culture grown at 37 ^{o}C was inoculated into ZYP-5052 autoinduction media[69] at a ratio of 50 mL overnight culture to 1 L final media volume with a minimum headspace-to-media ratio of 4:1. All media were supplemented with 50 μg/mL kanamycin. Cells were grown at 20 ^{o}C and shaken at 150 rpm for 20-24 hours before being harvested via centrifugation at 6,000xg and stored as cell pellets at -80 ^{o}C.

All pJ401 constructs were purified with an identical procedure. This purification protocol was similar but distinct to the protocol used for all SUMOstar constructs. All purification steps were carried out at 4 ^{o}C, unless otherwise specified, and pH values were set at room temperature. The initial purification steps for pJ401 and SUMOstar constructs were the same: cell pellets were thawed in Buffer A (25 mM HEPES, pH 7.5, 0.5 M NaCl, 10 mM imidazole, 1 mM TCEP), passed twice through a FRENCH Pressure Cell (Thermo Fisher Scientific; Waltham, MA, USA) at 1,100 psi for cell lysis, and cell debris was removed via high-speed centrifugation at 17,000xg. The clarified cell lysate was then incubated with NiNTA resin (Qiagen; Hilden, Germany), equilibrated in Buffer A, for one hour. The resin was first washed with 25 column volumes (CVs) of Buffer B (25 mM HEPES, pH 7.5, 0.1% (v/v) IgePal CA-630, 10 mM imidazole, 1 mM TCEP) to remove non-specific hydrophobically bound contaminants followed by 40 CVs of Buffer A. The protein was eluted with Buffer C (25 mM HEPES, pH 7.5, 0.5 M NaCl, 300 mM imidazole, 1 mM TCEP).

For constructs in pJ401 (FL and MD+CTD1-4), the protein was then incubated with amylose resin (NEB; Ipswich, MA, USA), equilibrated in Buffer D (25 mM HEPES, pH 7.5, 0.5 M NaCl, 1 mM TCEP), for one hour. The amylose resin was washed with 20 CVs of Buffer D

before being eluted with Buffer E (25 mM HEPES, pH 7.5, 0.5 M NaCl, 20% (w/v) glucose, 1 mM TCEP). The elution was concentrated to ~3 mL with a 10 kDa molecular-weight-cut-off (MWCO) stirred-cell and centrifugal concentrator (Millipore Sigma; Burlington, MA, USA). Three rounds of 15 mL Buffer D were added to the centrifugal concentrator to buffer exchange the protein out of glucose via centrifugal concentration. The protein was then digested with ~2 mg bovine thrombin (BioPharm; Bluffdale, UT, USA) overnight at room temperature before being treated with 1 mM phenylmethylsulfonyl fluoride (PMSF; 100 mM stock in 100% ethanol) and incubated with two consecutive rounds of amylose resin for an hour each, equilibrated in Buffer D, to remove residual uncut protein and MBP. The final flowthrough was concentrated to less than 1 mL and loaded onto a pre-packed Superdex 200 Increase 10/300 GL column (Cytiva; Marlborough, MA, USA), equilibrated in Buffer D, run at 0.5 mL/min. Fractions not in the void volume peak were concentrated to less than 1 mL and injected onto a pre-packed HiLoad Superdex 75 pg 16/600 (Cytiva), equilibrated in Storage/Crystallization Buffer (25 mM HEPES, pH 7.5 and 1 mM TCEP), and run at 0.5 mL/min.

For constructs in SUMOstar (with the exception of CTD2, see below), the protein eluted off of the first NiNTA resin was digested with SUMO protease overnight (expressed and purified in-house from Addgene plasmid pCDB302).[70] The cleaved protein was dialyzed against Buffer D for a minimum of 6 hours followed by another round of dialysis overnight to remove residual imidazole. The protein was then incubated with a second round of NiNTA resin, equilibrated in Buffer D, to remove the cleaved SUMO tag and any remaining uncleaved fusion. NiNTA flowthrough was concentrated to less than 1 mL and loaded onto a pre-packed HiLoad Superdex 75 pg 16/600, pre-equilibrated in Storage Buffer, run at 0.5 mL/min.

As the CTD2 NiNTA elution had a green hue, $ZnCl_2$ was added to the elution to a final concentration of 1 mM. After an hour of incubation, the solution became colorless. To prevent residual $ZnCl_2$ from interfering with the second NiNTA step, the elution was concentrated to ~3 mL, loaded onto a manually packed P-6DG desalting column (BioRad; Hercules, CA, USA) equilibrated in Buffer D, before being incubated with a second round of NiNTA resin and continuing on from the protocol for SUMOstar above.

The final purity of all proteins was qualitatively analysed using SDS-PAGE. Pure fractions off of the final size-exclusion column were concentrated, frozen in 20-30 μL pellets via direct

immersion in liquid nitrogen, and stored at -80 °C. Protein concentration was measured using 1% mass extinction coefficients theoretically determined from their primary sequence using the ProtParam program in Expasy (*Appendix A*).[71]

2.2.3 – Protein Crystallization

Crystallization screens were purchased from Molecular Dimensions (Holland, OH, USA). Chemicals used for protein crystallization were purchased from ChemImpex Inc. (Wood Dale, IL, USA). Crystal mounts, loops, and a UniPuck system were purchased from MiTeGen (Ithaca, NY, USA).

High-throughput crystallization screening for each of the successfully purified individual domains (NTD, MD, MD+CTD1, and CTD2) were carried out with commercially available screens in small-volume sitting-drop trays using a Crystal Gryphon LCP robot (Art Robbins Instruments; Sunnyvale, CA, USA). Drops consisting of 0.2 μL protein sample were mixed with 0.2 μL reservoir solution and left to equilibrate at room temperature for a few weeks. The reservoir volume was 30 μL.

The protein concentration used in crystallization screening was determined using the Pre-Crystallization Test (Hampton Research; Aliso Viejo, CA, USA). Briefly, 2 μL protein sample at various protein concentrations were mixed with 2 μL of two different mother liquor solutions and left to equilibrate at room temperature overnight in 24-well hanging-drop plates. The two solutions were 0.1 M Tris, pH 8.5 with 2.0 M ammonium sulfate and 0.1 M Tris, pH 8.5 with 0.2 M $MgCl_2$ and 30% (w/v) PEG 4,000. A suitable protein concentration for screening was determined based on the level of precipitation across these two solutions, with an optimal result of no precipitation in one condition and light precipitation in the other.

Initial hits grew after one to two weeks in 0.1 M BICINE, pH 9.3 and 25% (v/v) PEG Smear Medium (BCS B6)[72] for NTD and 0.1 M Bis-Tris, pH 6.5 and 25% (w/v) PEG 3350 (MCSG1 C12)[73] for MD+CTD1. As these crystals were not suitable for single-crystal diffraction, even when reproduced and scaled-up in 24-well hanging-drop plates, the conditions were further optimized by an additive screen (Hampton Research) at a ratio of 2 μL protein to 1.8 μL mother liquor and 0.2 μL non-volatile additive. Blocky crystals for the NTD resulted from a mix of 2.0

μL protein sample (7.5 mg/mL), 1.8 μL reservoir solution (0.1 M BICINE, pH 9.3 and 20% (v/v) PEG Smear Medium), and 0.2 μL 0.1 M $NiCl_2$. Clusters of hollow rectangular crystals for the MD+CTD1 resulted from a mix of 2.0 μL protein sample (8.0 mg/mL), 1.8 μL reservoir solution (0.1 M Bis-Tris, pH 6.5 and 12% (v/v) PEG 3350), and 0.2 μL 0.1 M praseodymium (III) acetate. These crystal clusters were manually manipulated to acquire single crystals suitable for diffraction. All crystals were cryoprotected in their respective reservoir solutions supplemented with 20% (v/v) glycerol. Due to the fragile nature of the MD+CTD1 crystals, they were cryoprotected stepwise with increasing concentrations of glycerol, from 5% to 20% (v/v) in steps of 5%. The MD+CTD1 E540A mutant was crystallized and cryoprotected identically to the wild type MD+CTD1 crystals.

2.2.4 – Metal-Chelation and Zinc-Supplementation Crystallographic Experiments

To obtain crystal structures of the enzyme bound to only zinc, Chelex 100 resin (BioRad) was directly added to protein solutions containing MD+CTD1 enzyme and left to incubate overnight at room temperature. To prepare the chelating resin, fresh Chelex 100 resin was added to a gravity column, washed with 50 CVs of 18.2 MΩcm water, 10 CVs of 0.5 M sodium phosphate buffer, pH 6.3, 50 CVs of 18.2 MΩcm water, then added directly to all solutions used for crystallization and cryoprotection, with the exception of the crystallization additives (praseodymium (III) acetate and $ZnCl_2$). All crystallization buffers and precipitant solutions were made with 18.2 MΩcm water. Chelex 100 resin that was added directly to any protein solution was further equilibrated in 50 CVs of Crystallization Buffer prior to use.

Crystal growth and manipulation were done as previously described with the exception of adding a small amount of Chelex 100 resin to the drop during crystallization. To supplement the crystals with zinc, metal-chelated single crystals were soaked in large 100 μL drops containing the mother liquor supplemented with 1 mM $ZnCl_2$ (Hampton Research) overnight at room temperature. Crystallization additive was not added to the soak solutions. Zinc-soaked crystals were then cryoprotected stepwise as described above with the addition of 1 mM $ZnCl_2$ to all solutions.

2.2.5 – Crystallographic Data Collection and Processing

Diffraction data for the NTD crystals were collected at the Cornell High-Energy Synchrotron Source (CHESS) ID7B2 beamline on a Detectris Pilatus3 S 6M with an exposure of 0.5°/0.25 sec/frame. Diffraction data for all MD+CTD1 crystals were collected at the Canadian Light Source BM beamline on a Detectris Pilatus3 S 6M. To confirm the location and identity of metals in the MD+CTD1 crystal structures, full datasets were collected in succession at the same spot on each crystal at three different wavelengths. A high-resolution native dataset was collected with an exposure of 0.2°/0.4 sec/frame as the primary dataset for model building and refinement. A second dataset was collected at an energy 150-200 keV above the zinc anomalous inflection (9.6586 keV) and a third dataset was collected at an energy 150-200 keV below the zinc anomalous inflection, both with an exposure of 0.2°/0.05 sec/frame.

All data were indexed, integrated, and scaled with DIALS (version 3.8.0)[74] and imported into the CCP4i suite (version 8.0.009)[75] with AIMLESS (version 0.7.9).[76] Phases for all native structures were solved using molecular replacement (MOLREP version 11.9.02)[77] with their corresponding AlphaFold2 model.[15,16] Refinement was done using phenix.refine (version 1.20.1_4487)[78] in conjunction with manual model building in COOT (version 0.8.9.2).[79] B-factors were refined isotropically with automatically determined translation-libration-screw parameters for structures with a resolution lower than 1.5 Å. Structures with a resolution of 1.5 Å or higher were refined with anisotropic B-factors. B-factors for water molecules were always refined isotropically. Alternate conformer and metal occupancies were refined with phenix.refine.[78] Model geometry was analysed and optimized based on suggestions by MolProbity (version 4.5.2).[80] Anomalous maps were generated with the cad[75] and fft[81–83] programs in the CCP4i suite using the finalized refined structures of their respective high-resolution datasets. Data collection and model statistics for all structures are summarized in *Table 2.3.2.1*.

2.3 – Results and Discussion

2.3.1 – Domain Prediction and Construct Design

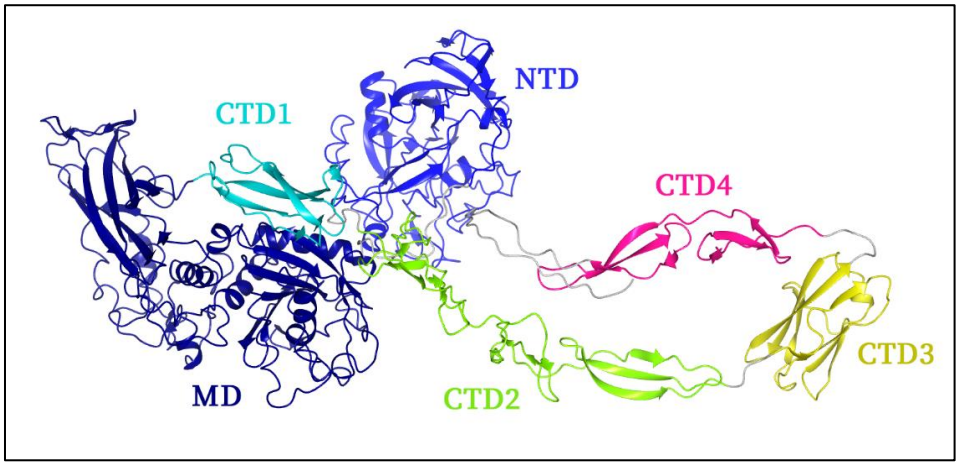


Figure 2.3.1.1 – Putative Domains of the *Thomasclavelia ramosa* Immunoglobulin A Protease. The AlphaFold2-predicted model was split into six putative domains linked together by flexible regions (grey): N-terminal domain (NTD; blue), middle domain or protease domain (MD; dark blue), C-terminal domain #1 (CTD1; cyan), C-terminal domain #2 (CTD2; green), C-terminal domain #3 (CTD3; yellow), and C-terminal domain #4 (CTD4; pink). These predictions formed the basis for the design of all *T. ramosa* IgAP protein constructs.

The AlphaFold2-predicted model (*Fig. 2.3.1.1*) was used as the basis for splitting the protease into its putative individual domains.[15,16] This model was generated using the entire secreted protease sequence (UniProt accession #Q9AES2 residues 31-1195), which was shown to be active when recombinantly expressed in *Escherichia coli*.[67] The predicted structure is reminiscent of the modular architecture of M26 IgA1Ps (*Section 2.1.4*), with domains of different functions strung together by protein linkers. The model was analyzed according to the following set of guidelines to assign domain boundaries:

1. AlphaFold2's predicted aligned error (PAE) metric was used as a measure of intradomain model confidence to assign initial boundaries for each domain.[15]
2. The starting and ending residues of a domain should make enthalpic contributions (*e.g.*, hydrogen bonds, π-stacking interactions, salt bridges) with the rest of the domain; N- or C-terminal residues that do not make enthalpic contributions to the domain in the predicted model are excluded.
3. Domain boundaries should not split any secondary structure elements.
4. Side chains generally are not considered in domain assignment, except for those hypothesized to have a structural and/or biological function.

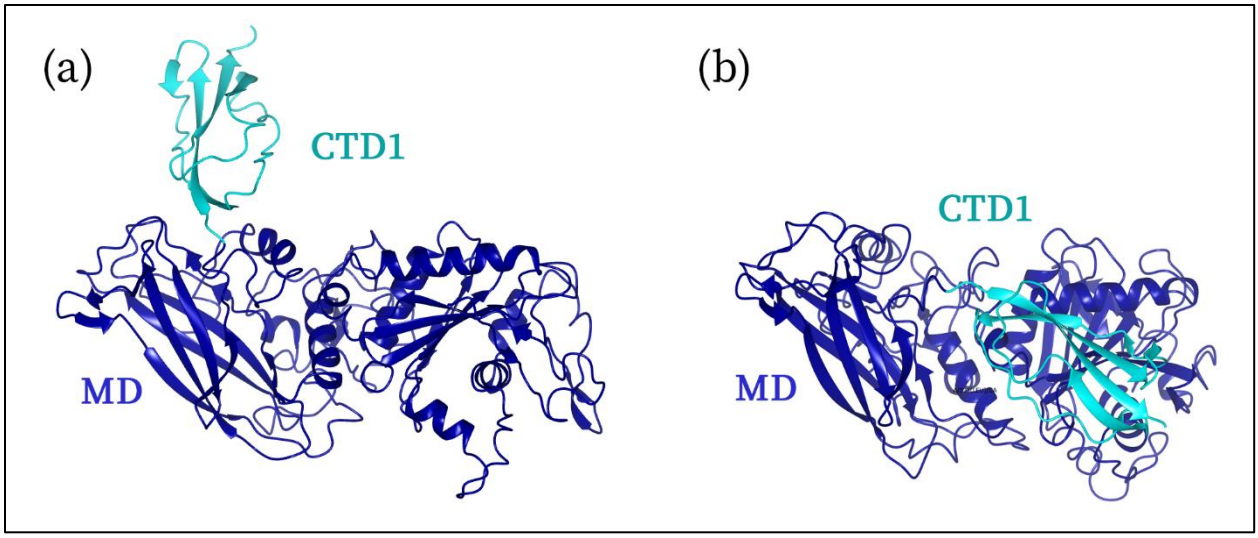


Figure 2.3.1.2 – RosettaFold-Predicted Model for the Protease Domain and Its Adjacent C-terminal Domain. The CTD1 of the *T. ramosa* IgAP is predicted to be a separate domain in the RosettaFold model **(a)** but not in the AlphaFold2 model **(b)**. This suggests that CTD1 might be a true domain despite the predicted confidence in the AlphaFold2 model.

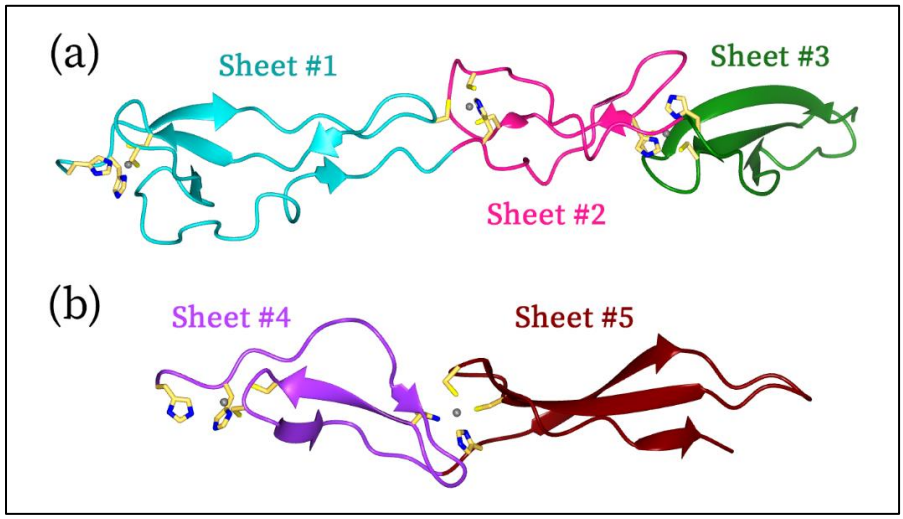


Figure 2.3.1.3 – AlphaFold2-Predicted Model for the C-terminal Putative Zinc-Binding Domains. The CTD2 **(a)** and CTD4 **(b)** domains of the *T. ramosa* IgAP contain several putative zinc-binding sites as a cluster of cysteine and histidine residues are modelled to be in close proximity. The side chains of these metal-binding residues are shown in their predicted conformation and zinc ions (grey spheres) were added post-hoc purely for visualization purposes. CTD2 contains two Cys_2His_2 and one Cys_3His_1 putative zinc-binding sites and CTD4 contains two Cys_2His_2 putative zinc-binding sites.

Following these general rules yielded six putative domains (*Fig. 2.3.1.1*), which are described accordingly:

1. *N-terminal domain (NTD; residues 31-313).* This domain is bioinformatically classified as a domain of unknown function (DUF6273; pfam19789).[84]
2. *Protease domain or middle domain (MD; residues 329-807).* This domain contains the active site and is classified as an M64 protease. Although the last N-terminal residue predicted to contribute enthalpically to the domain via backbone hydrogen bonds is Q331, it was found that an additional two-residue native linker at the N-terminus was needed for purification as the N-terminal tag could not be cleaved off suggesting it was too close to the start of the MD (data not shown).
3. *C-terminal domain #1 (CTD1; residues 808-876).* This domain is predicted to pack against the MD in the AlphaFold2 model. Despite the AlphaFold2 PAE metric[15] and PISA analysis (EMBL-EBI web access on May 1st, 2024)[85] suggesting that the MD and CTD1 form a domain together, the predictive model suggests that the interface is relatively hydrophilic and may be able stable on its own. This is supported by the RosettaFold prediction modelling the CTD1 as its own separate domain (*Fig. 2.3.1.2*).[86]
4. *C-terminal domain #2 (CTD2; residues 880-1008).* Although this domain at first appears to be comprised of three separate beta sheets strung together via small flexible linkers, the side chains of key cysteine and histidine residues between and within each beta sheet are predicted to come together in three putative zinc-binding pockets (*Fig. 2.3.1.3a*).[87] This suggests that the three beta sheets likely behave as a single rigid functional unit.
5. *C-terminal domain #3 (CTD3; residues 1011-1096).* This domain consists of two stacked beta sheets in an immunoglobulin-like fold (Ig fold) and is bioinformatically classified as an X2 carbohydrate-binding domain (pfam03442).[84]
6. *C-terminal domain #4 (CTD4; residues 1102-1166).* Similar to CTD2, this domain is comprised of two beta sheets that contain two putative zinc-binding sites coordinated by cysteine and histidine side chains (*Fig. 2.3.1.3b*)[87]. These beta sheets are thus predicted to behave as a single unit.

2.3.2 – Protein Expression, Purification, and Crystallization

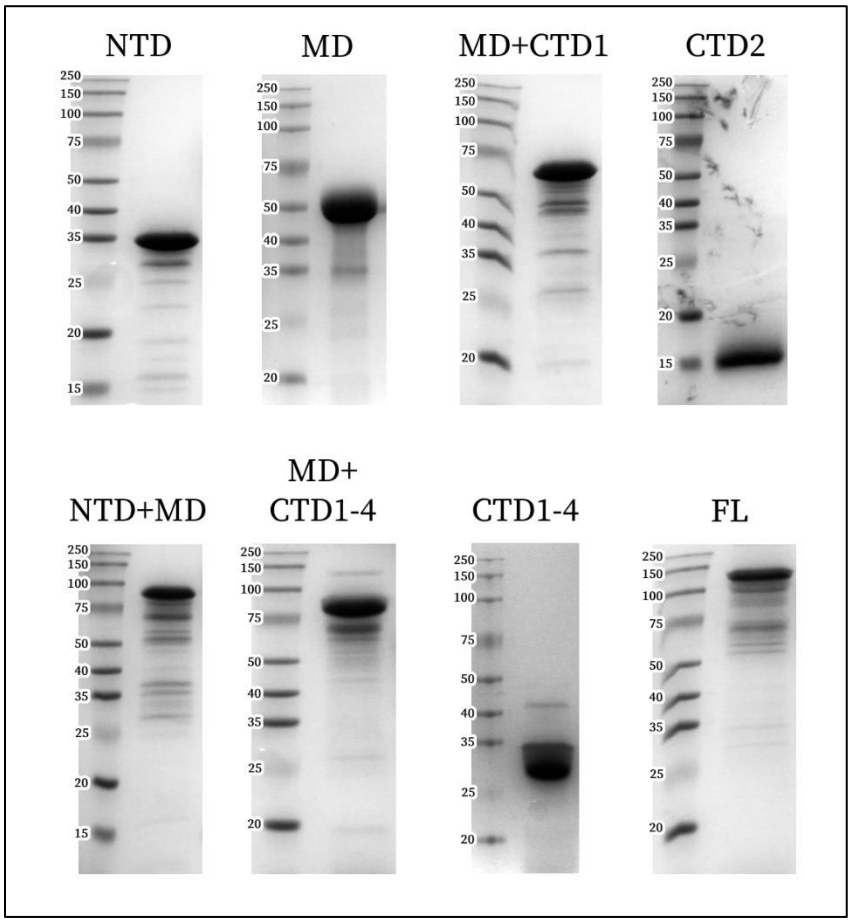


Figure 2.3.2.1 – Purity of Recombinantly Expressed *Thomasclavelia ramosa* Immunoglobulin A Protease Constructs. The purity of all protein constructs was assessed via SDS-PAGE. Constructs consisting of individual domains **(top row)** had a higher purity compared to that of multidomain constructs **(bottom row)**.

All protein constructs were purified to relatively high purity suitable for structural and biochemical characterization (*Fig. 2.3.2.1*), with the exception of CTD3 as it was proteolyzed during expression and unable to be purified with the same purification protocol (data not shown). The fact that most of these constructs were able to be purified without substantial degradation suggested that the domains assigned in *Section 2.3.1* are true domains, in that they can fold and function independent of the rest of the protein. Multidomain constructs, like the full-length enzyme and constructs that contains various CTDs, were less pure compared to constructs consisting of individual domains. Although these multidomain constructs were not suitable for X-ray crystallography, they were suitable for structural and functional studies via solution scattering (*Chapter 3*) and kinetic characterization (*Chapter 4*), respectively.

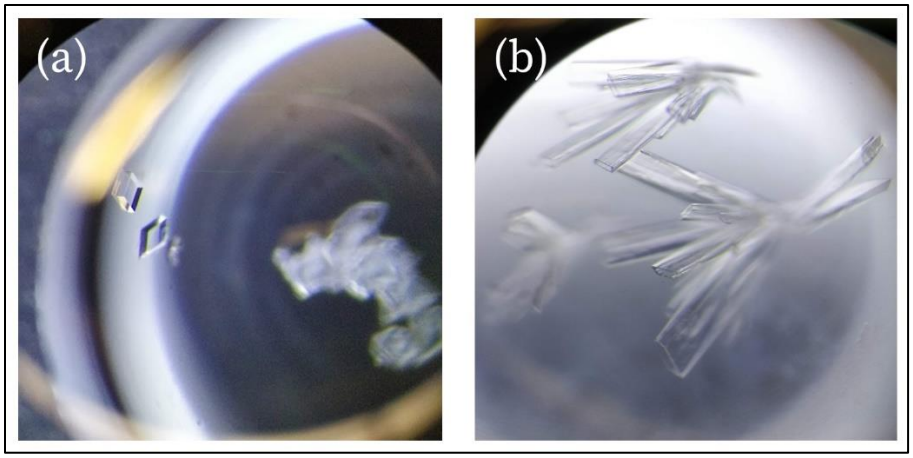


Figure 2.3.2.2 – Morphologies of the NTD and MD+CTD1 Crystals. (a) The NTD forms blocky crystals while **(b)** the MD+CTD1 forms hollow rectangular prisms.

The NTD, MD, and CTD2 proteins had the highest purity out of all the individual domain constructs and hence had the best chances of crystallization. Unfortunately, there were no high-throughput crystal screening hits for MD nor CTD2 both in the initial screening and subsequent re-screenings with varying protein concentrations and supplemented zinc. It was hypothesized that despite the CTD1 not being necessary for the fold and activity of the MD (*Sections 3.3.2* and *4.3.3*), it might have the ability to make the protein more globular and thus more crystallizable.

Initial hits for the NTD and MD+CTD1 proteins were found via high-throughput crystal screening, scaled-up to microliter volumes in hanging-drop plates, and optimized through the use of chemical additives. The optimized NTD crystals had a blocky crystal morphology (*Fig. 2.3.2.2a*) and grew with $NiCl_2$ as the chemical additive. The optimized MD+CTD1 crystals formed rectangular prisms with a zone of depletion in the centre of the crystal (*Fig. 2.3.2.2b*) and grew with praseodymium (III) acetate as the chemical additive.

Both NTD and MD+CTD1 crystals diffracted to high resolution (>1.8 Å) and their structures were solved via standard molecular replacement methods. Diffraction data for MD+CTD1 crystals were also collected at two anomalous wavelengths to assist in metal identification. The rationale for this will be detailed in *Section 2.3.5*. Crystal data collection and refinement statistics are summarized in *Table 2.3.2.1*.

Table 2.3.2.1 – Data Collection and Refinement Statistics. A summary of data and refinement statistics for all crystallographic datasets. Statistics for the highest-resolution bin are in parentheses.

Description of structure	*N-terminal domain*	*WT protease domain (native dataset)*	*WT protease domain (above Zn inflection)*	*WT protease domain (below Zn inflection)*
Wavelength (Å)	0.9686	1.1808	1.2705	1.2970
Resolution range	59.33 – 1.50 (1.53 – 1.50)	55.34 – 1.65 (1.67 – 1.65)	86.79 – 1.85 (1.88 – 1.85)	86.80 – 1.85 (1.88 – 1.85)
Space group	$P2_12_12_1$	$P2_1$	$P2_1$	$P2_1$
Unit cell (Å)	46.94 47.61 118.66	70.69 91.27 86.86 β = 91.09°	70.66 91.23 86.81 β = 91.10°	70.69 91.27 86.86 β = 91.09°
Total reflections	475108 (8054)	803949 (21824)	591061 (17258)	589909 (17004)
Unique reflections	41518 (1383)	122442 (3852)	88545 (3003)	88345 (2963)
Multiplicity	11.44 (5.82)	6.57 (5.67)	6.68 (5.75)	6.68 (5.74)
Completeness (%)	94.79 (64.09)	92.39 (58.61)	94.24 (64.55)	94.04 (63.75)
Mean I/σ(I)	5.1 (0.4)	14.4 (0.6)	12.0 (0.7)	11.9 (0.6)
Wilson B-factor	9.65	21.10	24.65	24.42
R_{merge}	0.162 (0.770)	0.083 (1.410)	0.120 (1.469)	0.119 (1.454)
R_{meas}	0.169 (0.847)	0.090 (1.554)	0.131 (1.617)	0.130 (1.601)
R_{pim}	0.048 (0.341)	0.035 (0.639)	0.050 (0.661)	0.050 (0.653)
$CC_{1/2}$	0.993 (0.712)	0.997 (0.591)	0.995 (0.645)	0.995 (0.661)
Number of reflections used in refinement	39325 (1755)	115075 (2329)		
Number of reflections used for R_{free}	2040 (96)	6315 (150)		
R_{work}	0.1735 (0.2407)	0.1959 (0.3069)		
R_{free}	0.2024 (0.2864)	0.2307 (0.3442)		
Molecules per asymmetric unit	1	2		
Number of atoms	2680	10205		
Protein	2304	8931		
Ligands	1	10		
Water	375	1264		
Average B-factors (Å²)				
Overall	14.83	27.36		
Protein	13.27	26.36		
Ligands	9.34	37.39		
Water	24.41	34.36		
Bond RMSD (Å)	0.010	0.006		
Angle RMSD (°)	1.13	0.80		
Ramachandran (%)				
Favored	96.43	97.52		
Allowed	3.57	2.48		
Outliers	0	0		
Rotamer outliers (%)	1.16	0.10		
Clashscore (percentile)	2.22 (99^{th})	2.62 (99^{th})		
MolProbity score (percentile)	1.28 (95^{th})	1.15 (99^{th})		

Table 2.3.2.1 Continued – *Data Collection and Refinement Statistics*.

Description of structure	*Metal-chelated WT protease domain (native dataset)*	*Metal-chelated WT protease domain (above Zn inflection)*	*Metal-chelated WT protease domain (below Zn inflection)*
Wavelength (Å)	1.1808	1.2651	1.3051
Resolution range	55.30 – 1.50 (1.53 – 1.50)	70.89 – 1.80 (1.83 – 1.80)	70.92 – 1.85 (1.88 – 1.85)
Space group	$P2_1$	$P2_1$	$P2_1$
Unit cell (Å)	70.74 91.55 86.85 β = 90.96°	70.80 91.63 86.90 β = 90.96°	70.83 91.66 86.94 β = 90.96°
Total reflections	1111333 (33717)	620884 (14669)	594648 (17082)
Unique reflections	174570 (7224)	98765 (3397)	93148 (3706)
Multiplicity	6.37 (4.67)	6.29 (4.32)	6.38 (4.61)
Completeness (%)	98.91 (82.51)	96.22 (66.32)	98.46 (79.07)
Mean I/σ(I)	6.3 (0.2)	11.1 (0.6)	11.4 (0.8)
Wilson B-factor	18.31	23.53	24.31
R_{merge}	0.089 (1.174)	0.103 (1.243)	0.103 (0.997)
R_{meas}	0.097 (1.326)	0.112 (1.418)	0.112 (1.128)
R_{pim}	0.037 (0.594)	0.044 (0.658)	0.043 (0.509)
$CC_{1/2}$	0.998 (0.499)	0.996 (0.478)	0.996 (0.604)
Number of reflections used in refinement	165774 (4365)		
Number of reflections used for R_{free}	8718 (240)		
R_{work}	0.1388 (0.2845)		
R_{free}	0.1810 (0.3214)		
Molecules per asymmetric unit	2		
Number of atoms	10467		
Protein	9009		
Ligands	6		
Water	1452		
Average B-factors (Å²)			
Overall	22.36		
Protein	20.54		
Ligands	30.50		
Water	33.61		
Bond RMSD (Å)	0.006		
Angle RMSD (°)	0.79		
Ramachandran (%)			
Favored	98.07		
Allowed	1.93		
Outliers	0		
Rotamer outliers (%)	0.10		
Clashscore (percentile)	2.37 (99^{th})		
MolProbity score (percentile)	1.02 (99^{th})		

Table 2.3.2.1 Continued – *Data Collection and Refinement Statistics*.

Description of structure	*Metal-chelated WT protease domain with 1 mM $ZnCl_2$ (native dataset)*	*Metal-chelated WT protease domain with 1 mM $ZnCl_2$ (above Zn inflection)*	*Metal-chelated WT protease domain with 1 mM $ZnCl_2$ (below Zn inflection)*
Wavelength (Å)	1.1808	1.2651	1.3051
Resolution range	62.73 – 1.78 (1.81 – 1.78)	54.86 – 2.05 (2.09 – 2.05)	54.86 – 2.05 (2.09 – 2.05)
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Unit cell (Å)	71.13 85.64 92.12	71.25 85.75 92.22	71.24 85.76 92.23
Total reflections	689077 (31395)	458850 (23381)	458158 (23257)
Unique reflections	54207 (2637)	35820 (1789)	35826 (1788)
Multiplicity	12.71 (11.91)	12.81 (13.07)	12.79 (13.01)
Completeness (%)	100 (100)	100 (100)	100 (100)
Mean I/σ(I)	11.0 (0.3)	10.1 (0.6)	9.9 (0.6)
Wilson B-factor	28.35	36.84	36.93
R_{merge}	0.102 (2.166)	0.170 (2.430)	0.174 (2.604)
R_{meas}	0.107 (2.264)	0.177 (2.529)	0.182 (2.710)
R_{pim}	0.030 (0.654)	0.049 (0.697)	0.050 (0.748)
$CC_{1/2}$	0.999 (0.586)	0.996 (0.538)	0.998 (0.501)
Number of reflections used in refinement	51393 (2605)		
Number of reflections used for R_{free}	2713 (143)		
R_{work}	0.1707 (0.3448)		
R_{free}	0.2072 (0.3737)		
Molecules per asymmetric unit	1		
Number of atoms	4999		
Protein	4502		
Ligands	13		
Water	490		
Average B-factors (Å^2)			
Overall	36.69		
Protein	36.18		
Ligands	45.63		
Water	41.18		
Bond RMSD (Å)	0.006		
Angle RMSD (°)	0.73		
Ramachandran (%)			
Favored	96.88		
Allowed	3.12		
Outliers	0		
Rotamer outliers (%)	0.20		
Clashscore (percentile)	2.71 (99th)		
MolProbity score (percentile)	1.25 (99th)		

Table 2.3.2.1 Continued – *Data Collection and Refinement Statistics*.

Description of structure	*E540A protease domain (native dataset)*	*E540A protease domain (above Zn inflection)*	*E540A protease domain (below Zn inflection)*
Wavelength (Å)	1.1808	1.2651	1.3051
Resolution range	62.57 – 1.42 (1.44 – 1.42)	62.52 – 1.65 (1.68 – 1.65)	62.58 – 1.70 (1.73 – 1.70)
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Unit cell (Å)	70.57 85.50 91.83	70.49 85.48 91.67	70.54 85.62 91.69
Total reflections	1276345 (63290)	650397 (5914)	594864 (5320)
Unique reflections	103871 (5071)	58910 (1383)	54111 (1265)
Multiplicity	12.29 (12.48)	11.04 (4.28)	10.99 (4.21)
Completeness (%)	98.68 (97.37)	87.17 (41.47)	87.18 (41.26)
Mean I/σ(I)	12.9 (1.9)	8.4 (0.2)	7.7 (0.1)
Wilson B-factor	15.53	26.93	29.80
R_{merge}	0.110 (0.752)	0.105 (1.252)	0.113 (1.149)
R_{meas}	0.115 (0.786)	0.110 (1.429)	0.118 (1.317)
R_{pim}	0.032 (0.225)	0.031 (0.665)	0.033 (0.620)
$CC_{1/2}$	0.998 (0.672)	0.998 (0.590)	0.997 (0.612)
Number of reflections used in refinement	98606 (3170)		
Number of reflections used for R_{free}	5136 (161)		
R_{work}	0.1507 (0.2375)		
R_{free}	0.1834 (0.2808)		
Molecules per asymmetric unit	1		
Number of atoms	5249		
Protein	4583		
Ligands	4		
Water	662		
Average B-factors (Å²)			
Overall	22.25		
Protein	20.71		
Ligands	23.62		
Water	32.86		
Bond RMSD (Å)	0.007		
Angle RMSD (°)	0.92		
Ramachandran (%)			
Favored	98.17		
Allowed	1.83		
Outliers	0		
Rotamer outliers (%)	0.19		
Clashscore (percentile)	2.42 (99^{th})		
MolProbity score (percentile)	1.02 (99^{th})		

Table 2.3.2.1 Continued – *Data Collection and Refinement Statistics*.

Description of structure	*Metal-chelated E540A protease domain (native dataset)*	*Metal-chelated E540A protease domain (above Zn inflection)*	*Metal-chelated E540A protease domain (below Zn inflection)*
Wavelength (Å)	1.1808	1.2651	1.3051
Resolution range	86.64 – 1.70 (1.73 – 1.70)	63.05 – 1.85 (1.88 – 1.85)	70.87 – 1.85 (1.88 – 1.85)
Space group	$P2_1$	$P2_1$	$P2_1$
Unit cell (Å)	70.77 91.66 86.66 β = 91.09°	70.78 91.67 86.65 β = 91.10°	70.78 91.69 86.66 β = 91.09°
Total reflections	769396 (37427)	584215 (17074)	582246 (16698)
Unique reflections	117374 (5752)	89028 (3079)	88789 (3016)
Multiplicity	6.56 (6.51)	6.56 (5.55)	6.56 (5.54)
Completeness (%)	96.62 (94.71)	94.31 (65.45)	94.03 (64.12)
Mean I/σ(I)	4.8 (0.3)	9.2 (0.5)	8.5 (0.4)
Wilson B-factor	18.20	25.94	26.64
R_{merge}	0.150 (1.114)	0.147 (1.934)	0.147 (2.137)
R_{meas}	0.163 (1.214)	0.161 (2.135)	0.160 (2.356)
R_{pim}	0.063 (0.475)	0.063 (0.881)	0.063 (0.968)
$CC_{1/2}$	0.987 (0.537)	0.986 (0.603)	0.989 (0.581)
Number of reflections used in refinement	111427 (3572)		
Number of reflections used for R_{free}	5750 (177)		
R_{work}	0.1694 (0.2731)		
R_{free}	0.2096 (0.3065)		
Molecules per asymmetric unit	2		
Number of atoms	10444		
Protein	8981		
Ligands	8		
Water	1455		
Average B-factors (Å²)			
Overall	24.07		
Protein	22.79		
Ligands	28.32		
Water	31.99		
Bond RMSD (Å)	0.006		
Angle RMSD (°)	0.76		
Ramachandran (%)			
Favored	97.98		
Allowed	2.02		
Outliers	0		
Rotamer outliers (%)	0		
Clashscore (percentile)	2.21 (99th)		
MolProbity score (percentile)	1.00 (100th)		

Table 2.3.2.1 Continued – *Data Collection and Refinement Statistics*.

Description of structure	*Metal-chelated E540A protease domain with 1 mM $ZnCl_2$ (native dataset)*	*Metal-chelated E540A protease domain with 1 mM $ZnCl_2$ (above Zn inflection)*	*Metal-chelated E540A protease domain with 1 mM $ZnCl_2$ (below Zn inflection)*
Wavelength (Å)	1.1808	1.2651	1.3051
Resolution range	62.53 – 1.45 (1.48 – 1.45)	62.60 – 1.75 (1.78 – 1.75)	56.32 – 1.75 (1.78 – 1.75)
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Unit cell (Å)	71.01 85.39 91.82	71.11 85.48 91.94	71.11 85.49 91.95
Total reflections	1230562 (49756)	717470 (31196)	714982 (30917)
Unique reflections	99220 (4874)	57036 (2875)	57026 (2874)
Multiplicity	12.40 (10.21)	12.58 (10.85)	12.54 (10.76)
Completeness (%)	100 (99.51)	100 (100)	100 (100)
Mean I/σ(I)	11.6 (0.2)	14.3 (0.4)	15.7 (0.5)
Wilson B-factor	22.67	28.55	29.49
R_{merge}	0.076 (2.474)	0.091 (2.024)	0.088 (2.107)
R_{meas}	0.079 (2.605)	0.095 (2.124)	0.092 (2.213)
R_{pim}	0.022 (0.805)	0.027 (0.638)	0.026 (0.668)
$CC_{1/2}$	0.999 (0.550)	0.998 (0.648)	0.998 (0.613)
Number of reflections used in refinement	94075 (3052)		
Number of reflections used for R_{free}	4964 (161)		
R_{work}	0.1683 (0.3449)		
R_{free}	0.2049 (0.3622)		
Molecules per asymmetric unit	1		
Number of atoms	5134		
Protein	4516		
Ligands	12		
Water	612		
Average B-factors (Å^2)			
Overall	29.00		
Protein	27.76		
Ligands	41.65		
Water	37.97		
Bond RMSD (Å)	0.005		
Angle RMSD (°)	0.73		
Ramachandran (%)			
Favored	97.98		
Allowed	2.02		
Outliers	0		
Rotamer outliers (%)	0		
Clashscore (percentile)	3.69 (96th)		
MolProbity score (percentile)	1.16 (98th)		

2.3.3 – Crystal Structure of the N-terminal Domain

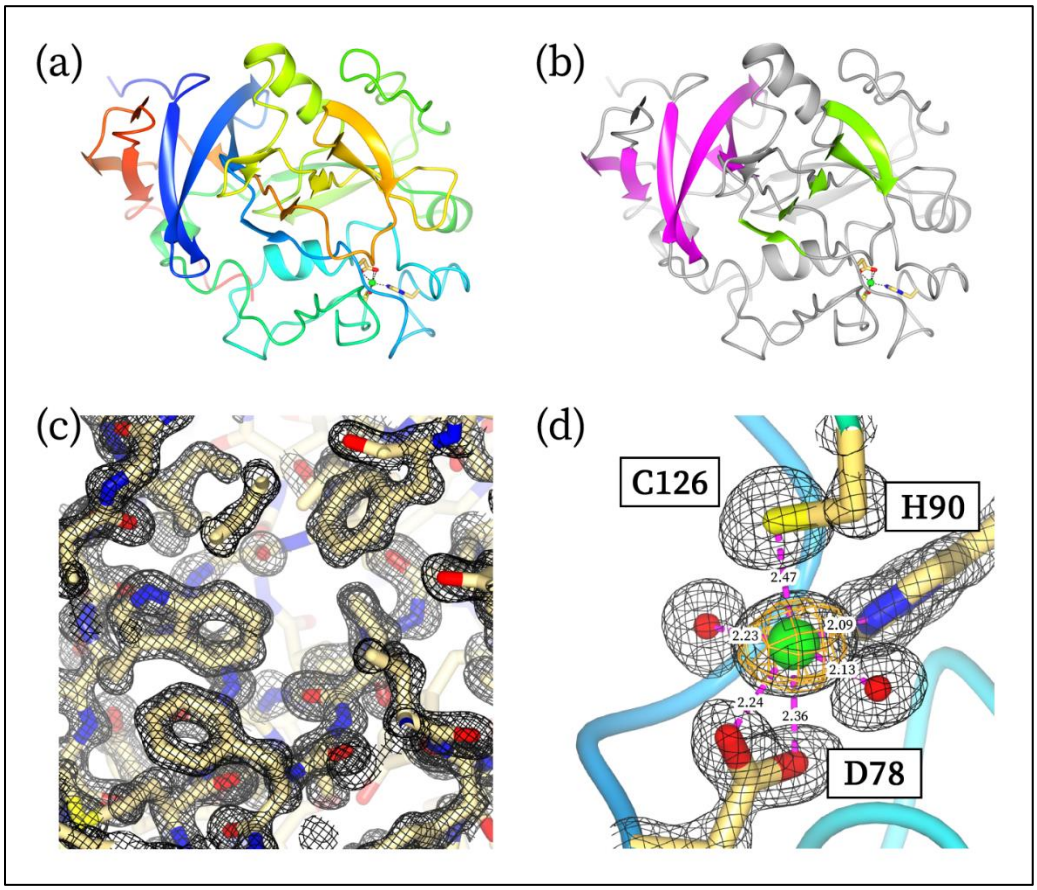


Figure 2.3.3.1 – Crystal Structure of the *Thomasclavelia ramosa* Immunoglobulin A Protease N-terminal Domain. (a) The ribbon structure of the NTD is colored from N- (blue) to C-terminus (red) and shows that the protein is comprised of a heterogeneous mix of mixed beta sheets, some alpha helices, and a significant number of structured loops. **(b)** Two five-stranded beta sheets (magenta and lime green) make up the protein's main secondary structure and hydrophobic core. An example of the good map-to-model fit for the NTD structure is shown in panel **(c)** (black density rendered at 2.0 σ $2F_o$-F_c). **(d)** A nickel ion is present in the crystal structure and its location is supported by an anomalous signal (orange; 5.0 σ anomalous difference density). It is coordinated by three side chains and two water molecules (black density; 2.2 σ $2F_o$-F_c) and is found closer to the core of the enzyme (also depicted in ***a*** and ***b***). Distances between the nickel ion and coordinating atoms (magenta dashes) are shown in Å.

The *T. ramosa* IgAP NTD has a heterogeneous secondary but globular tertiary structure (*Fig. 2.3.3.1*). Although part of its secondary structure is comprised of regular beta sheets and alpha helices, a significant portion of the protein is made up of ordered loops that do not fall into well-defined secondary-structure elements. Despite this, the hydrophobic core largely consists of two large beta sheets (*Fig. 2.3.3.1b*) that are a mix of parallel and anti-parallel beta strands from both the N- and C-terminal ends of the protein (*Fig. 2.3.3.1a*).

The NTD crystal structure also contains a well-coordinated metal ion (*Fig. 2.3.3.1d*) that is found closer to the core of the protein. This was assumed to be nickel as $NiCl_2$ was used as a crystallization additive and, as such, it is unclear whether this is biologically relevant. However, the fact that the metal-binding site is not solvent accessible may give credence to it serving a more structural function. Regardless, it is unlikely that nickel is the biologically relevant metal as it is difficult to obtain, sparely available in most human organs, and thought to be related to pathogenicity via nickel-dependent virulence factors.[88] Zinc may be more relevant as there is precedent for zinc binding in many of the other *T. ramosa* IgAP domains and the geometry of the metal-binding site can likely accommodate both nickel and zinc (CheckMyMetal; web server accessed April 27, 2024).[89]

The NTD is bioinformatically classified as a domain of unknown function (DUF6273; pfam19789). Bioinformatic analysis via InterPro (entry #IPR046240) reveals that it is primarily found in proteins from bacteria within the order *Eubacteriales*, of which *T. ramosa* is a species of, and is typically produced by itself without any additional appended domains.[84] Like in the *T. ramosa* IgAP, it can also be found adjacent to other domains with various predicted functions, ranging from protein kinase to surface-adhesion domains, suggesting that it has its own separate function in the context of these proteins.[84]

5 to 10% of all annotated DUF6273 domains can also be found in proteins from viruses, the majority of which are in the class *Caudoviricetes*, a class of bacteriophages.[84] Predicted bacteriophage genes that contain DUF6273 domains have been postulated to be involved in viral hypermutation.[90] These hypermutating mechanisms are commonplace in viruses of the gut virome and are thought to act as a mechanism to quickly adapt to changing environments.[90] Baykov *et al.* sequenced the full genomes of two gut bacteriophages, and the genes targeted by its conserved hypermutation elements were identified to modify DUF6273 domains.[90] Through structural prediction via AlphaFold2 and structural homology searches using DALI, these researchers postulated that these viral DUF6273 domains (*Fig. 2.3.3.2a*) resemble C-type lectins,[90] which are involved in carbohydrate binding, cell adhesion, and receptor binding.[91] However, when I re-analyzed these viral domains and the *T. ramosa* IgAP NTD with DALI,[92] the structural alignment only covered around half of the domain's structure (*Fig. 2.3.3.2b*); the highest-scored hits were oligomeric coagulation factors and toxins (C-type lectins were present with lower scores on the

list), and their low Z-scores (< 6.0) suggest that these hits may not be truly homologous.[93] This re-analysis thus casts doubts on whether or not these viral proteins and the NTD function as C-type lectins.

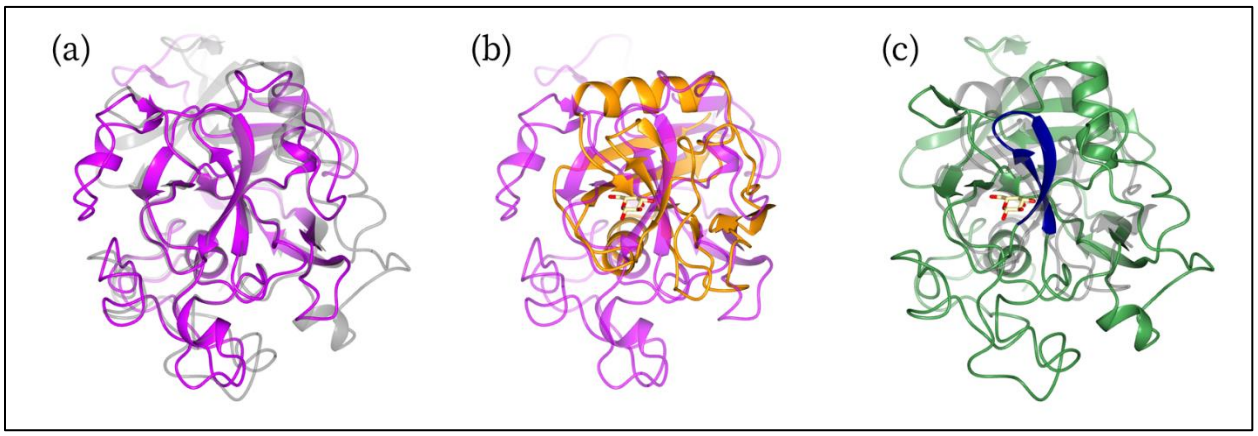


Figure 2.3.3.2 – The N-terminal Domain and Its Relationship to Bacteriophage Proteins and C-type Lectins. **(a)** The *T. ramosa* IgAP NTD (semi-transparent grey; crystal structure) and some bacteriophage proteins of unknown function (purple; AlphaFold2 model) share a similar fold and are a part of the DUF6273 family of domains.[15,16] This viral protein is encoded in the genome of a newly sequenced gut-associated bacteriophage called nd4 (protein is encoded from 21198 to 22826 bp in the genome) and has been postulated to be related to C-type lectins.[90] **(b)** This is hypothesized based on the structural homology between the bacteriophage protein (purple) and C-type lectins, an example of one is shown in this panel (orange; PDB 1TLG). **(c)** A beta-hairpin structure (blue) that is structurally conserved between the NTD (green), viral DUF6273 domain, and the example C-type lectin (grey) is the region where the viral genome is hypothesized to undergo hypermutation. This beta-hairpin surface may be important for ligand interactions as it is the site of galactose binding (shown as cylinders) in the C-type lectin shown here (PDB 1TLG).[94]

Despite this, the fact that DUF6273 domains may be a target in viral hypermutation gives some insight into what structural elements may be important for biological function. The site of hypermutation in the DUF6273 domains is a beta-hairpin structure found on the surface of the protein (*Fig. 2.3.3.2c*).[90] This structure is where C-type lectins and DUF6273 domains are most closely structurally aligned and also corresponds to the ligand-binding surface of some C-type lectins.[90,94] As such, it is possible that this surface plays an important role in the biological function of DUF6273 domains and further research on potential ligands for these domains will provide context for its role in bacteria and viruses as well as its relevance to the *T. ramosa* IgAP.

2.3.4 – Crystal Structure of the Protease Domain

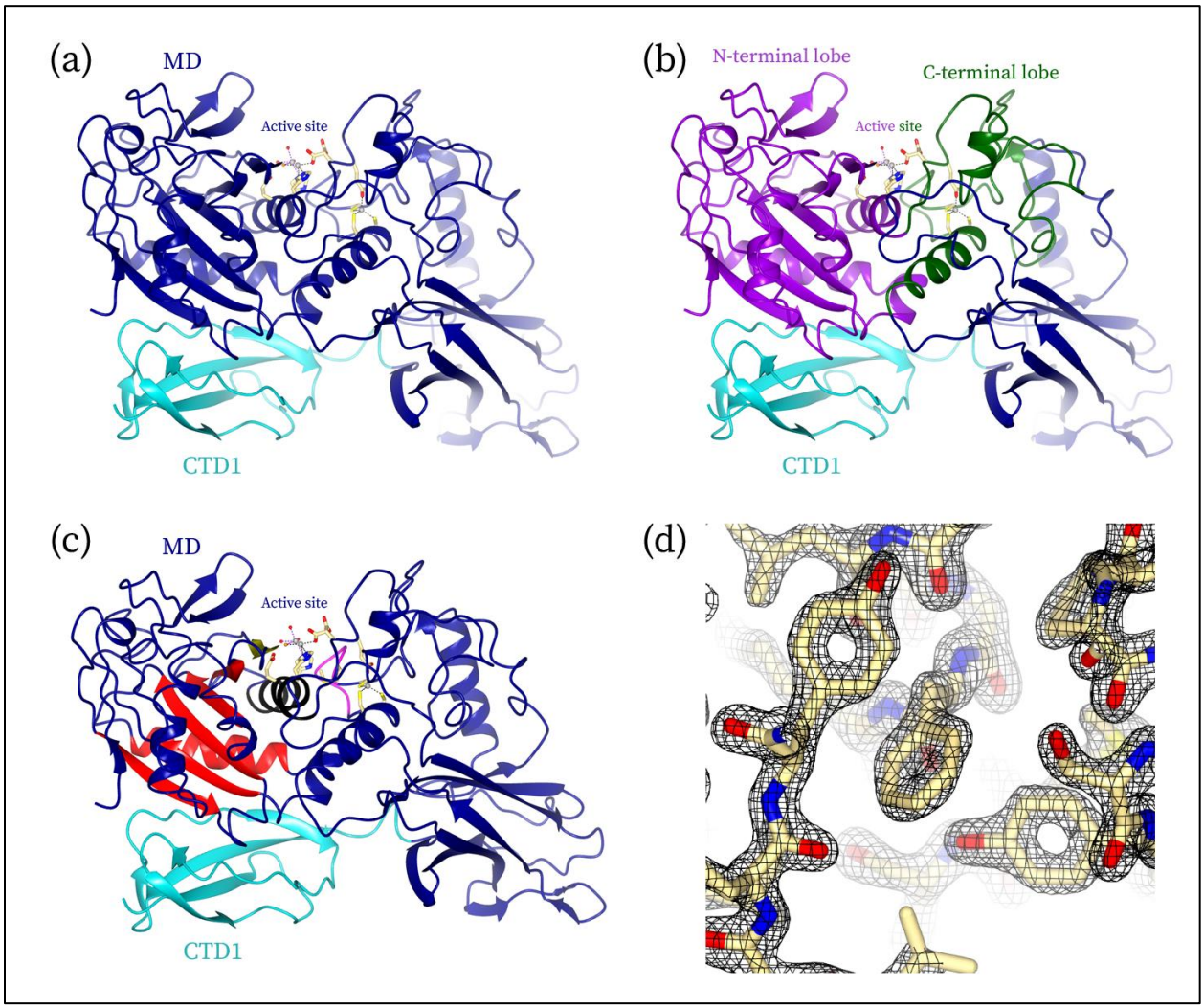


Figure 2.3.4.1 – Crystal Structure of the *Thomasclavelia ramosa* Immunoglobulin A Protease Middle Domain. (a) The crystal structure of the protease domain (middle domain; MD; dark blue) shows features characteristic of a metzincin metalloprotease fold. The MD was only able to crystallize with CTD1 (cyan) as it likely made the enzyme more globular and crystallizable. The structure depicted is of the wild-type metal-chelated MD+CTD1 protease that was supplemented with 1 mM $ZnCl_2$, or in other words the biological holoenzyme. Residues involved with zinc binding (grey spheres) are shown as cylinders. The active-site zinc is poorly ordered and occupies two slightly adjacent positions. **(b)** The active site is found at the interface between two lobes (purple and green) that contain either the HEXXH active-site helix or the C-terminal aspartic acid. The C-terminal lobe also contains a second zinc-binding pocket. **(c)** The metzincin fold has the following common features that are shared in the *T. ramosa* IgAP. 1) An alpha helix backing a large beta sheet (red). The incoming peptide substrate makes an antiparallel beta-strand interaction with the last strand of this beta sheet adjacent to the active site (yellow). 2) An active-site alpha helix (black) that contains the HEXXH motif. And 3) a loop containing a conserved methionine called the Met-loop (magenta). An example of the good map-to-model fit for the MD+CTD1 structure is shown in panel **(d)** (blue density rendered at 2.5 σ $2F_o$-F_c).

The structure of the *T. ramosa* IgAP MD (*Fig. 2.3.4.1a*) falls under a larger family of metalloproteases known as metzincins. Metzincins have several structural features that make up the catalytic fold that are conserved in the MD (*Fig. 2.3.4.1bc*).[95] The active site houses a single catalytic zinc atom that is coordinated by residues found on an active-site alpha helix and a C-terminal glutamic or aspartic acid.[95] The peptide-binding cleft is sandwiched between two lobes of the enzyme that contain either the HEXXH helix or the C-terminal acidic residue and runs parallel to the active-site helix (*Fig. 2.3.4.1b*).[95] This division is exemplified in the *Streptococcus pneumoniae* and *Gemella haemolyans* IgA1Ps, where the two halves can be purified separately and combined *in vitro* to regain activity (*Section 2.1.4*).[11,12,46] The consensus sequence for the metzincin active-site motif extends past the traditional metalloprotease HEXXH motif to include a flexible glycine and the C-terminal metal-binding residue to become **H**EXX**H**XXGXX**D/E** (bolded residues are involved in metal binding).[95] The metzincin fold also contains a large beta sheet backed by a supporting alpha helix in the N-terminal lobe, which stabilizes the incoming peptide substrate by forming an extended antiparallel beta sheet over the active site (*Fig. 2.3.4.1c*).[95] Finally, the metal-binding site is structurally supported by a conserved methionine on a loop in the C-terminal lobe called the Met-turn (*Fig. 2.3.4.1c*).[95]

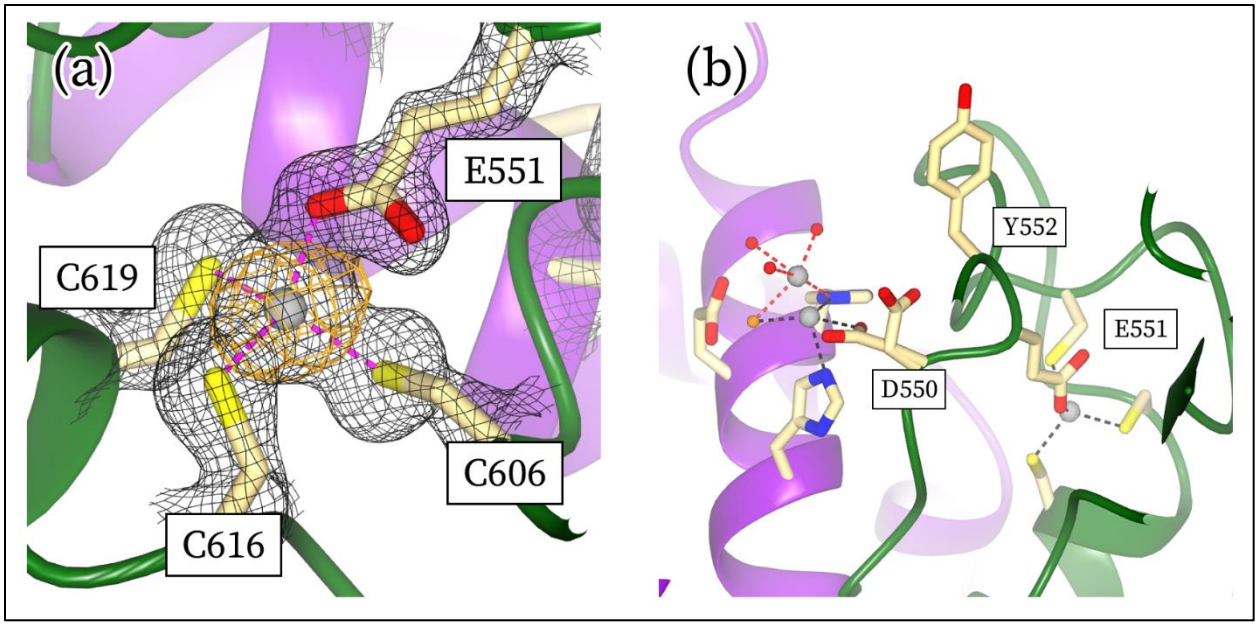


Figure 2.3.4.2 – The Non-Catalytic Zinc-Binding Site in the Protease Domain. (a) Unlike most other metzincins, the *T. ramosa* IgAP MD contains a second zinc-binding site adjacent to the active site. This zinc is coordinated (magenta dashes) by the side chains of three cysteines and one glutamic acid (black density rendered at 2.5 σ $2F_o$-F_c; anomalous difference map rendered at 5.0 σ in orange) and is linked to the active site by a loop that bridges both metal-binding sites **(b)**. This bridging loop is a part of the consensus motif (DEY) found C-terminal to the standard metzincin consensus sequence. The structure depicted is of the wild-type zinc-supplemented MD+CTD1, or in other words the biological holoenzyme.

Unlike the majority of metzincins, the *T. ramosa* IgAP MD contains an additional non-catalytic zinc-binding site in the C-terminal lobe (*Fig. 2.3.4.2a*). This second metal-binding site is adjacent to the active site and is bridged by a loop that coordinates both metals (*Fig. 2.3.4.2b*). Hence the active-site consensus sequence is extended beyond the metzincin sequence to include this loop and the second metal-binding residues to become **H**EXX**H**XXXGXG**DE**Y$(X)_{53}$**C**$(X)_9$ **C**XX**C**. This distinct subgroup of metzincins is classified as the M64 family of proteases and the role and importance of the *T. ramosa* IgAP within that family will be discussed in *Section 2.3.8*.

2.3.5 – The Role of Zinc in the Protease Domain

Zinc plays an important structural and functional role in the *T. ramosa* IgAP as the NTD, MD, and several CTDs contain metal-binding sites. The second zinc-binding site in the MD is distinct amongst metzincins and it is only found in the M64 family as well as some archaeal proteases.[95,96] These archaemetzincins also share a loop that coordinates both metal-binding sites, however the glutamic acid that coordinates the second site in the M64 proteases is substituted by

a cysteine in archaemetizincins.[96] The distances in the primary sequence between each of the residues that coordinate the second site also differs between the two families.[96]

It is unclear if this second zinc-binding site has any biological role in addition to structural stabilization. The second site in archaemetzincins has been postulated to act as a redox sensor whereby oxidation of the cysteines and release of zinc may trigger large conformational changes that are required for activity.[96,97] However, this hypothesis is based on the redox-sensitive structure of the *E. coli* Hsp33 protein whose Cys_4 zinc-binding site is located on a flexible, solvent-exposed region of the protein, contrary to the more buried second zinc-binding sites of the archaemetzincins and *T. ramosa* IgAP protease domain.[98]

Given the proximity of the second zinc to the active site and the presence of a conserved loop coordinating both metals, it would be surprising if the second site did not have a regulatory function. One can speculate that the second site may play a role in modulating the affinity of zinc to the active-site as removing zinc from the second site, even without cysteine oxidation, will likely impact the dynamics of the bridging loop and thus one of the active-site zinc ligands. Regardless, the role of the second zinc-binding site remains speculative and more experimental evidence is needed to make any claims on function.

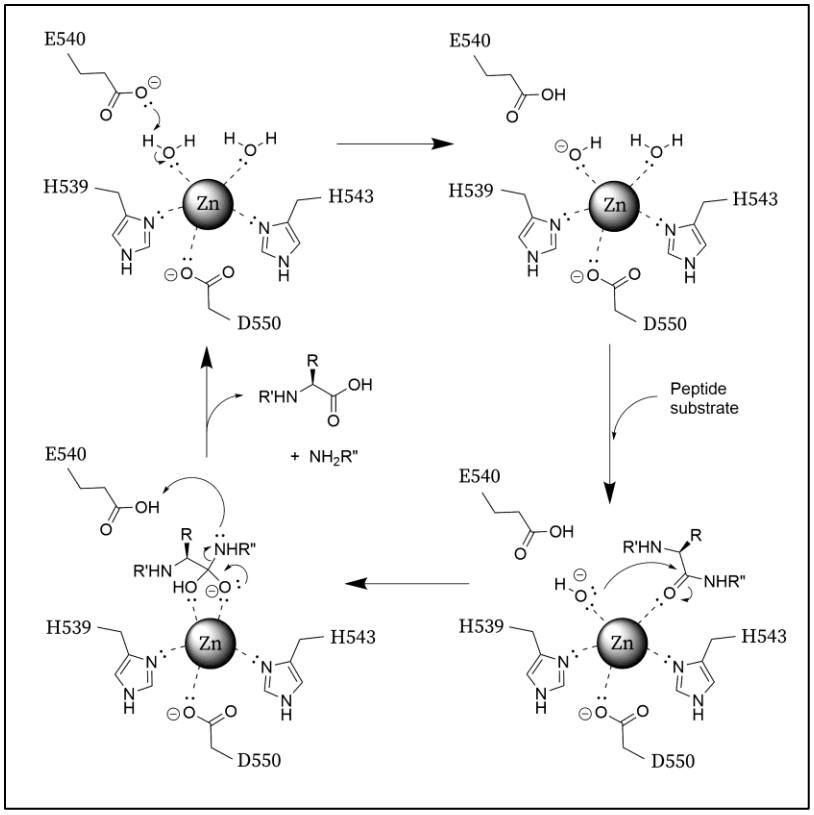


Figure 2.3.5.1 – Catalytic Mechanism of Zinc Metalloproteases. The conserved chemical mechanism used by zinc metalloproteases is shown with four distinct steps. **(a, top left)** The active-site zinc is coordinated by H539, H543, D550, and two water molecules. **(b, top right)** The water molecule closest to the catalytic base, E540, is deprotonated to form the nucleophilic zinc-hydroxide complex. **(c, bottom right)** The peptide substrate binds to the active-site groove with the carbonyl of the scissile bond being stabilized by the zinc ion. The nucleophilic hydroxide then attacks the peptide bond. **(d, bottom left)** A tetrahedral intermediate is formed before breaking down into the two proteolysis products.

The active-site zinc, on the other hand, is well understood in terms of its mechanism and biological role. The chemical mechanism for metalloproteases has been summarized previously in *Section 2.1.4* with respect to M26 IgA1Ps, however for the sake of clarity and more detail, it will be reiterated with respect to specifics of the *T. ramosa* IgAP (*Fig. 2.3.5.1*).[51] H539 and H543 on the active-site helix (**H**EXX**H** motif; *Fig. 2.3.4.1c*) and D550 on the bridging loop (*Fig. 2.3.4.2b*) are the three ligands that coordinate the active-site zinc (*Fig. 2.3.5.1a*). A fourth zinc ligand is a water molecule that becomes the nucleophile in the proteolysis reaction.[51] This water molecule is polarized by the zinc and is deprotonated by E540 on the active-site helix (H**E**XXH motif; *Fig. 2.3.4.1c*) to form a Zn^{2+}-hydroxide complex (*Fig. 2.3.5.1b*).[51] Once the substrate enters the active site, the carbonyl oxygen of the to-be-hydrolyzed peptide bond becomes another ligand of the active-site zinc and is polarized, like the water molecule, to make it susceptible to nucleophilic attack (*Fig. 2.3.5.1c*).[95] The hydroxide nucleophile then attacks the carbonyl to form a tetrahedral intermediate, whose negative charge is stabilized by the zinc ion, acting like an oxyanion hole (*Fig. 2.3.5.1d*).[95] The intermediate breaks down into the N- and C-terminal products which are then released to regenerate the free enzyme.[95]

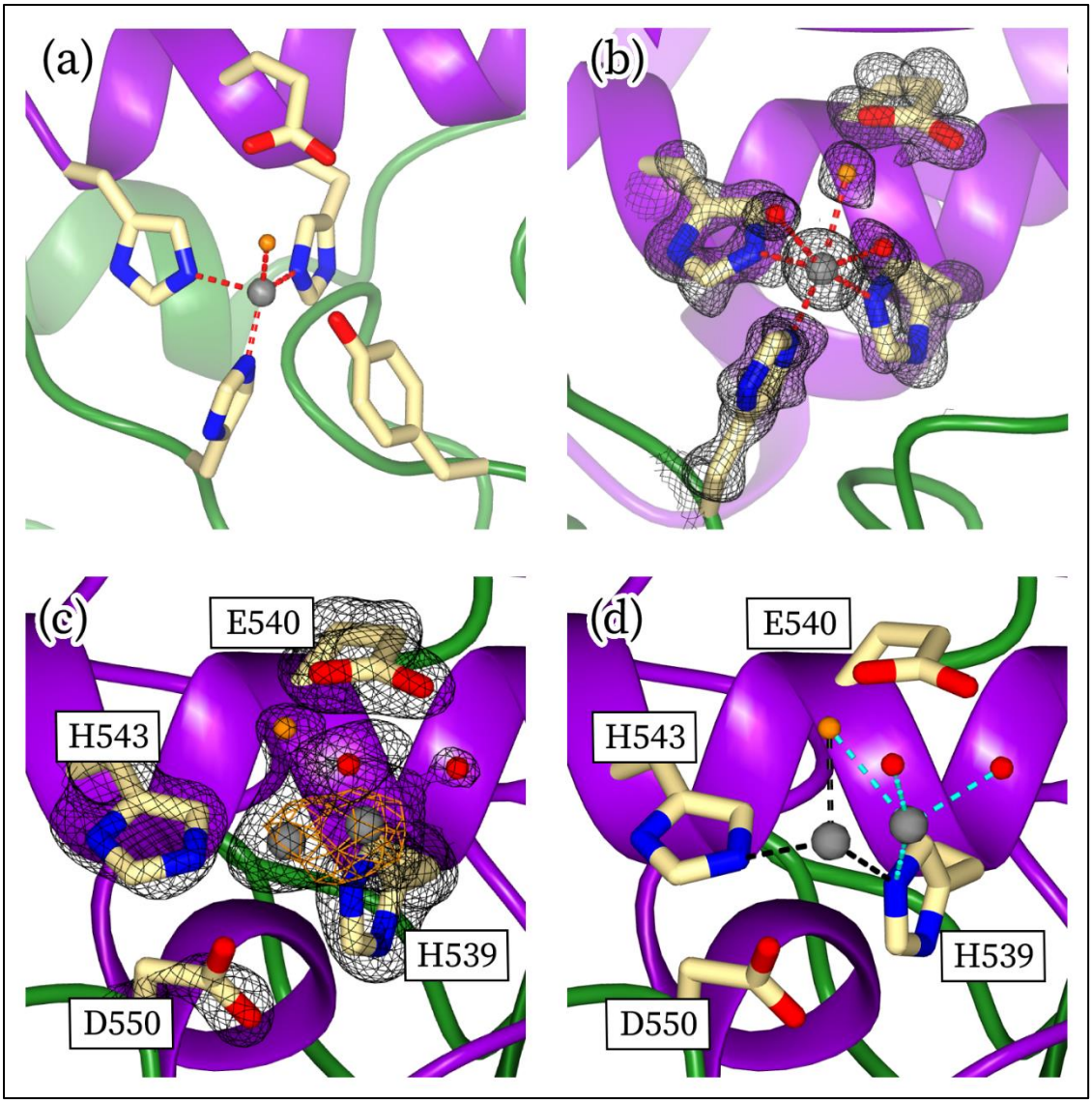


Figure 2.3.5.2 – Conformational Flexibility in the Active Site of the Wild Type Protease Domain. Compared to the holoenzyme structures of astacin from *Astacus astacus* **(a)** (PDB 1AST; no publicly available density maps)[101] and human matrix metalloprotease-12 **(b)** (PDB 2OXU; black $2F_o$-F_c density contoured at 1.5 σ),[99] the structure of the wild type as-purified MD+CTD1 enzyme is relatively disordered at the active site. **(c)** Zinc (grey spheres) can be modelled in two positions, with the side chain of D550 too far to coordinate with the active-site zinc. This is supported by anomalous density (orange) at 4.0 σ and $2F_o$-F_c density (black) at 1.5 σ. **(d)** When zinc is found deeper into the active site (left sphere), it interacts with H539, H543, and the nucleophilic water that is hydrogen bonded with catalytic base (black dashes). The zinc at the alternate position (right sphere) has a separate set of coordinating ligands (cyan dashes), with H539, the nucleophilic water, and several other water molecules that are too far to coordinate with the zinc at the other position. The nucleophilic water molecules for all panels are colored in orange.

Early studies on model metalloproteases have shown that the active-site zinc, coordinating water molecule(s), and zinc-binding residues are well ordered and in their catalytically competent state in preparation for an incoming substrate (*Fig. 2.3.5.2ab*).[95,99–101] In contrast to these structures, the first set of crystal structures solved of the *T. ramosa* IgAP protease domain depicts a conformationally flexible active site with a partially disordered zinc ion. The zinc ion was modelled in two positions, with one of the zinc-binding residues (D550) being too far to coordinate the metal, and the water molecules coordinating the zinc are somewhat disordered (*Fig. 2.3.5.2cd*).

It was hypothesized that this disorder was due to the enzyme binding to a variety of metals during expression and purification, as zinc was only added in trace amounts during expression and was not added in purification. To resolve this, the purified enzyme was incubated with a metal-chelating resin and crystallized in its metal-free state. These crystals were then used as-is or soaked with 1 mM $ZnCl_2$ prior to cryoprotection and freezing. The crystal structure of the as-purified enzyme was compared to the metal-chelated and zinc-supplemented structures to look for structural changes, particularly at the two metal-binding sites. In addition to the high-resolution dataset used for structural refinement, two datasets were collected at an energy either 150-200 keV higher (termed the "above-peak" dataset) or lower (termed the "below-peak" dataset) than the zinc anomalous inflection energy (9.6586 keV) to experimentally confirm the presence or absence of zinc and other metals at both sites.[102] A 150 keV energy buffer was used to account for minute changes in the zinc absorption peak energy relative to that of its theoretical energy[102] and for the larger wavelength bandwidth of the CMCF-BM beamline's optics (personal communication, Dr. Dean Lang, beamline scientist at CMCF). Anomalous density in the above-peak data but not in the below-peak data confirms the presence of zinc at those locations.[102] On the other hand, anomalous density in both above- and below-peak data suggest that there are other non-zinc transition metals bound at that site.[102]

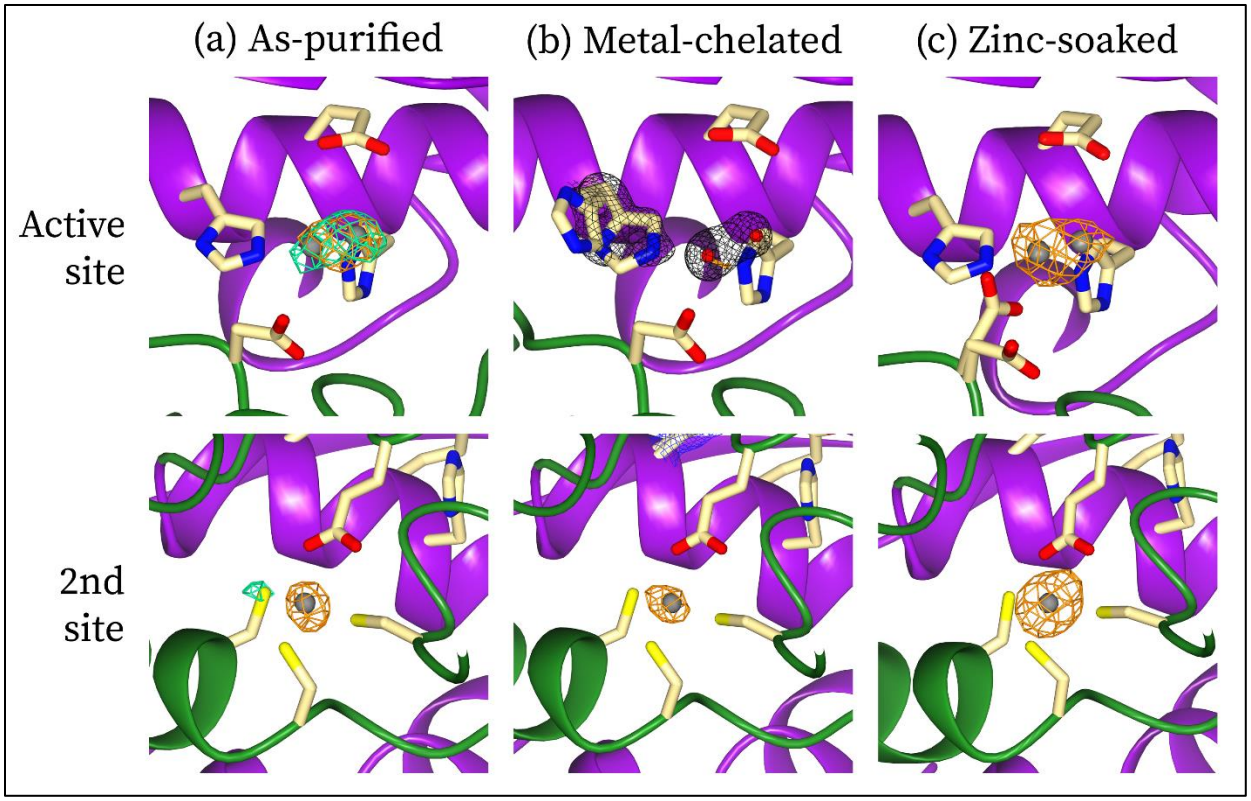


Figure 2.3.5.3 – Anomalous Density Analysis for the Wild Type Protease Domain. The anomalous density maps for the active (top row) and second (bottom row) sites are shown for the wild type protease domain. Maps for the above- (orange) and below-peak (green) data are contoured at ~0.10 e^-/A^3 and overlayed over protease domain structure (N-terminal lobe is purple and C-terminal lobe is dark green). **(a)** The as-purified enzyme is bound to a heterogeneous mixture of metals at the active site, a notable portion of which is not zinc (3.9 σ above-peak; 3.8 σ below-peak). The second site is always selective for zinc as no anomalous signal is seen in the below-peak data. **(b)** The metal-chelated enzyme has no metal at the active site (4.3 σ above-peak; 4.4 σ below-peak) and instead has two water molecules that occupy the same space (black density; 1.3 σ $2F_o$-F_c). H543 takes on two conformations due to the lack of active-site metal. **(c)** Once zinc is soaked back into the crystal, both pockets become saturated with zinc (5.3 σ above-peak; 5.2 σ below-peak) and display strong anomalous densities in the above-peak but not in the below-peak data. H543 no longer has an alternate conformation and D550 is shown to partially coordinate the loosely bound zinc.

Anomalous density analysis of the crystal structure of the wild type as-purified protease domain supports the hypothesis that the enzyme is bound to a heterogeneous mixture of transition metals at the active site (*Fig. 2.3.5.3a*). The second metal-binding site, on the other hand, is selective for zinc as there no anomalous signal in the below-peak data for any of the structures collected on the MD+CTD1 crystals. When the wild type enzyme was treated with metal-chelating resin, the active-site metal was completely chelated while the second site was not completely stripped of zinc (*Fig. 2.3.5.3b*), likely due to it being less solvent exposed. However, both sites became fully occupied with zinc (*Fig. 2.3.5.3c*) upon addition of 1 mM $ZnCl_2$ and thus this structure represents the holoenzyme of biological interest and further discussion.

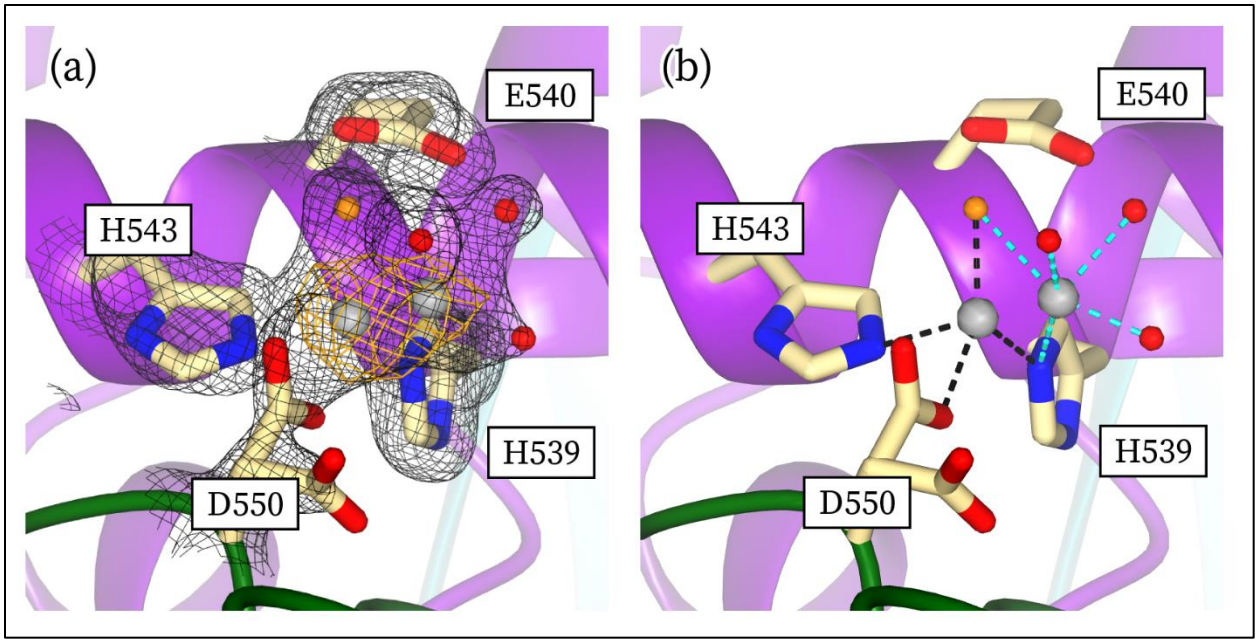


Figure 2.3.5.4 – Inherent Disorder in the Active-Site Zinc-Binding Pocket of the Protease Domain. The crystal structure of the wild type zinc-supplemented MD+CTD1 enzyme shows disorder at the active-site zinc. **(a)** Zinc (grey spheres) can be modelled in two positions, supported by anomalous density (orange) at 5.0 σ, and D550 can be modelled in two conformations (black $2F_o$-F_c density contoured at 1.3 σ) corresponding to these two zinc positions. **(b)** When zinc is found in its active position (left sphere), it is properly coordinated (black dashes) with H539, H543, D550, and the nucleophilic water (orange) that is hydrogen bonded with the catalytic base (E540). The zinc at the alternate position (right sphere) is coordinated (cyan dashes) with H539, the nucleophilic water, and several other water molecules that are too far to coordinate with the zinc at the active position.

The active site of the zinc-bound holoenzyme was still relatively disordered despite the full occupancy of only zinc at the active site (*Fig. 2.3.5.4*). The same features were seen in both as-purified and zinc-supplemented structures, with the active-site zinc occupying the same two locations and the water molecules coordinating the zinc remaining relatively disordered. However unlike in the as-purified structure (*Fig. 2.3.5.2cd*), the side chain of D550 in the zinc-supplemented structure was shown to coordinate the active-site zinc, albeit only with a ~50% occupancy between the coordinating and non-coordinating conformations (*Fig. 2.3.5.4*). Overall, this suggests that the active site of the holoenzyme is disordered unlike many other metalloproteases (*Fig. 2.3.5.2*) and that this disorder is inherent to the protein and not due to the binding of other metals at the active site. This disorder is hypothesized to have significant functional consequences on how it binds to and selects for its substrate. These details and the residues proposed to be involved in binding the peptide will be the focus of the *Section 2.3.7.*

2.3.6 – The E540A Mutation Quenches Disorder in the Protease Domain Active Site

The catalytic base was mutated to an alanine (E540A) in the background of the MD+CTD1 construct to generate a catalytically inactive protein for the purposes of forming substrate and product complexes with the protease domain. Although no stable complexes were able to be successfully made, the structure of the E540A MD+CTD1 mutant revealed essential structure-function relationships in the enzyme.

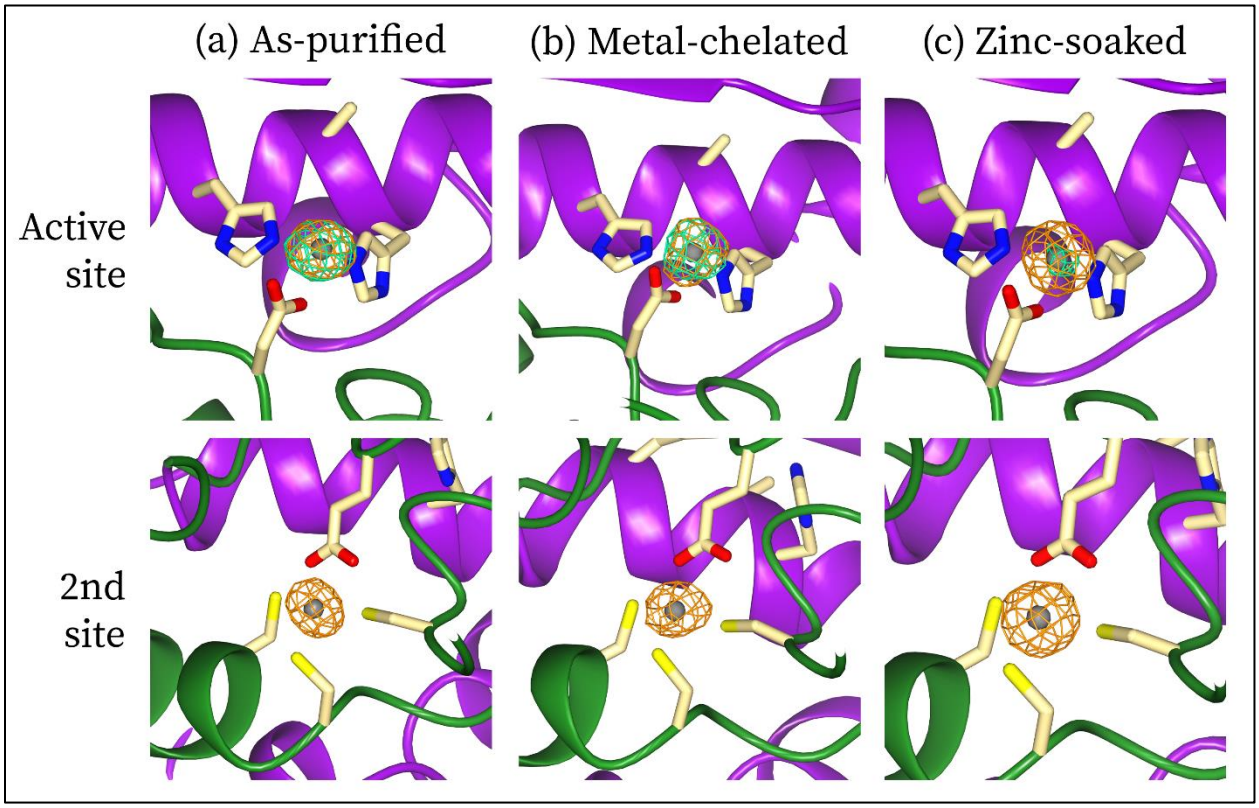


Figure 2.3.6.1 – Anomalous Density Analysis for the Protease Domain E540A Mutant. The anomalous density maps for the active (top row) and second (bottom row) sites are shown for E540A protease domain mutant. Maps for the above- (orange) and below-peak (green) data are contoured at ~0.10 e^-/A^3 and overlayed over the protease domain structure (N-terminal lobe is purple and C-terminal lobe is dark green). **(a)** The as-purified protein is bound to a heterogeneous mixture of metals at the active site, a significant portion of which is not zinc (8.6 σ above-peak; 9.7 σ below-peak). The second site is always selective for zinc as no anomalous signal is seen in the below-peak data. **(b)** The metal-chelated protein is still bound to a heterogeneous mixture of metals at the active site (3.2 σ above-peak; 3.5 σ below-peak). **(c)** Once zinc is soaked back into the crystal, both pockets become saturated (6.8 σ above-peak; 7.5 σ below-peak). Compared to the as-purified structure **(a)**, the active site is bound to a larger proportion of zinc relative to contaminating metals, as seen in the difference in anomalous peak intensities for the data above- and below-peak.

Similar to the wild type enzyme described above, the initial structure of the as-purified E540A MD+CTD1 protein showed that its active site was bound to a heterogeneous mixture of metals (*Fig. 2.3.6.1a*) while the second metal-binding site was selective for zinc. It came as a surprise that, unlike the wild type enzyme, the active-site metal was not able to be fully removed from the E540A mutant, even when the protein was treated with the same experimental conditions as the wild type enzyme (*Fig. 2.3.6.1b*). Contaminating non-zinc metal(s) were thus still seen when the metal-chelated crystals were soaked with 1 mM $ZnCl_2$ (*Fig. 2.3.6.1c*). However, because the absolute intensity of the anomalous density for the zinc-soaked crystals above the zinc peak was higher than below the zinc peak (*Fig. 2.3.6.1c*), the active site had the capacity to bind exogenous zinc despite the inability to chelate all contaminating transition metals.

The increased order of the E540A mutant active-site metal not only manifests in the architecture of the active site, as discussed below, but also more globally in the context of the crystal. The global tertiary structure of the E540A MD+CTD1 protein is nearly identical to the wild type enzyme, with an alpha-carbon root mean square deviation (C_α-RMSD) of ~0.1 Å between the two structures (as determined by UCSF Chimera).[103] Local differences, however, allow MD+CTD1 proteins to crystallize in two related space groups, $P2_1$ and $P2_12_12_1$, with a nearly identical crystal packing and unit cell (*Table 2.3.2.1*). As a general trend, the protein tends to crystallize in $P2_1$ when the protein active site is not bound to metal at full occupancy either due to inherent dynamics in the active site (as-purified wild type structure) or because of metal chelation. This heterogeneity breaks the $P2_12_12_1$ crystallographic symmetry, slightly skews one of the 90^o angles to become the unique beta angle of ~91^o in the monoclinic unit cell, and results in two molecules in the asymmetric unit related by non-crystallographic symmetry as opposed to one molecule. $P2_1$ datasets can easily be erroneously indexed and scaled in $P2_12_12_1$ due to their close geometrical relationship, however the merging R-factors for $P2_1$ datasets processed in $P2_12_12_1$ are notably higher (overall R_{merge} of ~0.25-0.35) relative to that in their true space group (0.07-0.15; *Table 2.3.2.1*).

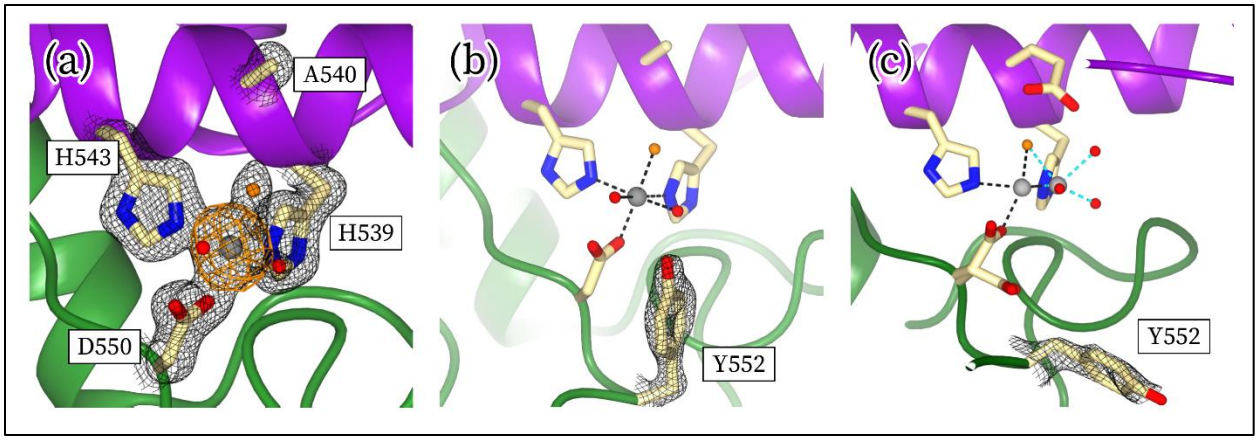


Figure 2.3.6.2 – The E540A Mutation Orders the Active Site of the Protease Domain. (a) The crystal structure of the E540A zinc-supplemented protease domain shows a well-coordinated active-site zinc (grey sphere; orange anomalous density at 5.0 σ; black $2F_o$-F_c density contoured at 1.5 σ). **(b)** The active-site zinc is properly coordinated (black dashes) with H539, H543, D550, and the nucleophilic water (orange) that would normally be hydrogen bonded with the catalytic base (E540). Y552 is found in a conformation that points towards the active-site zinc (black $2F_o$-F_c density contoured at 1.5 σ). This conformation is seen in the metalloprotease astacin from *Astacus astacus* (*Fig. 2.3.5.2a*), where it is found closer to the active-site zinc as a ligand.[101] **(c)** Y552 is primarily found in a conformation that points away from the active-site zinc in the wild type holoenzyme (*Fig. 2.3.4.2b*), however like the active-site zinc, it is relatively disordered with supporting $2F_o$-F_c density (black) only found below 1.2 σ (0.8 σ depicted).

In stark contrast with the zinc-supplemented wild type enzyme, the active site of the zinc-supplemented E540A structure is well ordered and reminiscent of the holoenzyme structures of many other metalloproteases (*Fig. 2.3.5.2ab*). The active-site zinc is coordinated by the side chains of residues H539, H543, and D550 and three water molecules in an octahedral configuration (*Fig. 2.3.6.2ab*). This configuration is also seen in the holoenzyme structure of the human matrix metalloprotease-12 (*Fig. 2.3.5.2b*).[99] None of these residues have alternate conformations and have well-defined density to support their conformation, unlike D550 in the zinc-supplemented structure (*Fig. 2.3.5.4*).

There are also differences in the conformation of Y552 between the zinc-supplemented wild type and E540A structures. Like the active-site zinc in the wild type structure, Y552 is disordered and points away from the zinc ion (*Fig. 2.3.6.2c*). However, the mutation locks Y552 into a single conformation that points towards the zinc (*Fig. 2.3.6.2b*). This tyrosine is located on the loop that bridges the two metal-binding sites (DE**Y**) and makes up part of the conserved sequence of the M64 protease family (*Fig. 2.3.4.2b*). A structurally homologous tyrosine is also seen in the astacin subfamily of metzincins (*Fig. 2.3.5.2a*) where it serves as an additional ligand to the active-site zinc and is hypothesized to stabilize the scissile carbonyl in binding and during

catalysis.[101] In the E540A structure, Y552 is too far away from the active-site zinc to interact but it may play an essential role in substrate selectivity as it may form part of the enzyme's subsites for substrate binding.

2.3.7 – Hypothetical Mechanisms of Substrate Recognition in the Protease Domain

The fact that an order-to-disorder transition is seen between the zinc-supplemented wild type and E540A structures suggests that the inherent disorder around the active-site zinc seen in the wild type active site may be biologically relevant. A working hypothesis suggests that in the absence of stabilizing interactions, the active site and its subsites are malformed and thus cannot bind an incoming peptide substrate. Following this logic, the ordered mutant structure may be a proxy for the conformation of the enzyme-substrate complex, simulating the structure of the protein as if it was bound in regions distal to the active site to stabilize its conformation.

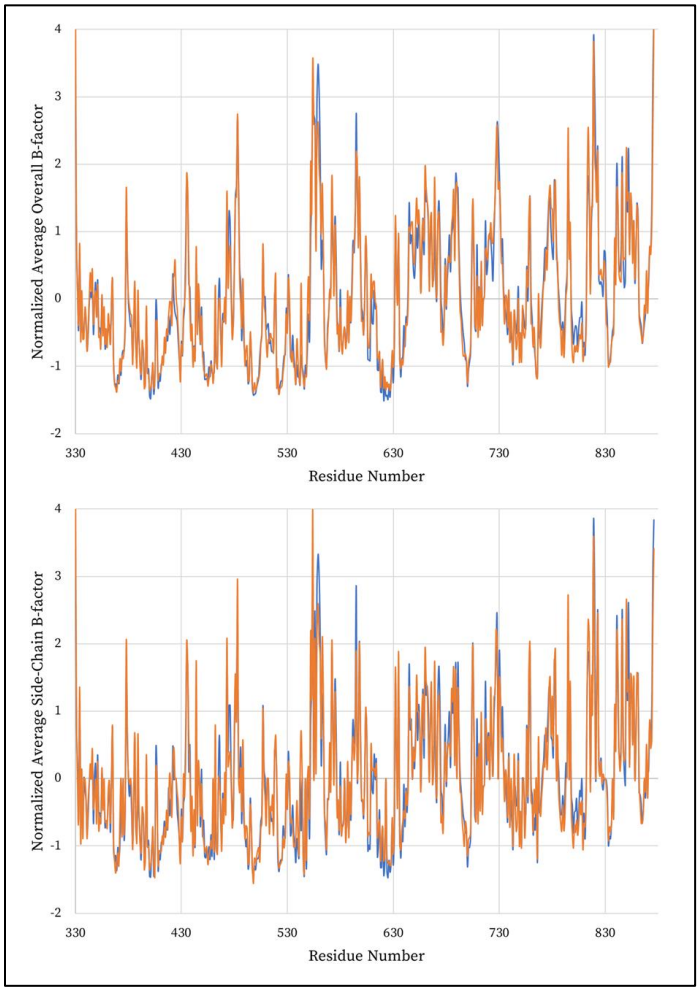


Figure 2.3.7.1 – Normalized B-factor Comparisons Between the Wild Type and E540A Protease Domain Structures. The normalized average overall **(top)** and side-chain **(bottom)** B-factors for each residue in the wild type (orange) and E540A (blue) MD+CTD1 structures were compared to look for potential differences in their ground state. Although there is more variability in the normalized side-chain B-factors, there are no noticeable changes that would indicate dynamic differences. B-factors were extracted with the Baverage program in the CCP4i suite[75] and normalized assuming a Gaussian distribution.[104] Normalized B-factors (B') were calculated for each i^{th} residue as $B'(i) = [B(i) - \mu]/\sigma$ where μ is the sample mean and σ is the sample standard deviation over all B-factors.

This working model implies that the wild type structure should have higher per-residue B-factors compared to the E540A mutant. However, the normalized B-factors[104] for each residue and its side chain are nearly identical between the two structures (*Fig. 2.3.7.1*), meaning that the E540A mutant may not be a good representation of the enzyme's substrate-ready state or that the working model is incorrect. There is also the possibility that these dynamics are quenched or restricted in the context of the crystal, which can be seen when comparing crystal structures to models determined by protein NMR,[105] or by contaminating non-zinc metals found in the active-site pocket (*Fig. 2.3.6.1c*).

Despite not knowing what conformation the E540A structure biochemically represents, the ordered geometry of its active site allowed for the rigid-body modelling of the IgA hinge peptide into peptide-binding groove to discern putative enzyme subsites that stabilize the substrate for cleavage. This was guided by general information known about metzincins (*Section 2.3.4*), characteristics of how they bind their substrate, as well as previously solved structures of metzincins with peptides of inhibitors. This peptide-bound model was then used to dock the partial and full molecules of IgA1 and IgA2m(1) onto the protease domain to make hypotheses for regions of the protease that may be involved in substrate recognition.

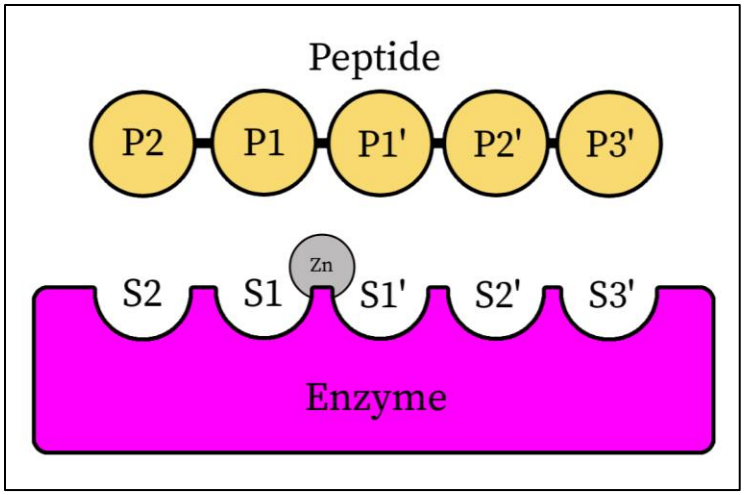


Figure 2.3.7.2 – Protease and Peptide Substrate Nomenclature. The nomenclature for describing peptide (P) and enzyme subsite (S) positions is simplistically depicted in a "lock-and-key" model.[106] The peptide is cleaved at the location of the active-site zinc. Residues and subsites C-terminal to the cleavage site are denoted with prime.

The peptide and subsite nomenclature that will be used in the following discussion is that from Schechter and Berger (*Fig. 2.3.7.2*).[106] Residues in the substrate are denoted "P" and subsites on the enzyme that interact with these residues are denoted "S". [106] They are also numbered starting from one at the cleavage site and increase when moving away from the scissile bond.[106] Residues and subsites that are C-terminal to the scissile bond are denoted with prime while those N-terminal to the bond do not have any additional nomenclature.[106]

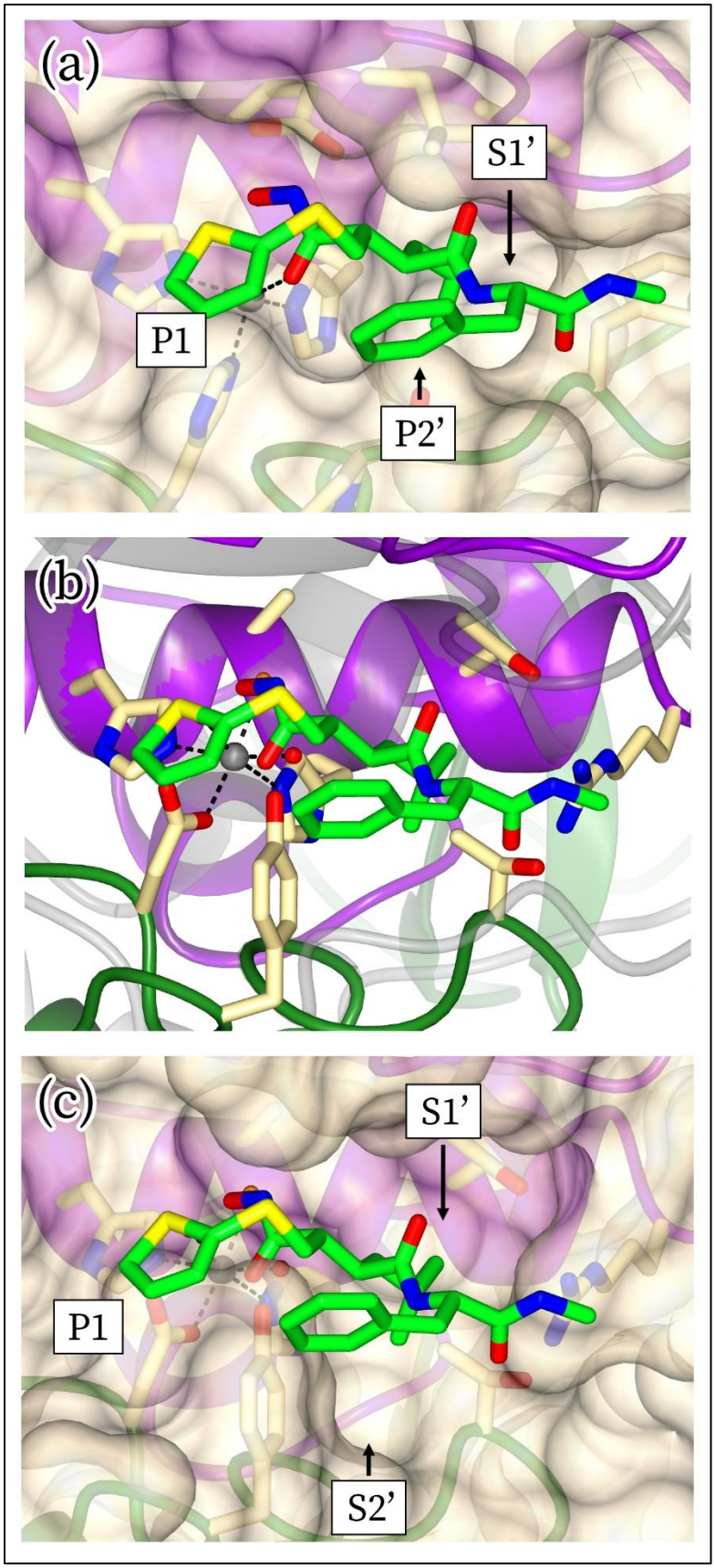


Figure 2.3.7.3 – Hypothesized Subsites in the Active-Site Groove of the Protease Domain. (a) The structure of human matrix metalloprotease-8 (MMP-8; purple for its N-terminal lobe and green for its C-terminal lobe) in complex with a peptidomimetic inhibitor (shown as cylinders in green; PDB 1MMB; no publicly available density) was used as a reference for rigid-body peptide docking.[107] The inhibitor's P1-analogue carbonyl coordinates with the active-site zinc and is stabilized by interactions in the enzyme's S1' and S2' pockets. **(b)** The subsites for the *T. ramosa* protease domain were inferred based on the inhibitor pose when superimposed with the E540A MD+CTD1 crystal structure. The P1-analogue carbonyl and hydroxyl groups superimpose onto two of the zinc-coordinating water molecules. The MMP-8 structure is shown in gray. **(c)** The S1' and S2' pockets for the protease domain were hypothesized based on the MMP-8-inhibitor structure.

A peptide with the IgA1 cleavage-site sequence P|VPST ("|" denoting the *T. ramosa* IgAP cut site; *Fig. 1.4.1b*) was modelled with AlphaFold2[15,16] and manually fit into the active-site cleft with Chimera.[103] The active-site helix (HEXXH motif) was used as the reference for the superimposition of the E540A MD+CTD1 structure with a structure of the human matrix metalloprotease-8 in complex with a peptidomimetic hydroxamate inhibitor (MMP-8; *Fig. 2.3.7.3a*).[107] Subsites on the E540A MD+CTD1 structure were assigned based on structural homology to the subsites of MMP-8 and the pose of the hydroxamate inhibitor (*Fig. 2.3.7.3bc*).[107] The AlphaFold2 peptide was rigid-body docked into the subsites, with a focus on optimizing the fit in the S1' and S2' pockets (*Fig. 2.3.7.3c*). It was important to ensure that the peptide bound antiparallel to the last strand of the beta sheet found in the N-terminal lobe (*Fig. 2.3.4.1c*) and that the carbonyl oxygen of the scissile peptide bond displaced one of the water molecules coordinating the active-site zinc.

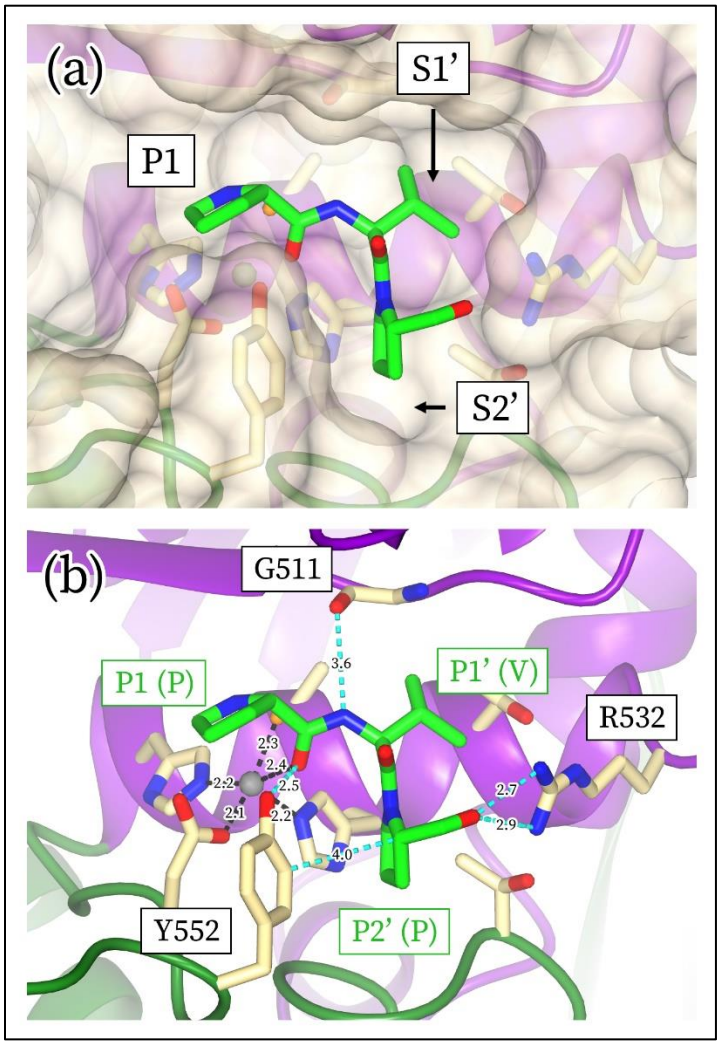


Figure 2.3.7.4 – Working Model of the Protease-Peptide Complex. (a) The rigid-body-docked model of the IgA hinge peptide with the protease domain depicts the peptide in a kinked orientation, anchored in place by well-formed S1' and S2' pockets. The S2' pocket is particularly deep and is primarily comprised of the aromatic surface of Y552. **(b)** The scissile carbonyl of P1 replaces one of the water molecules that previously coordinated the active-site zinc (black dashes) and is in close proximity to the nucleophilic water molecule (orange) in preparation for attack. Several enthalpic interactions (cyan dashes) are proposed to occur to stabilize the peptide in the active-site groove. These include an anti-parallel beta-strand hydrogen bond with G511, a proline-π interaction with Y552, and hydrogen bonds between Y552 and the P1 carbonyl and R532 and the P2' carbonyl. Bond and interaction distances are shown in Å.

The resultant model suggests that the IgA peptide binds to well-formed S1' and S2' pockets in the E540A protease domain mutant (*Fig. 2.3.7.4a*). The S1' pocket is defined by the side chains of T536 and R532 as well as the peptide backbone of residues 509-511 in the adjacent anti-parallel beta strand. The S2' pocket is defined by the side chains of H539, T611, and Y552 as well as the backbone of the end of the Met-turn (residues 608-609). This working model suggests that the peptide is stabilized by several enthalpic interactions that stabilize the P1, P1', P2' residues (*Fig. 2.3.7.4b*). The backbone carbonyl of P1 is stabilized by the active-site zinc and the hydroxyl group of Y552. The alpha-amino group of P1' makes a hydrogen bond to the backbone carbonyl of G511 on the adjacent anti-parallel beta strand. This type of interaction is seen in the MMP-8-inhibitor structure and is a common strategy for how metzincins stabilize their substrates.[95,107] The slightly positive hydrogen atom attached to the P2' alpha carbon is hypothesized to make a proline-π interaction with the slightly negative surface of Y552[108] and its backbone carbonyl hydrogen bonds to the side chain of R532. P3' and P4' residues do not make any contacts with the protease domain in this working model. This model may also explain why the protease is able to cleave both IgA1 and IgA2m(1) as the active-site groove only recognizes the side chains of P1'-P2', which are conserved between the two IgA molecules (*Fig. 1.4.1b*), and suggests that the O-linked glycans on the C-terminal portion of the IgA1 hinge are not involved in substrate binding to the protease domain.

Y552 plays a critical role in shaping the S2' pocket and stabilizing the P2' proline (*Fig. 2.3.7.3a*). In line with the working hypothesis, in the absence of distal stabilizing interactions with IgA, the active site is unable to bind the hinge due to disorder in the active-site zinc, its coordinating residues, and the zinc's water ligands. Y552 is also disordered and is mainly in a conformation that points away from the active-site zinc (*Fig. 2.3.6.2c*). Under this hypothesis, not only does this disorder make it entropically unfavorable for the peptide to bind but the conformation of Y552 blocks where the S2' subsite serves as an additional steric barrier to entry. Y552 is thus hypothesized to be a key residue in substrate gating and recognition. Its importance in the substrate recognition is supported by the fact that it is part of the M64 consensus sequence.

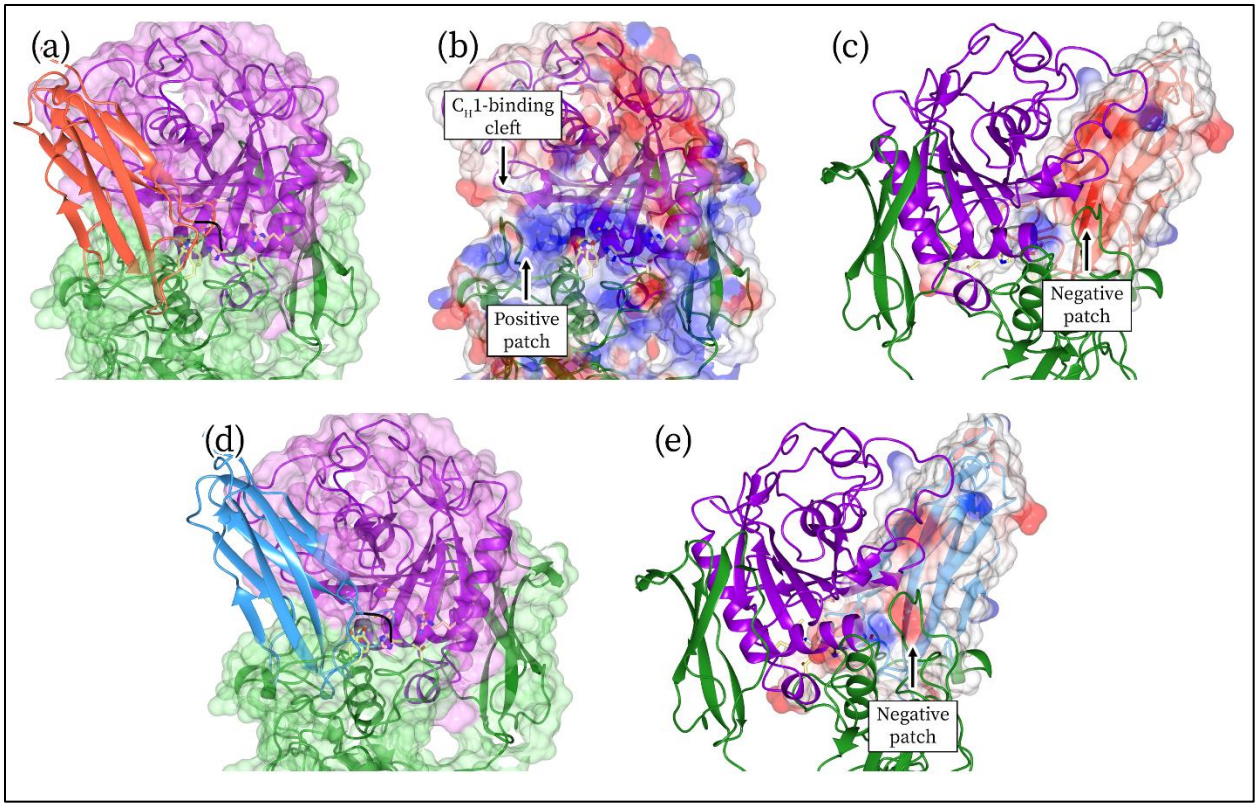


Figure 2.3.7.5 – Working Model of the IgA C_H1-Protease Domain Complex. (a) The AlphaFold2 model of the IgA1 C_H1 domain (dark orange) with part of the hinge (black) was superimposed onto the peptide-bound model (*Fig. 2.3.7.4*). The N- (purple) and C-terminal (green) lobes of the protease domain are shown as surfaces. **(b)** Although the IgA1 C_H1 domain clashes with parts of both lobes in this rigid-body-fit model, the model suggests that the protease domain contains a C_H1-binding cleft on the non-prime side of the active-site groove. This cleft contains a positively charged patch (blue surface), that extends from the active site. **(c)** This patch is nearby a negatively charged surface (red surface) on the C_H1 domain and may make favorable electrostatic interactions in a more optimized model. **(d)** The IgA2m(1) C_H1 domain (light blue) is also hypothesized to bind in the same cleft and **(e)** contains a similar negatively charged surface to IgA1. ***c*** and ***e*** are rotated 180° horizontally from ***a*** and ***d***.

AlphaFold2 was used to generate predictive models of the IgA1 and IgA2m(1) heavy chain C_H1 domain (*Fig. 1.3.1*) linked with the hinge peptide (PVP)[15,16] as current structures of these molecules are based on solution scattering data which do not contain any side chain information to model the hinge.[109,110] These structures were superimposed with the protease domain using the peptide-bound model as reference. The resultant model of the IgA1 and IgA2m(1) C_H1-protease complexes suggest that the C_H1 domains may bind in a cleft shared by the N- and C-terminal lobes of the protease domain (*Fig. 2.3.7.5ad*) that extends from the non-prime side of the active site (*Fig. 2.3.7.5b*). This cleft is positively charged and may make favorable electrostatic interactions with nearby negatively charged surfaces on the C_H1 domain (*Fig. 2.3.7.5ce*).

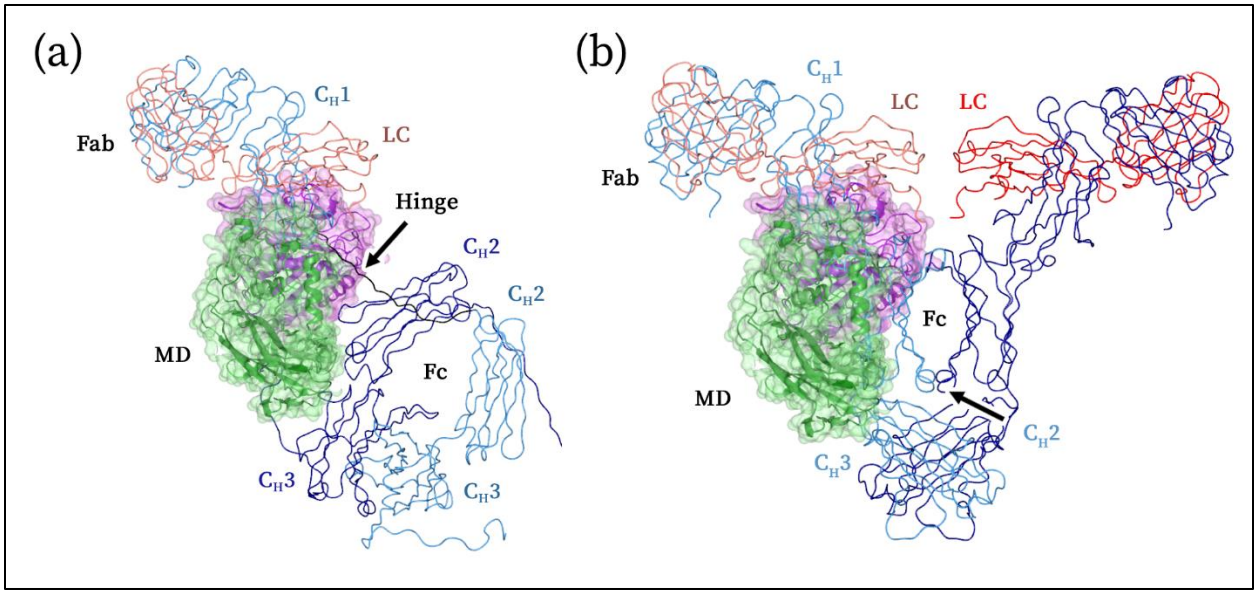


Figure 2.3.7.6 – Working Model of the IgA-Protease Domain Complex. The solution-scattering structures of IgA1 (PDB 1IGA)[110] **(a)** and IgA2m(1) (PDB 1R70)[109] **(b)** were superimposed onto the C_H1-bound protease domain model (*Fig. 2.3.7.4*). The N- (purple) and C-terminal (green) lobes of the protease domain are shown as surfaces. **(a)** The heavy chains of IgA1 are depicted in dark and light blue; the light blue heavy chain is modelled with its hinge (black) in the active site of the protease. **(b)** The heavy chains of IgA2m(1) are depicted in dark and light blue; the light blue heavy chain is modelled with its hinge (not visible) in the active site of the protease.

Full-molecule models of the substrate-protease complex were generated by superimposing the solution structures of IgA1[110] and IgA2m(1)[109] onto the C_H1-complexed models (*Fig. 2.3.7.6*). These models support the hypothesized binding cleft of the IgA C_H1 domains as the light chains (LC) do not sterically clash with the protease domain. Unfortunately, due to the inflexible and crude nature of these rigid-body-fit models, there are a significant number of clashes between the IgA1 and IgA2m(1) C_H1 and protease domains (*Fig. 2.3.7.5ad*). There are also clashes between the IgA2m(1) C_H2 and protease domains (*Fig. 2.3.7.6b*). Because the hinge cut site is directly next to the IgA Fab, the C-terminal region of the Fab overlaps with the non-prime side of the active site. A particular severe clash involves the P2 cysteine residue as it forms an inter-strand disulfide bond that cannot be sterically accounted for on the non-prime side of the active site. There are also less severe clashes on the surfaces of both the N- and C-terminal lobes. As such, it is important to take these models with a grain of salt. Given the disordered active site in the wild type protease domain, it is likely the case that IgA is recognized distal to the active site, perhaps at the proposed C_H1-binding cleft, which may result in structural changes that order the active site and accommodate for the otherwise severe steric clashes in the rigid-body models.

Initial biochemical characterization of the *T. ramosa* IgAP found that the protease was able to cleave IgA2m(1) but not IgA2m(2) (*Section 2.1.5*).[14] This was hypothesized to be a result of the differences in hinge primary sequence between the two allotypes, with IgA2m(2) having an arginine at P1 compared to a P1 proline in the A2m(1) allotype.[14] In the absence of structural data, the large side-chain of a P1 arginine was thought to prevent the hinge from binding into the protease active site and hence be the reason for protease resistance.[14] A subsequent study assessed the ability of the *T. ramosa* IgAP to cleave ape IgA and revealed that it was able to cleave gorilla IgA1, which has an arginine at P1 (R|VP).[111] These researchers suggested that the *T. ramosa* IgAP may be recognizing a P2 cysteine conserved amongst all ape IgAs tested.[111] However, as discussed above, this cysteine forms a inter-strand disulfide bond in the context of the IgA Fab and hence its conformation may be variable with respect to each species' Fab.

The previously discussed working model for substrate recognition and binding provides a plausible explanation for these observations. The peptide-bound model shown in *Fig. 2.3.7.4* has the P1 residue on the surface of the active site with its side chain pointed away from the active-site zinc. As such, in the case where P1 is arginine, there would be no steric restrictions that prevent it from binding to the cleft. In addition, the main stabilizing contributions proposed by the peptide-bound model are at the S2' subsite, which binds the P2' proline conserved amongst ape IgA1s.[111] The ability of the *T. ramosa* IgAP to cleave human IgA2m(1) but not IgA2m(2) can be attributed to differences in IgA structure distal to the hinge and how those interactions order the active site for substrate binding.

2.3.8 – The Protease Domain and Its Relationship to the M64 Protease Family

Members of the family of M64 proteases are identified based on the presence of the extended metalloprotease and metzincin sequence as well as a second zinc-binding pocket coordinated by one glutamic acid and three cysteines (*Section 2.3.4*).[66,95] Currently, there are only two M64 protease structures in the Protein Data Bank: one from *Bacteroides ovatus* (PDB 3P1V) and another from *B. fragilis* (PDB 4DF9) that are 74% identical and 88% similar to one another. These structures were solved by the now-dissolved Joint Center for Structural Genomics and thus do not have any associated publications or supplementary information.

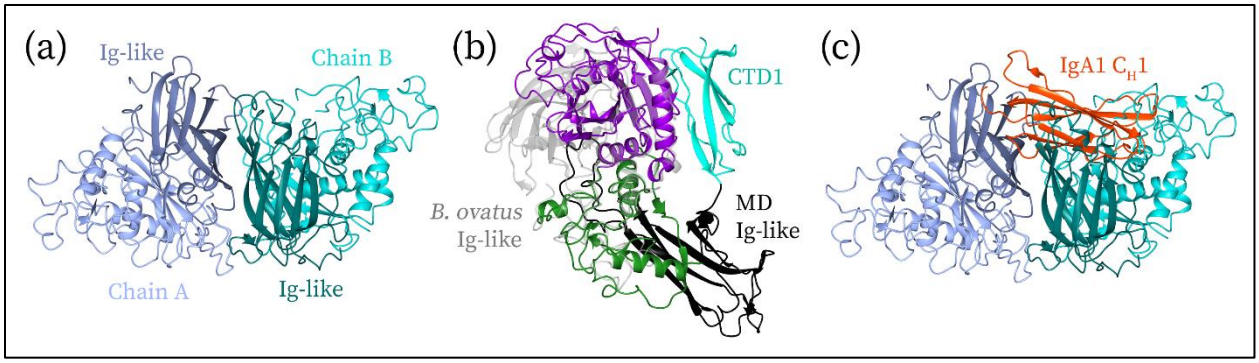


Figure 2.3.8.1 – Structural Comparisons Between Members of the M64 Protease Family. (a) The *Bacteroides ovatus* M64 protease is found as a dimer (PDB 3P1V; light blue and turquoise) that is stabilized through interactions with each monomer's N-terminal Ig-like fold (dark blue and blue green). **(b)** A similar Ig-like fold is seen in the *T. ramosa* protease domain but is found on the domain's C-terminus (black), adjacent to CTD1 (cyan). This is opposed to the Ig-like fold being found N-terminal to the *B. ovatus* M64 domain (grey). **(c)** This N-terminal Ig-like fold sterically clashes with the working model of how the IgA1 C_H1 domain (red) binds to the enzyme.

Both *B. ovatus* and *B. fragilis* M64 proteases crystallize as dimers, which are stabilized through an N-terminal beta-sheet sandwich on each monomer (*Fig. 2.3.8.1a*; commonly referred to as an Ig-like fold). A similar Ig-like fold is also found in the *T. ramosa* protease domain, but it is located on the C-terminal end of the MD, immediately N-terminal to CTD1 (*Fig. 2.3.8.1b*). The *B. ovatus* (UniProt Accession #A7LS78) and *B. fragilis* (UniProt Accession #Q5L9L3) M64 proteases also do not contain any additional C-terminal domains; their primary sequences only contain a signal peptide, the N-terminal Ig-like fold, and the M64 protease domain.

Substrates for these other M64 proteases have yet to be identified with experimental data. All M64 proteases have been bioinformatically categorized as being IgA proteases based on their structural and sequence similarity to the *T. ramosa* IgAP.[19,66] Despite this, preliminary data on the recombinantly expressed and purified *B. ovatus* M64 protease domain showed that it does not cleave IgA under the same *in vitro* assay conditions as the *T. ramosa* IgAP (data not shown; assay conditions summarized in *Chapter 4*). In light of the working model of the IgA C_H1 domain in complex with the protease domain (*Fig. 2.3.7.5*), this inability to cleave IgA1 may be due to the presence of the N-terminal Ig-like fold as it sterically occludes the hypothesized C_H1-binding cleft in the context of the dimer (*Fig. 2.3.8.1c*). In addition, unlike the *T. ramosa* IgAP, the active sites in both of these other M64 holoenzymes are well ordered, suggesting that the mechanisms that are hypothesized to be involved in IgA binding and recognition are not present in these other proteases. This gives more credence to the idea that they are not IgAPs.

As such, it is important to take caution when making conclusions based on bioinformatic annotations of proteins, and in particular proteases. Many proteins with structural and sequence similarity to the *T. ramosa* protease domain have likely been incorrectly annotated as IgAPs despite the lack of experimental evidence confirming activity. The *B. ovatus* M64 protease is just one of these examples. This assumption has already led to the claim of this protease having IgAP activity in a review article on metzincins.[95] There is also the possibility that the *B. ovatus* protease can cleave IgA1 but only under fastidious experimental conditions, like in the presence of co-factors or at a specific pH. Unfortunately, it is difficult to obtain definitive evidence for a negative hypothesis. Regardless, the function of these proteases should not be assumed until substrate(s) are found and experimentally verified.

2.4 – Chapter Conclusions and Future Directions

This chapter focused on understanding the *T. ramosa* IgAP from a high-resolution structural-biology perspective. The enzyme was broken into six domains, two of which were structurally characterized using X-ray crystallography. These data revealed that the domains were true domains in that they can fold independently and have separate functions. Although the function of the NTD remains unknown, structural analyses suggest that a beta-hairpin structure may be an important ligand-binding surface that can be the focus of future investigations. The inherent disorder seen in the protease domain active site gave insight into the substrate-recognition mechanisms of the protease and led to a working model in which protease-substrate interactions distal to the active site may be required to order the active site into a catalytically competent conformation. The biological role of the second zinc-binding site in the protease domain still remains a mystery and should the focus of future studies on M64 proteases. Insights into the putative substrate-binding poses of the hinge peptide and IgA strongly suggest that the bioinformatic categorization of M64 proteases may be misleading, especially as the M64 protease from *Bacteroides ovatus* showed no IgAP activity. As such, the structural differences between the *B. ovatus* and *T. ramosa* proteases is another line of questioning for future work.

2.5 – Chapter References

1. Mistry, D. & Stockley, R. A. IgA1 protease. *Int J Biochem Cell Biol* **38**, 1244–1248 (2006).

2. Mulks, M. H., Kornfeld, S. J., Frangione, B. & Plaut, A. G. Relationship Between the Specificity of IgA Proteases and Serotypes in Haemophilus influenzae. *J Infect Dis* **146**, 266–274 (1982).

3. Kilian, M., Mestecky, J. & Schrohenloher, R. E. Pathogenic Species of the Genus Haemophilus and Streptococcus pneumoniae Produce Immunoglobulin A1 Protease. *Infect Immun* **26**, 143–149 (1979).

4. Mulks, M. H. & Plaut, A. G. IgA Protease Production as a Characteristic Distinguishing Pathogenic from Harmless Neisseriaceae. *N Engl J Med* **299**, 973–976 (1978).

5. Burton, J., Wood, S. G., Lynch, M. & Plaut, A. G. Substrate Analogue Inhibitors of the IgA1 Proteinases from Neisseria Gonorrhoeae. *J Med Chem* **31**, 1647–1651 (1988).

6. Wood, S. G., Lynch, M., Plaut, A. G. & Burton, J. Tetrapeptide Inhibitors of the IgA1 Proteinases from Type I Neisseria gonorrhoeae. *J Med Chem* **32**, 2407–2411 (1989).

7. Garner, A. L., Fullagar, J. L., Day, J. A., Cohen, S. M. & Janda, K. D. Development of a high-throughput screen and its use in the discovery of Streptococcus pneumoniae immunoglobulin A1 protease inhibitors. *J Am Chem Soc* **135**, 10014–10017 (2013).

8. Choudary, S. K., Qiu, J., Plaut, A. G. & Kritzer, J. A. Versatile Substrates and Probes for IgA1 Protease Activity. *ChemBioChem* **14**, 2007–2012 (2013).

9. Shehaj, L., Choudary, S. K., Makwana, K. M., Gallo, M. C., Murphy, T. F. & Kritzer, J. A. Small-Molecule Inhibitors of Haemophilus influenzae IgA1 Protease. *ACS Infect Dis* **5**, 1129–1138 (2019).

10. Bachovchin, W. W., Plaut, A. G., Flentke, G. R., Lynch, M. & Kettner, C. A. Inhibition of IgA1 proteinases from Neisseria gonorrhoeae and Hemophilus influenzae by peptide prolyl boronic acids. *Journal of Biological Chemistry* **265**, 3738–3743 (1990).

11. Wang, Z., Rahkola, J., Redzic, J. S., Chi, Y. C., Tran, N., Holyoak, T., Zheng, H., Janoff, E. & Eisenmesser, E. Mechanism and inhibition of Streptococcus pneumoniae IgA1 protease. *Nat Commun* **11**, (2020).

12. Redzic, J. S., Rahkola, J., Tran, N., Holyoak, T., Lee, E., Martín-Galiano, A. J., Meyer, N., Zheng, H. & Eisenmesser, E. A substrate-induced gating mechanism is conserved among Gram-positive IgA1 metalloproteases. *Commun Biol* **5**, (2022).

13. Johnson, T. A., Qiu, J., Plaut, A. G. & Holyoak, T. Active-Site Gating Regulates Substrate Selectivity in a Chymotrypsin-Like Serine Protease. The Structure of Haemophilus influenzae Immunoglobulin A1 Protease. *J Mol Biol* **389**, 559–574 (2009).

14. Fujiyama, Y., Kobayashi, K., Senda, S., Benno, Y., Bamba, T. & Hosoda, S. A Novel IgA Protease from Clostridium sp. Capable of Cleaving IgA1 and IgA2 A2m(1) but not IgA2 A2m(2) allotype paraproteins. *The Journal of Immunology* **134**, 573–576 (1985).

15. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

16. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S. & Steinegger, M. ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679–682 (2022).

17. Mortensen, S. B. & Kilian, M. Purification and Characterization of an Immunoglobulin A1 Protease from Bacteroides melaninogenicus. *Infect Immun* **45**, 550–557 (1984).

18. Kilian, M., Mestecky, J., Kulhavy, R., Tomana, M. & Butler, W. T. IgA1 proteases from Haemophilus influenzae, Streptococcus pneumoniae, Neisseria meningitidis, and Streptococcus sanguis: comparative immunochemical studies. *The Journal of Immunology* **124**, 2596–2600 (1980).

19. Rawlings, N. D., Barrett, A. J. & Bateman, A. MEROPS: The database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res* **40**, (2012).

20. Rawlings, N. D. & Bateman, A. How to use the MEROPS database and website to help understand peptidase specificity. *Protein Science* **30**, 83–92 (2021).

21. Mulks, M. H., Kornfeld, S. J. & Plaut, a G. Specific proteolysis of human IgA by Streptococcus pneumoniae and Haemophilus influenzae. *J Infect Dis* (1980).

22. Arzese, A. & Botta, G. A. Comparative Studies of the IgA1 Protease Genes from Haemophilus influenzae, Neisseria gonorrhoeae, and Prevotella melaninogenica. *Clinical Infectious Diseases* **20**, S169–S171 (1995).

23. McGillivary, G., Smoot, L. M. & Actis, L. A. Characterization of the IgA1 protease from the Brazilian purpuric fever strain F3031 of Haemophilus influenzae biogroup aegyptius. *FEMS Microbiol Lett* **250**, 229–236 (2005).

24. Plaut, A. G. & Bachovchin, W. W. IgA-Specific Prolyl Endopeptidases: Serine Type. *Methods Enzymol* **244**, 137–151 (1994).

25. Poulsen, K., Reinholdt, J., Jespersgaard, C., Boye, K., Brown, T. A., Hauge, M. & Kilian, M. A Comprehensive Genetic Study of Streptococcal Immunoglobulin A1 Proteases: Evidence for Recombination within and between Species. *Infect Immun* **66**, 181–190 (1998).

26. Plaut, A. G., Gilbert, J. V, Artenstein, M. S. & Capra, J. D. Neisseria gonorrhoeae and Neisseria meningitidis: Extracellular Enzyme Cleaves Human Immunoglobulin A. *Science* **190**, 1103–1105 (1975).

27. Mistry, D. V. & Stockley, R. A. The cleavage specificity of an IgA1 protease from Haemophilus influenzae. *Virulence* **2**, (2011).

28. Henderson, I. R., Navarro-Garcia, F., Desvaux, M., Fernandez, R. C. & Ala'Aldeen, D. Type V Protein Secretion Pathway: the Autotransporter Story. *Microbiology and Molecular Biology Reviews* **68**, 692–744 (2004).

29. Grundy, F. J., Plaut, A. & Wright, A. Haemophilus influenzae Immunoglobulin A1 Protease Genes: Cloning by Plasmid Integration-Excision, Comparative Analyses, and Localization of Secretion Determinants. *J Bacteriol* **169**, 4442–4450 (1987).

30. Pohlner, J., Halter, R., Beyreuther, K. & Meyer, T. F. Gene Structure and Extracellular Secretion of Neisseria gonorrhoeae IgA Protease. *Nature* **325**, 458–462 (1987).

31. Natale, P., Brüser, T. & Driessen, A. J. M. Sec- and Tat-mediated protein secretion across the bacterial cytoplasmic membrane-Distinct translocases and mechanisms. *Biochim Biophys Acta* **1778**, 1735–1756 (2008).

32. Tran, N. Establishing Viable Systems for the Recombinant Production of Haemophilus influenzae Immunoglobulin A1 Protease. (2020).

33. Dautin, N. & Bernstein, H. D. Protein Secretion in Gram-Negative Bacteria via the Autotransporter Pathway. *Annu Rev Microbiol* **61**, 89–112 (2007).

34. Rojas-Lopez, M., Zorgani, M. A., Kelley, L. A., Bailly, X., Kajava, A. V., Henderson, I. R., Polticelli, F., Pizza, M., Rosini, R. & Desvaux, M. Identification of the Autochaperone Domain in the Type Va Secretion System (T5aSS): Prevalent Feature of Autotransporters with a β-Helical Passenger. *Front Microbiol* **8**, (2018).

35. Qiu, J., Hendrixson, D. R., Baker, E. N., Murphy, T. F., St. Geme, J. W. & Plaut, A. G. Human milk lactoferrin inactivates two putative colonization factors expressed by Haemophilus influenzae. *Proc Natl Acad Sci U S A* **95**, 12641–12646 (1998).

36. Plaut, A. G., Qiu, J. & St. Geme, J. W. Human lactoferrin proteolytic activity: analysis of the cleaved region in the IgA protease of Haemophilus influenzae. *Vaccine* **19**, S148-152 (2001).

37. Roussel-Jazédé, V., Arenas, J., Langereis, J. D., Tommassen, J. & van Ulsen, P. Variable processing of the IgA protease autotransporter at the cell surface of Neisseria meningitidis. *Microbiology (N Y)* **160**, 2421–2431 (2014).

38. Junker, M., Schuster, C. C., Mcdonnell, A. V, Sorg, K. A., Finn, M. C., Berger, B. & Clark, P. L. Pertactin β-helix folding mechanism suggests common themes for the secretion and folding of autotransporter proteins. *Proc Natl Acad Sci* **103**, 4918–4923 (2006).

39. Poulsen, K., Reinholdt, J. & Kilian, M. Characterization of the Streptococcus pneumoniae Immunoglobulin A1 Protease Gene (iga) and Its Translation Product. *Infect Immun* **64**, 3957–3966 (1996).

40. Schneewind, O. & Missiakas, D. Sec-secretion and sortase-mediated anchoring of proteins in Gram-positive bacteria. *Biochim Biophys Acta* **1843**, 1687–1697 (2014).

41. Paukovich, N., Redzic, J. S., Chi, Y. C., Rahkola, J. T., Issaian, A., Blue, A., Hansen, K. C., Janoff, E. N. & Eisenmesser, E. Z. Streptococcus pneumoniae G5 domains bind different ligands. *Protein Science* **28**, 1797–1805 (2019).

42. Bateman, A., Holden, M. T. G. & Yeats, C. The G5 domain: a potential N-acetylglucosamine recognition domain involved in biofilm formation. *Bioinformatics* **21**, 1301–1303 (2005).

43. Roche, F. M., Meehan, M. & Foster, T. J. The Staphylococcus aureus surface protein SasG and its homologues promote bacterial adherence to human desquamated nasal epithelial cells. *Microbiology* **149**, 2759–2767 (2003).

44. Bender, M. H. & Weiser, J. N. The atypical amino-terminal LPNTG-containing domain of the pneumococcal human IgA1-specific protease is required for proper enzyme localization and function. *Mol Microbiol* **61**, 526–543 (2006).

45. Yamaguchi, M., Nakata, M., Sumioka, R., Hirose, Y., Wada, S., Akeda, Y., Sumitomo, T. & Kawabata, S. Zinc metalloproteinase ZmpC suppresses experimental pneumococcal meningitis by inhibiting bacterial invasion of central nervous systems. *Virulence* **8**, 1516–1524 (2017).

46. Chi, Y. C., Rahkola, J. T., Kendrick, A. A., Holliday, M. J., Paukovich, N., Roberts, T. S., Janoff, E. N. & Eisenmesser, E. Z. Streptococcus pneumoniae IgA1 protease: A metalloprotease that can catalyze in a split manner in vitro. *Protein Science* **26**, 600–610 (2017).

47. Reinholdt, J. & Kilian, M. Comparative Analysis of Immunoglobulin A1 Protease Activity among Bacteria Representing Different Genera, Species, and Strains. *Infect Immun* **65**, 4452–4459 (1997).

48. Blue, C. E., Paterson, G. K., Kerr, A. R., Bergé, M., Claverys, J. P. & Mitchell, T. J. ZmpB, a Novel Virulence Factor of Streptococcus pneumoniae that Induces Tumor Necrosis Factor Alpha Production in the Respiratory Tract. *Infect Immun* **71**, 4925–4935 (2003).

49. Bek-Thomsen, M., Poulsen, K. & Kilian, M. Occurrence and Evolution of the Paralogous Zinc Metalloproteases IgA1 protease, ZmpB, ZmpC, and ZmpD in Streptococcus pneumoniae and Related Commensal Species. *mBio* **3**, (2012).

50. Schulz, E. C. & Ficner, R. Knitting and snipping: Chaperones in β-helix folding. *Curr Opin Struct Biol* **21**, 232–239 (2011).

51. Vallee, B. L. & Auld, D. S. Active-site zinc ligands and activated H2O of zinc enzymes. *Proc. Natl. Acad. Sci. U S A* **87**, 220–224 (1990).

52. Holdeman, L. V, Cato, E. P. & Moore, W. E. C. Clostridium rumosum (Vuillemin) comb. nov.: Emended Description and Proposed Neotype Strain. *Int J Syst Bacteriol* **21**, 35–39 (1971).

53. Lawson, P. A., Perez, L. S. & Sankaranarayanan, K. Reclassification of Clostridium cocleatum, Clostridium ramosum, Clostridium spiroforme and Clostridium saccharogumia as Thomasclavelia cocleata gen. nov., comb. nov., Thomasclavelia ramosa comb. nov., gen. nov., Thomasclavelia spiroformis comb. nov. and Thomasclavelia saccharogumia comb. nov. *Int J Syst Evol Microbiol* **73**, (2023).

54. Senda, S., Fujiyama, Y., Ushijima, T., Hodohara, K., Bamba, T., Hosoda, S. & Kobayashi, K. Clostridium ramosum, an IgA Protease-Producing Species and Its Ecology in the Human Intestinal Tract. *Microbiology and Immunity* **29**, 1019–1028 (1985).

55. Bao, D., Xu, X., Wang, Y., Zhu, F., Wu, Y. & Li, H. Clostridium ramosum Bacteremia in an Immunocompetent Patient with SARS-CoV-2 Infection: A Case Report. *Infect Drug Resist* **16**, 4455–4461 (2023).

56. Legaria, M. C., García, S. D., Tudanca, V., Barberis, C., Cipolla, L., Cornet, L., Famiglietti, A. M. R., Stecher, D. & Vay, C. A. Clostridium ramosum rapidly identified by MALDI-TOF MS. A rare gram-variable agent of bacteraemia. *Access Microbiol* **2**, (2020).

57. Forrester, J. D. & Spain, D. A. Clostridium ramosum Bacteremia: Case Report and Literature Review. *Surg Infect* **15**, 343–346 (2014).

58. Brook, I. Clostridial infection in children. *J Med Microbiol* **42**, 78–82 (1995).

59. Shinzato, T., Yonaha, T., Oshiro, Y. & Ishiki, H. Clostridium ramosum bacteremia: A case series at a general acute care hospital. *Journal of Infection and Chemotherapy* **29**, 78–81 (2023).

60. Male, C. J. Immunoglobulin A1 Protease Production by Haemophilus influenzae and Streptococcus pneumoniae. *Infect Immun* **26**, 254–261 (1979).

61. Plaut, A. G., Genco, R. J. & Tomasi, T. B. Isolation of an Enzyme from Streptococcus sanguis Which Specifically Cleaves IgA. *J Immunol* **113**, 289–291 (1974).

62. Mehta, S. K., Plaut, A. G., Calvanico, N. J. & Tomasi, T. B. Human Immunoglobulin A: Production of an Fc Fragment by an Enteric Microbial Proteolytic Enzyme. *J Immunol* **111**, 1274–1276 (1973).

63. Brandtzaeg, P., Kett, K. & Rognum, T. O. Subclass Distribution of IgG- and IgA-Producing Cells in Secretory Tissues and Alterations Related to Gut Diseases. in *Recent Advances in Mucosal Immunology* (eds. Mestecky, J., McGhee, J. R., Bienenstock, J. & Ogra, P. L.) 321–333 (1987).

64. Woof, J. M. & Mestecky, J. Mucosal immunoglobulins. *Immunol Rev* **206**, 64–82 (2005).

65. Kobayashi, K., Hagiwara, K., Kondoh, H. & Fujiyama, Y. Resistance of Normal Serum IgA and Secretory IgA to Bacterial IgA Proteases: Evidence for the Presence of Enzyme-Neutralizing Antibodies in Both Serum and Secretory IgA, and Also in Serum IgG. *Microbiol Immunol* **31**, 1097–1106 (1987).

66. Kosowska, K., Reinholdt, J., Rasmussen, L. K., Sabat, A., Potempa, J., Kilian, M. & Poulsen, K. The Clostridium ramosum IgA Proteinase Represents a Novel Type of Metalloendopeptidase. *J Biol Chem* **277**, 11987–11994 (2002).

67. Xie, X., Li, J., Liu, P., Wang, M., Gao, L., Wan, F., Lv, J., Zhang, H. & Jin, J. Chimeric Fusion between Clostridium Ramosum IgA Protease and IgG Fc Provides Long-Lasting Clearance of IgA Deposits in Mouse Models of IgA Nephropathy. *Journal of the American Society of Nephrology* **33**, 918–935 (2022).

68. Potempa, J. & Poulsen, K. IgA protease of Clostridium ramosum. in *Handbook of Proteolytic Enzymes* (eds. Barrett, A. J., Rawlings, N. D. & Woessner, J. F.) 759–762 (Academic Press, 2004).

69. Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expr Purif* **41**, 207–234 (2005).

70. Lau, Y.-T. K., Baytshtok, V., Howard, T. A., Fiala, B. M., Johnson, J. L. M., Carter, L. P., Baker, D., Lima, C. D. & Bahl, C. D. Discovery and engineering of enhanced SUMO protease enzymes. *J Biol Chem* **293**, 13224–13233 (2018).

71. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D. & Bairoch, A. Protein Analysis Tools on the ExPASy Server. in *The Proteomics Protocols Handbook* (ed. Walker, J. M.) 571–607 (Humana Press Inc., Totowa, 2005).

72. Chaikuad, A., Knapp, S. & von Delft, F. Defined PEG smears as an alternative approach to enhance the search for crystallization conditions and crystal-quality improvement in reduced screens. *Acta Crystallogr* **D71**, 1627–1639 (2015).

73. Fazio, V. J., Peat, T. S. & Newman, J. A drunken search in crystallization space. *Acta Crystallogr* **F70**, 1303–1311 (2014).

74. Winter, G., Waterman, D. G., Parkhurst, J. M., Brewster, A. S., Gildea, R. J., Gerstel, M., Fuentes-Montero, L., Vollmar, M., Michels-Clark, T., Young, I. D. *et al.* DIALS: Implementation and evaluation of a new integration package. *Acta Crystallogr* **D74**, 85–97 (2018).

75. Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr* **D67**, 235–242 (2011).

76. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr* **D69**, 1204–1214 (2013).

77. Vagin, A. & Teplyakov, A. MOLREP: an Automated Program for Molecular Replacement. *J Appl Crystallogr* **30**, 1022–1025 (1997).

78. Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H. & Adams, P. D. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr* **D68**, 352–367 (2012).

79. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr* **D66**, 486–501 (2010).

80. Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., Verma, V., Keedy, D. A., Hintze, B. J., Chen, V. B. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science* **27**, 293–315 (2018).

81. Ten Eyck, L. F. Crystallographic Fast Fourier Transforms. *Acta Cryst* **A29**, 183–191 (1973).

82. Read, R. J. & Schierbeek, A. J. A phased translation function. *J. Appl. Cryst* **21**, 490–495 (1988).

83. Immirzi, A. Crystallographic Computing Techniques. in (ed. Ahmed, F. R.) 399 (Munksgaard, 1966).

84. Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G. A., Bileschi, M. L., Bork, P., Bridge, A., Colwell, L. *et al.* InterPro in 2022. *Nucleic Acids Res* **51**, D418-427 (2023).

85. Krissinel, E. & Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J Mol Biol* **372**, 774–797 (2007).

86. Baek, M., Dimaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

87. Krishna, S. S., Majumdar, I. & Grishin, N. V. Structural classification of zinc fingers. *Nucleic Acids Res* **31**, 532–550 (2003).

88. Maier, R. J. & Benoit, S. L. Role of Nickel in Microbial Pathogenesis. *Inorganics (Basel)* **7**, (2019).

89. Gucwa, M., Lenkiewicz, J., Zheng, H., Cymborowski, M., Cooper, D. R., Murzyn, K. & Minor, W. CMM - An enhanced platform for interactive validation of metal binding sites. *Protein Science* **32**, (2023).

90. Baykov, I. K., Tikunov, A. Y., Babkin, I. V., Fedorets, V. A., Zhirakovskaia, E. V. & Tikunova, N. V. Tentaclins—A Novel Family of Phage Receptor-Binding Proteins That Can Be Hypermutated by DGR Systems. *Int J Mol Sci* **24**, (2023).

91. Mnich, M. E., van Dalen, R. & van Sorge, N. M. C-Type Lectin Receptors in Host Defense Against Bacterial Pathogens. *Front Cell Infect Microbiol* **10**, (2020).

92. Holm, L. DALI and the persistence of protein shape. *Protein Science* **29**, 128–140 (2020).

93. Holm, L. Using Dali for Protein Structure Comparison. in *Structural Bioinformatics: Methods and Protocols. Methods in Molecular Biology* (ed. Gáspári, Z.) vol. 2112 29–42 (Humana Press Inc., 2020).

94. Poget, S. F., Legge, G. B., Proctor, M. R., Butler, P. J. G., Bycroft, M. & Williams, R. L. The Structure of a Tunicate C-type Lectin from Polyandrocarpa misakiensis Complexed with. *J Mol Biol* **290**, 867–879 (1999).

95. Cerdà-Costa, N. & Gomis-Rüth, F. X. Architecture and function of metallopeptidase catalytic domains. *Protein Science* **23**, 123–144 (2014).

96. Waltersperger, S., Widmer, C., Wang, M. & Baumann, U. Crystal structure of archaemetzincin AmzA from Methanopyrus kandleri at 1.5 Å resolution. *Proteins* **78**, 2720–2723 (2010).

97. Graef, C., Schacherl, M., Waltersperger, S. & Baumann, U. Crystal Structures of Archaemetzincin Reveal a Moldable Substrate-Binding Site. *PLoS One* **7**, (2012).

98. Janda, I., Devedjiev, Y., Derewenda, U., Dauter, Z., Bielnicki, J., Cooper, D. R., Graf, P. C. F., Joachimiak, A., Jakob, U. & Derewenda, Z. S. The Crystal Structure of the Reduced, Zn2+-Bound Form of the B. subtilis Hsp33 Chaperone and Its Implications for the Activation Mechanism. *Structure* **12**, 1901–1907 (2004).

99. Bertini, I., Calderone, V., Fragai, M., Luchinat, C., Maletta, M. & Kwon, J. Y. Snapshots of the Reaction Mechanism of Matrix Metalloproteinases. *Angew Chem Int Ed* **45**, 7952–7955 (2006).

100. Gomis-Rüth, F.-X., Kress, L. F. & Bode, W. First structure of a snake venom metalloproteinase: a prototype for matrix metalloproteinases/collagenases. *EMBO J* **12**, 4151–4157 (1993).

101. Bode, W., Gomis-Rüth, F. X., Zwilling, R. & Stöcker, W. Structure of astacin and implications for activation of astacins and zinc-ligation of collagenases. *Nature* **358**, 164–167 (1992).

102. Handing, K. B., Niedzialkowska, E., Shabalin, I. G., Kuhn, M. L., Zheng, H. & Minor, W. Characterizing metal-binding sites in proteins with X-ray crystallography. *Nat Protoc* **13**, 1062–1090 (2018).

103. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. UCSF Chimera - A visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605–1612 (2004).

104. Carugo, O. & Argos, P. Correlation between side chain mobility and conformation in protein structures. *Protein Eng* **10**, 777–787 (1997).

105. Alderson, T. R. & Kay, L. E. NMR spectroscopy captures the essential role of dynamics in regulating biomolecular function. *Cell* **184**, 577–595 (2021).

106. Schechter, I. & Berger, A. On the Size of the Active Site in Proteases. I. Papain. *Biochem Biophys Res Commun* **27**, (1967).

107. Grams, F., Cri, M., Hinnes, L., Huxley, P., Pieper, J. M., Tschesche, H. & Bode, W. Structure Determination and Analysis of Human Neutrophil Collagenase Complexed with a Hydroxamate Inhibitor. *Biochemistry* **34**, 14012–14020 (1995).

108. Zondlo, N. J. Aromatic-Proline Interactions: Electronically Tunable CH/π Interactions. *Acc Chem Res* **46**, 1039–1049 (2013).

109. Furtado, P. B., Whitty, P. W., Robertson, A., Eaton, J. T., Almogren, A., Kerr, M. A., Woof, J. M. & Perkins, S. J. Solution Structure Determination of Monomeric Human IgA2 by X-ray and Neutron Scattering, Analytical Ultracentrifugation and Constrained Modelling: A Comparison with Monomeric Human IgA1. *J Mol Biol* **338**, 921–941 (2004).

110. Boehm, M. K., Woof, J. M., Kerr, M. A. & Perkins, S. J. The Fab and Fc Fragments of IgA1 Exhibit a Different Arrangement from that in IgG: A Study by X-ray and Neutron Solution Scattering and Homology Modelling. *J Mol Biol* **286**, 1421–1447 (1999).

111. Qiu, J., Brackee, G. P. & Plaut, A. G. Analysis of the Specificity of Bacterial Immunoglobulin A (IgA) Proteases by a Comparative Study of Ape Serum IgAs as Substrates. *Infect Immun* **64**, 933–937 (1996).

112. McNicholas, S., Potterton, E., Wilson, K. S. & Noble, M. E. M. Presenting your structures: the CCP4mg molecular-graphics software. *Acta Crystallogr* **D67**, 386–394 (2011).

Chapter 3 – Structurally Characterizing the *Thomasclavelia ramosa* Immunoglobulin A Protease with Small-Angle X-ray Scattering Techniques

3.1 – Chapter Introduction

3.1.1 – Importance of Solution Scattering in Understanding Immunoglobulin A Proteases

Many pathogenic and commensal bacteria that inhabit human mucosal tissues produce a group of enzymes known as immunoglobulin A proteases (IgAPs) to cleave and inactivate the main antibody found on mucosal surfaces, immunoglobulin A (IgA).[1] These proteases have clinical applications as not only a drug target[2–4] but also as protein-based therapeutics to treat immunoglobulin-related diseases.[5,6] As such, it is important to understand IgAPs from a structural and functional perspective to encourage and further this clinically relevant research.

IgAPs are classified into three distinct enzyme families. Each family has their own structural fold and biochemical mechanism (*Section 2.1.2*) which have convergently evolved to adapt to their specific mucosal environment (*Section 2.1.5*).[1] The S6 and M26 families of IgA1Ps are produced from mucosal pathogens and are selective to the IgA1 subclass primarily found on respiratory tissues (*Section 2.1.3* and *2.1.4*).[7,8] This is in contrast to the third IgAP family represented by the protease from the commensal gut bacteria *Thomasclavelia ramosa*, which is selective to the IgA1 and IgA2m(1) subclasses both predominantly found along the intestinal tract (*Section 2.1.5*).[9,10]

The fact that there is a distinct divide between the IgAPs produced from pathogens versus commensals suggests that the structural diversity seen in the three IgAP families may have evolved under biological pressures influenced by their host-microorganism relationship. This divide becomes even more apparent when the proteases are described in the context of how they interact with the host. Pathogen-produced IgA1Ps are highly immunogenic, meaning that the host immune system has a baseline immunity against these proteases.[11] This is in contrast to the commensal-produced IgAP, which has low immunogenicity and does not illicit a strong immune response.[11] Furthermore, the ability of *T. ramosa* to produce its IgAP is not correlated with intestinal disease unlike S6 and M26 IgA1Ps which are considered virulence factors in respiratory infections.[12–14]

As with many biological phenomena, biological function manifests from protein structure. It is thus important to investigate this in the context of the three IgAP families to understand the

functional consequences of their divergent protease structures and how those differences relate to their host-microorganism relationship. This will also lay the groundwork for developing the aforementioned clinical avenues of IgAP research.

The most apparent structural difference between the pathogen- and commensal-produced IgAPs is their protein architecture. The pathogenic S6 and M26 IgA1Ps have a globular non-reducible fold while the *T. ramosa* IgAP has a modular and potentially minimizable protein architecture.[15,16] The modular nature of the *T. ramosa* IgAP is thus an insightful first point of investigation into the intricacies of IgAPs' role in host-microorganism relationships. The previous chapter showed that the *T. ramosa* IgAP could be split into individual domains that were able to be produced independent of the full enzyme. The crystal structures of some of these domains were determined and aided in synthesizing a working model for how the protease functioned. Due to the limitations of the approach in analyzing conformationally flexible proteins, most *T. ramosa* IgAP constructs were not amenable to structural determination via crystallography.

Small-angle X-ray scattering (SAXS) was used as a complementary and alternative approach to determine the low-resolution structures of various *T. ramosa* IgAP constructs in order to understand how individual domains behaved in the context of the full-length enzyme. SAXS is an appropriate structural method for analyzing multidomain structures as their inherent flexibility does not preclude the use of the tool and the measurements are done on particles in solution, where the proteins are able to undergo their full range of motion. These SAXS data revealed information about the spatial arrangement of the individual domains relative to each other and the potential biological function of the non-proteolytic domains.

3.1.2 – Introduction to Small-Angle X-ray Scattering and Its Limitations

SAXS is a technique that takes advantage of the physical phenomenon of X-ray scattering by electrons to gather low-resolution information about the shape, density, and size of particles in a sample.[17,18] This phenomenon results from the same underlying physics that leads to X-ray diffraction (*i.e.*, the physical process that occurs in an X-ray crystallography experiment), with the major difference being the lack of any preferred orientation for particles in a solution sample compared to that in a protein crystal.[19] As the purpose of crystallization is to directionally align the particles to amplify the scattered signal, SAXS data are inherently noisy and, due to the random

orientation of particles in solution, lack the directionality that is required to provide a unique structure for the scattering data (*Section 3.1.4*).[20] As such, SAXS should be used primarily to support hypotheses built on a pre-existing model rather than generate that model *ab initio*.

3.1.3 – The Small-Angle X-ray Scattering Curve and Its Transformations

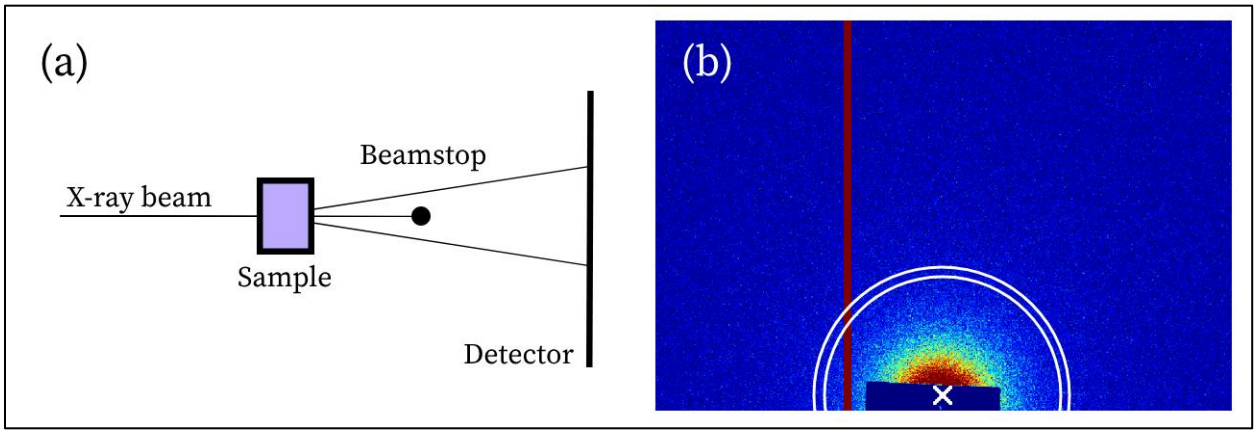


Figure 3.1.3.1 – Experimental Setup for Small-Angle X-ray Scattering. (a) The experimental setup for a SAXS experiment is nearly identical to that for X-ray crystallography. An X-ray beam, typically from synchrotron radiation, is passed through a sample (protein solution or buffer control) before it is blocked by a beamstop. The X-rays that are scattered by the sample solution are measured on a detector. **(b)** An example of raw SAXS data is shown as a heatmap of pixel intensity (colored from blue to red for zero to an arbitrarily high value, respectively) for a detector image after a sample was exposed to X-ray radiation. The intensities are integrated in thin slices (exaggerated slice shown as the region between the two white arcs) from the beam center (shown as an X) towards the edge of the image.

The raw data obtained from a SAXS experiment are a set of detector images that capture the X-ray scatter of protein molecules in solution (*Fig. 3.1.3.1a*).[19] As the protein molecules do not have a preferred orientation, the scattering is radially symmetric and hence the images are azimuthally integrated from the center of the X-ray direct beam outwards towards the edge of the detector (*Fig. 3.1.3.1b*) to obtain a scattering curve.[19] The scattering curve (*Fig. 3.3.2.1a*) plots the integrated intensity as a function of the normalized scattering vector, q, which is related to the scattering angle or, in crystallography terms, the interplanar spacing that leads to that scatter.[19] q also accounts for experimental set-up differences in X-ray wavelength such that these plots can be compared between beamlines.[19]

The fact that SAXS data are inherently noisy means that the non-proteinaceous components of the solution contribute significantly to the overall X-ray scatter.[20] As such, each protein sample must come with an exact or near-exact buffer control whose SAXS intensities are subtracted from

the scattering curve of the protein-buffer solution to obtain the scattering from just the protein.[20] These background-subtracted data are then used for analysis as described below.

The scattering curves are transformed in several ways to analyze different biophysical properties of the proteins in solution. The very low to low q range ($q*R_g$ values of 0 to ~1.0 to 1.7) is called the Guinier region (*Fig. 3.3.2.1b*), which can be empirically fit to a linear equation when the logarithm of intensity is plotted against q.[19] This regression is used to extrapolate information about the intensity scattered at zero angle, called I(0), and calculate the radius of gyration, called R_g.[19] I(0) provides insight into the concentration of protein in the sample and R_g represents the mass distribution of atoms in space and approximates particle size.[19] Both I(0) and R_g are not very insightful on their own but are parameters that are used to derive more useful information on particle compactness and shape,[21] which are more pertinent to the discussions below. Because the Guinier fit is empirical, the region of the low-q scatter that remains linear is dependent on the shape of the particles, with elongated rod-like particles deviating from linearity faster than globular particles.[19] Thus, it is important to take *a priori* knowledge on particle shape into consideration for Guinier analysis.

The medium q range is used in Kratky analysis (*Fig. 3.3.2.1c*) to garner information about how well-folded protein molecules are in solution.[21] This is based on how the scattering curve at medium q decays with respect to q.[21] Kratky curves must be normalized with respect to R_g and I(0) before they are compared between samples.[21] Compact proteins that are well-folded and globular have a characteristic bell-shaped Kratky curve that goes to zero at high $q*R_g$.[21] Proteins that are semi-folded tend to reach a non-zero baseline at high $q*R_g$, and fully disordered proteins tend to increase indefinitely, even at high $q*R_g$.[21]

Finally, the whole scattering curve is used in pair-distance distribution analysis (*Fig. 3.3.2.1d*) to provide information about protein shape.[19] An indirect Fourier transform is used to mathematically convert the scattering curve into a pair-distance distribution function, also known as a P(r) function.[19] This process is analogous to the Fourier transformation of crystallographic diffraction intensities into the Patterson function.[19] This means that P(r) represents a discrete distribution of distances between all combinations of atoms within a protein and, as such, proteins with particular shapes have characteristic P(r) distributions.[19] The maximum distance between two

atoms in the particle, D_{max}, can also be derived from the P(r) function and is another parameter that provides an estimate of particle size.[19]

3.1.4 – Structural Validation with Small-Angle X-ray Scattering Data

Kratky and P(r) analyses provide valuable information on the compactness and shape of proteins in solution, respectively,[19,21] which can be used to support hypotheses based on predictive models or crystal structures. Another method of validating structural hypotheses is by calculating the theoretical scatter from a model and seeing how well it fits the SAXS data.[20] This can be easily done as the physics behind X-ray scattering is well understood and can be accurately simulated with a starting model.[20]

Molecular dynamics or relaxation can be used in conjunction with this method to generate structural models that better account for interdomain flexibility.[18,20] The program CORAL (complexes with random loops) uses a library of loops to limit the distances between domains as the model is relaxed using simulated annealing to best fit the scattering density.[22] Since the domains of the *T. ramosa* IgAP already have their structures solved by crystallography (*Chapter 2*) or can be confidently predicted using AlphaFold2,[23,24] the individual domains can be docked and fit with randomized loop conformations into the SAXS data using CORAL as the length and location of the flexible linkers between domains is known.

3.1.5 – Chapter Objectives

The previous chapter focused on understanding the crystal structures of the *T. ramosa* IgAP and what these structures have to say about the functions of their individual domains. These sections explored aspects of their structures from a high-resolution perspective, zooming in on the conformations of individual side chains, secondary structure elements, and subdomains. X-ray crystallography is unfortunately not suited to structural analysis of multidomain *T. ramosa* IgAP constructs due to the inherent flexibility between domains. As such, I will further explore the structures of the *T. ramosa* IgAP using SAXS as it is able to provide information on this flexibility to gain insight into the biological function of this multidomain architecture and how it may differ compared to the other two IgAP families.

This chapter aims to (1) validate the NTD and MD+CTD1 crystal structures by showing that their theoretical scatter reasonably fits their solution scattering data. (2) confirm that the MD is well-folded without the need for CTD1, despite it needing CTD1 to crystallize. These first two goals will provide more evidence that the NTD, MD, and CTD1 are true domains in that they are well folded when produced in isolation of the rest of the enzyme. This will also support the validity of their crystal structures as being representative of their structures in solution. (3) model a low-resolution solution SAXS structure for CTD2 as its structure was unable to be solved by crystallography. And most importantly, (4) model a low-resolution solution structure of the full-length enzyme to understand how the individual domains behave in the context of the rest of the protein.

3.2 – Materials and Methods

3.2.1 – Small-Angle X-ray Scattering Sample Preparation and Data Collection

SAXS data were collected at the CHESS ID7A beamline using an X-ray wavelength of 1.1013 Å on an EIGER 4M detector.[25] All proteins were dialyzed against 25 mM HEPES, pH 7.5 with 1 mM TCEP overnight at 4 ^{o}C using a 12-14 kDa MWCO D-Tube Mini Dialyzer (Millipore Sigma) to obtain a matching solution for buffer subtraction. A 30 μL volume of sample was injected into the sample capillary for analysis. Samples were oscillated within the capillary during data collection to reduce radiation damage (2-3 oscillations/sec). 15 images were collected for each sample, with the detector being exposed for one second per image with no attenuation.

3.2.2 – Small-Angle X-ray Scattering Data Processing

Images were azimuthally averaged and buffer subtracted in RAW (version 2.1.4).[26] Processed data were subsequently analysed in RAW for Guinier and Kratky analysis,[26] GNOM (ATSAS package, version 3.2.1) for P(r) analysis,[27] and FoXS for simulating model scatter (web server, accessed February 27th, 2024).[28] The P(r) function was calculated using all measured data. The P(r) function was only calculated on truncated data (high q cutoff at ~0.3 or 0.4) if the P(r) fit to the scattering curve showed skewed residuals at $q > 0.3$ or 0.4, respectively.

CORAL was used to optimize the multidomain models to the SAXS scattering curves.[22] Where applicable, the individual domains of the NTD, MD+CTD1, CTD2, CTD3, and CTD4 were modelled with flexible linker lengths of 15, 4, 4, and 5 residues between each domain, respectively. The full-length enzyme was the only construct that was modelled with a 29-residue C-terminal flexible region. The NTD and MD+CTD1 domains were modelled using their crystal structures and CTD2, CTD3, and CTD4 were modelled with their AlphaFold2-predicted models,[23,24] which were truncated based on the boundaries outlined in *Chapter 2*.

3.3 – Results and Discussion

3.3.1 – Summary of Small-Angle X-ray Scattering Data

Table 3.3.1.1 – Small-Angle X-ray Scattering Data Details and Statistics.

Protein construct	NTD	MD	MD+CTD1	CTD2
Concentration (mg/mL)	3.16	1.83	1.60	3.0
q-measurement range	0.00584 – 0.50081	0.00584 – 0.50081	0.00821 – 0.51346	0.00821 – 0.51346
Guinier analysis				
I(0) (arbitrary units)	0.0541	0.0405	0.0705	0.0305
R_g (Å)	18.94	25.66	26.67	24.99
q-range ($Å^{-1}$)	0.0193 – 0.0712	0.0091 – 0.0511	0.0141 – 0.0506	0.0165 – 0.0531
$q_{min}*R_g$	0.365	0.233	0.376	0.389
$q_{max}*R_g$	1.349	1.311	1.349	1.007
R^2 fit	0.996	0.989	0.992	0.916
P(r) analysis				
I(0) (arbitrary units)	0.0541	0.0407	0.0707	0.0312
R_g (Å)	18.91	26.12	26.97	26.93
D_{max} (Å)	60	88	92	96
R^2 fit	0.951	0.882	0.653	0.690
q-range ($Å^{-1}$)	0.0193 – 0.4010	0.0091 – 0.5008	0.0141 – 0.4135	0.0156 – 0.5135
Volume analysis (kDa)				
Theoretical	31.63	54.78	62.53	13.36
Porod volume (V_p)	34.0	54.5	65.7	9.1
Correl. volume (V_c)	25.0	46.3	56.0	9.2

Protein construct	NTD+MD	CTD1-4	MD+CTD1-4	FL
Concentration (mg/mL)	1.14	3.05	1.10	0.74
q-measurement range	0.00584 – 0.50081	0.00821 – 0.51346	0.00821 – 0.51346	0.00821 – 0.51346
Guinier analysis				
I(0) (arbitrary units)	0.0350	0.0900	0.0791	0.0656
R_g (Å)	32.91	53.23	57.48	55.28
q-range ($Å^{-1}$)	0.0078 – 0.0400	0.0092 – 0.0190	0.0089 – 0.0175	0.0089 – 0.0185
$q_{min}*R_g$	0.258	0.490	0.514	0.494
$q_{max}*R_g$	1.305	1.011	1.007	1.023
R^2 fit	0.978	0.931	0.862	0.808
P(r) analysis				
I(0) (arbitrary units)	0.0352	0.0917	0.0824	0.0680
R_g (Å)	33.48	57.09	65.07	62.63
D_{max} (Å)	108	215	240	245
R^2 fit	0.944	0.635	0.672	0.619
q-range ($Å^{-1}$)	0.0078 – 0.3009	0.0092 – 0.5135	0.0111 – 0.5135	0.0089 – 0.5135
Volume analysis (kDa)				
Theoretical	88.0	38.7	93.5	129.6
Porod volume (V_p)	82.7	74.3	125.6	147.3
Correl. volume (V_c)	71.1	30.7	74.9	98.7

Sample concentrations and SAXS data statistics for proteins that will be discussed in this chapter are summarized in *Table 3.3.1.1*.

The NTD, NTD+MD, MD, and MD+CTD1 proteins were assumed to be globular based on the crystal structures of the NTD and MD+CTD1. CTD2 was assumed to be rod shaped due to the three predicted zinc-binding sites that connect each of the three beta sheets (*Fig. 2.3.1.3*). All constructs that contained CTD2 (CTD1-4, MD+CTD1-4, and FL) were also assumed to have an elongated shape. To account for these differences in shape, Guinier analyses were done with a $q_{max}*R_g$ of ~1.3 for globular constructs and ~1.0 for elongated constructs as these are the values of $q_{max}*R_g$ at which the scattering data theoretically begin to substantially differ from linearity for spheres and rods, respectively.[19] The Guinier correlation coefficients were high ($R^2 > 0.90$) for constructs with R_g values less than 35 Å. Larger constructs with R_g values greater than 50 Å showed lower values for the Guinier correlation coefficient ($R^2 \sim 0.8$-0.9). This may be due to the lack of very low q data required for larger particles as an inherent experimental restriction of the beamline. This is supported by the fact that the Guinier fits for these larger constructs have a $q_{min}*R_g$ value of ~0.5 as opposed to ~0.2 to 0.4 for the smaller constructs despite using lower q data to fit to the Guinier equation.

The consequences of this experimental restriction may have also manifested in the P(r) analysis, where larger constructs generally had lower χ^2 values relative to smaller individual domains. As the P(r) function is calculated from all scattering data, the lack of very low q data is important in determining D_{max} and may affect how well the Fourier transformation of P(r) fits to the scattering curve. There was, however, good agreement of the I(0) and R_g values as determined by both Guinier and P(r) analysis.

Finally, the solution scattering data suggested that all constructs were monomers in solution. This is supported by the calculated molecular weights (MWs) of the particles, which were generally within the range of the theoretical MW of a monomer. The only exceptions were the Porod volume values for all constructs that contained the entire C-terminal tail (CTD1-4, MD+CTD1-4, and FL) as there is more error associated with elongated particles relative to globular proteins when using this method to approximate MW.[29]

3.3.2 – Solution Scattering Data Supports a Modular Domain Architecture for the Protease

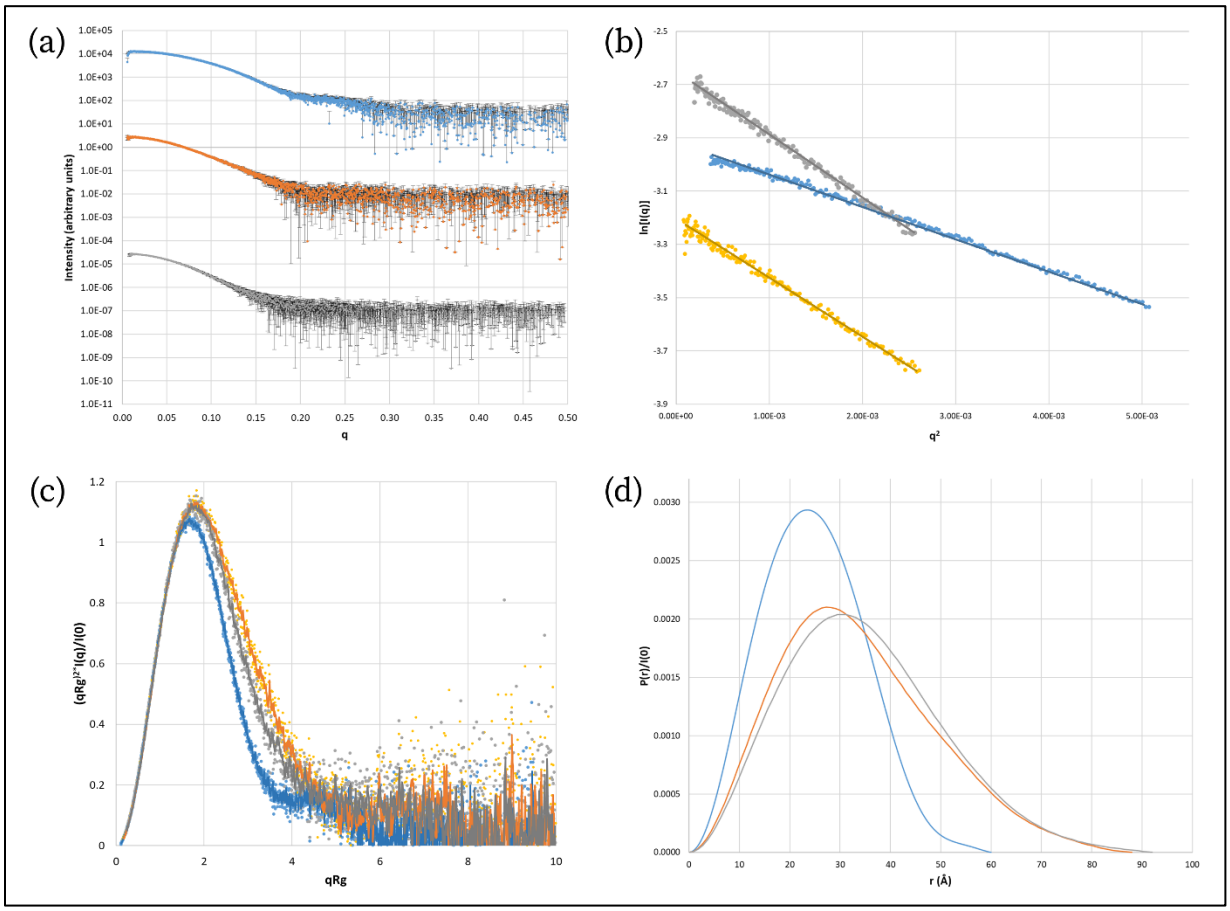


Figure 3.3.2.1 – SAXS Profiles and Transformations of the Globular N-terminal and Protease Domains. The scattering profiles of the *T. ramosa* IgAP NTD (blue), MD (orange), and MD+CTD1 (grey) proteins are consistent with that of well-folded globular domains. **(a)** The scattering curves were arbitrarily offset in I(q) to aid with visual comparison. Data in the scattering curves are presented as mean values ± one standard deviation across 15 azimuthally averaged detector images. These errors were calculated by RAW,[26] which follows standard statistical best practices for error propagation. **(b)** Guinier analysis for all data were fit to a $q_{max}*R_g$ of ~1.3, which is the rule-of-thumb cutoff for globular proteins. **(c)** Normalized Kratky analysis shows a bell-shaped curve characteristic of well-folded proteins. **(d)** The normalized P(r) curves are also consistent with globular proteins with a near-spherical (NTD) or ovoid shape (MD and MD+CTD1).[17]

The scattering profiles for NTD, MD, and MD+CTD1 are all consistent with well-folded globular proteins (*Fig. 3.3.2.1*). Their normalized Kratky plots (*Fig. 3.3.2.1c*) depict a bell-shaped curve, characteristic of a well-folded protein and the plots decay to zero at high $q*R_g$.[21] In addition, the peak of each curve has a value of approximately 1.1 at a $q*R_g$ of ~1.73, which coincides with the theoretical peak for globular proteins.[21] The Kratky plot peaks for MD and MD+CTD1 are slightly off from the theoretical globular peak as the crystal structure of the MD+CTD1 (*Fig. 2.3.4.1*) and, by inference, the structure of the MD is more elongated compared to an ideal sphere.

The NTD, on the other hand, aligns very well with the theoretical peak as seen in the globular fold of its crystal structure (*Fig. 2.3.3.1*). These differences in shape are also reflected in the P(r) plots (*Fig. 3.3.2.1d*) where the NTD has a near-perfect bell-shaped curve, characteristic of an idealized sphere, and the MD and MD+CTD1 proteins have a slightly tailed distribution as a result of being more ovoid.[17] Taken in concert, these results suggest that the NTD, MD, and MD+CTD1 are true domains in that they can fold and function independent of the context of the rest of the protein. Results of the MD, in particular, indicate that it can fold without CTD1 despite it being predicted as part of the same domain (*Section 2.3.1*) with AlphaFold2 and it being required for crystallization. As will be discussed in *Chapter 4*, this is further supported by the fact that the MD is active on its own, outside of the context of CTD1.

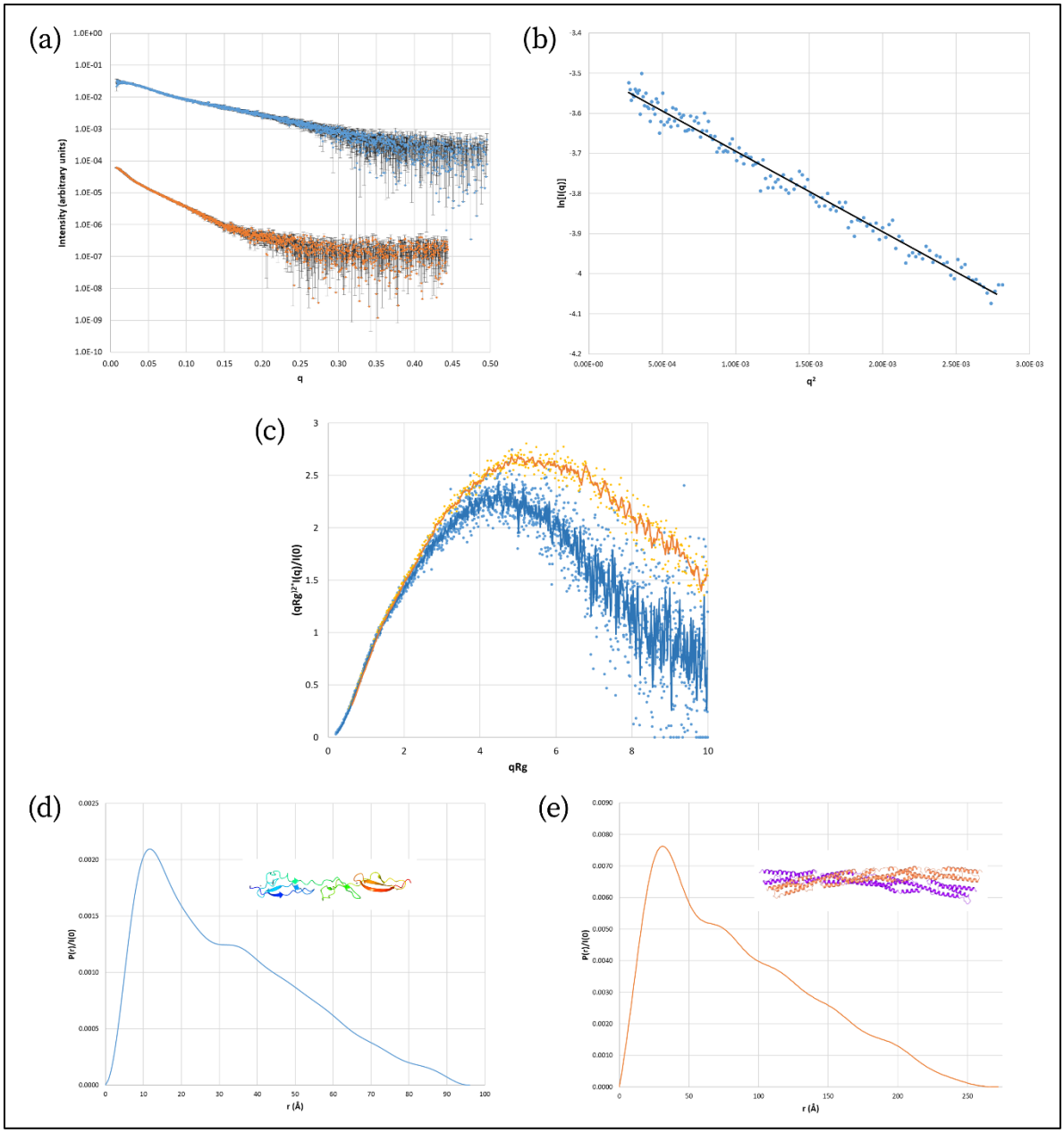


Figure 3.3.2.2 – SAXS Profiles and Transformations of the Rod-Like C-terminal Domain 2 and Human α-Actinin. The scattering profiles of the *T. ramosa* IgAP CTD2 (blue) and human α-actinin[31] (orange) proteins are consistent with that of well-folded rod-like domains. **(a)** The scattering curves were arbitrarily offset in I(q) to aid with visual comparison. Data in the CTD2 scattering curve are presented as mean values ± one standard deviation across 15 azimuthally averaged detector images. These errors were calculated by RAW,[26] which follows standard statistical best practices for error propagation. Data for human α-actinin were taken from the SASBDB (entry #SASDJL6).[31,37] **(b)** Guinier analysis for CTD2 data were fit to a q_{max}*R_g of ~1.0, which is the rule-of-thumb cutoff for rod-shaped proteins. **(c)** Normalized Kratky analysis shows a bell-shaped curve characteristic of well-folded proteins. The Kratky plots for both curves are cut off at a q*R_g value of 10 as the curves become significantly noisier at higher q*R_g values. However, both curves eventually reach baseline to indicate that the proteins are well-folded. The normalized P(r) curves for CTD2 **(d)** and α-actinin **(e)** are consistent with the distinct skewed distribution of elongated rod-like proteins.[17] The AlphaFold2 model of CTD2 and crystal structure of α-actinin (PDB 1HCI)[30] are inset within their respective P(r) curves.

As hypothesized based on several predicted zinc-binding motifs (*Fig. 2.3.1.3a*), the SAXS data for CTD2 are consistent with an elongated rod-like structure (*Fig. 3.3.2.2*) and its normalized Kratky and P(r) plots are distinct from those seen for the globular domains (*Fig. 3.3.2.1cd*). The SAXS profile for human α-actinin is shown as a reference for a rod-like protein with a solved crystal structure (*Fig. 3.3.2.2ce*).[30,31] The Kratky plot for CTD2 (*Fig. 3.3.2.2c*) extends much further than previously seen for the globular domains but still decays to zero at high $q*R_g$ to indicate that the protein is well-folded. The domain's rod-like structure is most supported by the distinctly tailed P(r) function (*Fig. 3.3.2.2d*), which is very similar in distribution to that of the extended human α-actinin but on a shorter D_{max} scale (*Fig. 3.3.2.2e*).

The CTD2 is structurally homologous to a family of repetitive and modular bacterial domains that are involved in biofilm formation and protein display on bacterial surfaces.[32] These so-called SHIRT domains were structurally characterized in proteins found in *Streptococcus gordonii* that contained multiple structural repeats of a beta-sheet-rich module.[32] Some of these modules are connected by short but rigid proline-rich linkers that allow for the protein to adopt an extended conformation.[32,33] As such, proteins containing these domains have been called periscope proteins as they have been hypothesized to play a role in extending functional parts of the protein away from the bacterial cell surface.[32] Although the CTD2 is predicted to contain three of these homologous beta-sheet-rich modules, they are not connected via proline-rich linkers but rather use zinc-binding motifs to bridge and rigidify the connection between the modules. This has yet to be described in any of the currently characterized periscope domain repeats and may be another conserved mechanism by which periscope domains operate.[32–34] The function of CTD2 is thus hypothesized to help display the NTD and MD further away from the surface of the bacterium to aid in IgA proteolysis. The two modules found in CTD4 are also hypothesized to act in the same manner and may have the same function as CTD2 based on their structural homology.

The NTD and MD+CTD1 crystal structures alongside the AlphaFold2 model of CTD2 and implied structure of MD were validated by their corresponding SAXS data. The FoXS web server was used to generate theoretical SAXS data from each domain's individual structures.[28] The predicted scattering curves were superimposed against their corresponding real data and each domain's structure was shown to have a good fit with their SAXS data (*Fig. 3.3.2.3*).

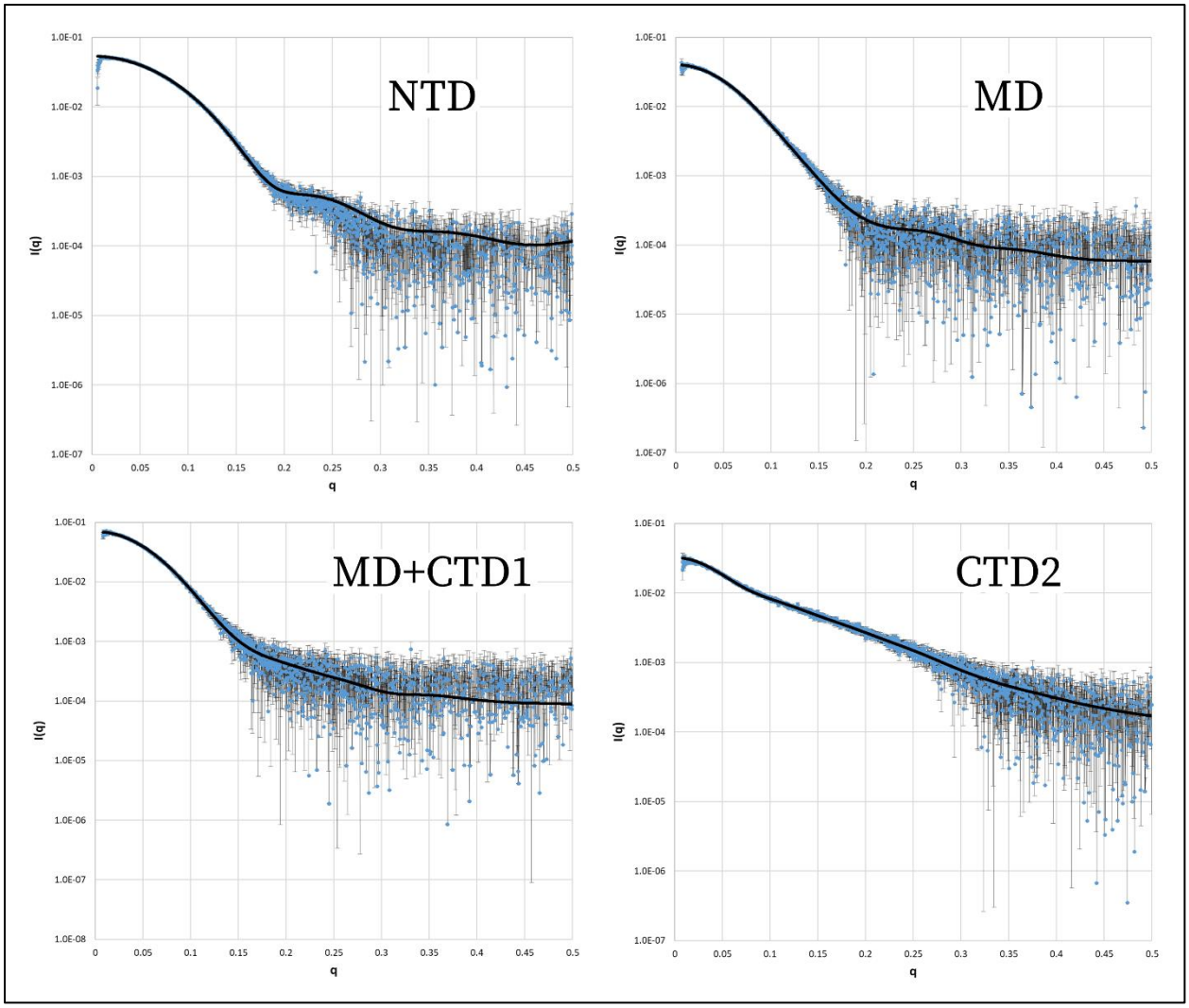


Figure 3.3.2.3 – Simulated Scatter of Crystal Structures and AlphaFold2 Models Fit Raw Scattering Data. The FoXS web server [28] was used to generate the simulated solution scattering intensity (black line) of the NTD and MD+CTD1 crystal structures and the CTD2 AlphaFold2 model with manually added zinc ions. The simulated scatter of MD was generated from the crystal structure of MD+CTD1 with CTD1 removed from the model. Data in the scattering curves are presented as mean values ± one standard deviation across 15 azimuthally averaged detector images. These errors were calculated by RAW,[26] which follows standard statistical best practices for error propagation. All simulated profiles showed good data-to-model fits, with FoXS χ^2 values of 1.94, 1.14, 0.74, 0.90 for NTD, MD, MD+CTD1, and CTD2, respectively.

Overall, there is good agreeance on all accounts between the high-resolution crystallographic and AlphaFold2-predicted models and the low-resolution SAXS data. This once again strongly suggests that the NTD, MD, CTD1, and CTD2 are well-folded and can function independent of the rest of the enzyme. This also suggests that the crystal structures are representative of what the proteins look like in solution, outside of the context of the crystal.

3.3.3 – The Full-Length Protease Has an Extended Conformation in Solution

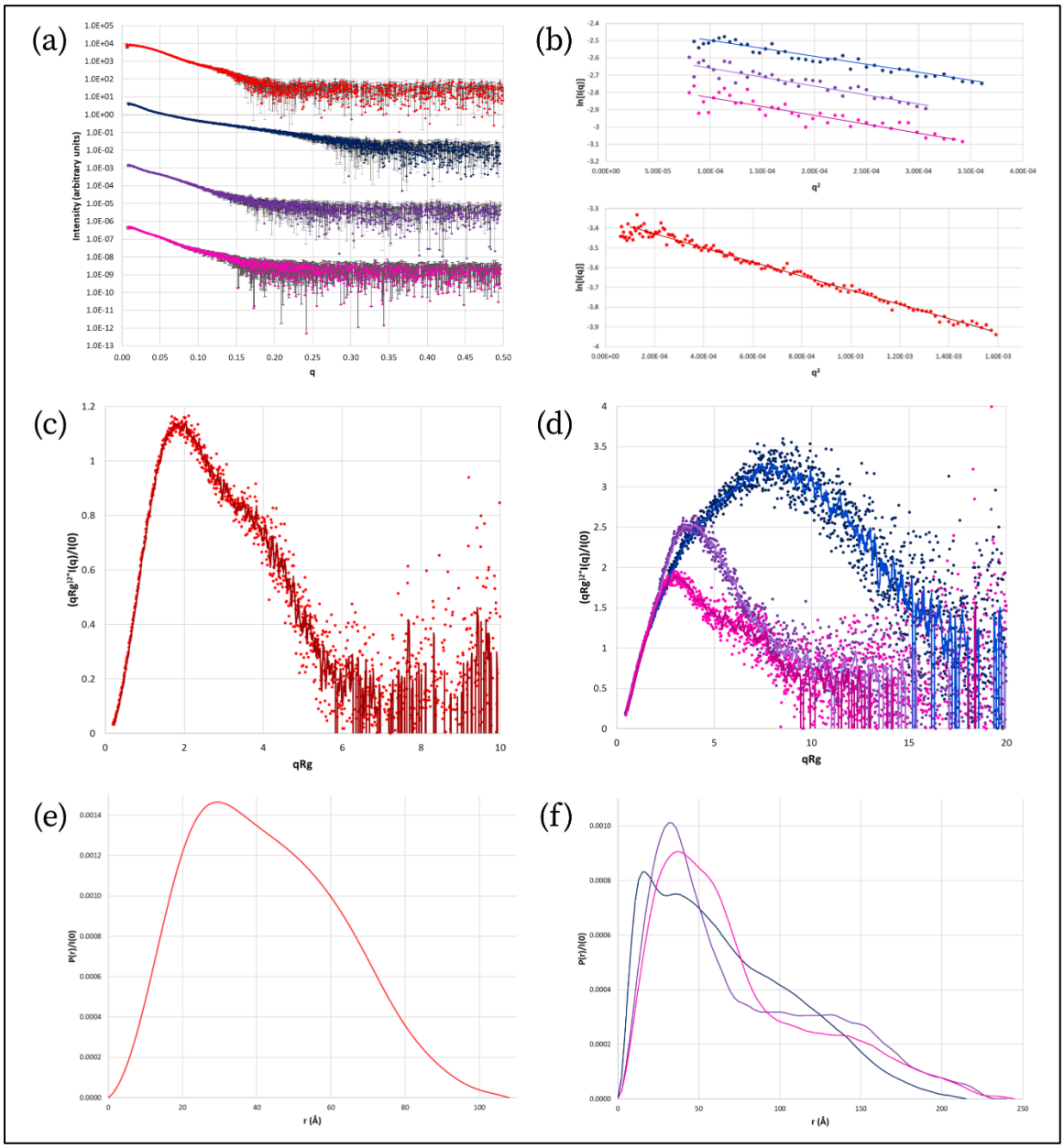


Figure 3.3.3.1 – SAXS Profiles of Multidomain *Thomasclavelia ramosa* Immunoglobulin A Protease Constructs. The scattering profile of the *T. ramosa* IgAP NTD+MD (red) protein is consistent with a globular two-domain protein while the *T. ramosa* IgAP CTD1-4 (dark blue), MD+CTD1-4 (purple), and FL (magenta) constructs are consistent with that of an elongated protein. **(a)** The scattering curves were arbitrarily offset in I(q) to aid with visual comparison. Data are presented as mean values ± one standard deviation across 15 azimuthally averaged detector images. These errors were calculated by RAW,[26] which follows standard statistical best practices for error propagation. **(b)** Guinier analysis for elongated constructs were fit to a $q_{max}*R_g$ of ~1.3 (top subsection) and ~1.0 for the globular NTD+MD construct (bottom subsection). **(c)** Normalized Kratky analysis of the NTD+MD data shows a curve characteristic of a well-folded globular protein. **(d)** Normalized Kratky analysis of the other data shows curves characteristic of elongated particles. **(e)** The bimodal normalized P(r) curve for NTD+MD is consistent with a particle with a flexible linker between two well-folded domains. **(f)** The right-skewed normalized P(r) curves for the other data is consistent with a general rod-like shape.

The scattering profile for NTD+MD is consistent with a globular two-domain protein with an internal flexible linker (*Fig. 3.3.3.1*). Like the NTD, MD, and MD+CTD1 data in *Fig. 3.3.2.1c*, its Kratky plot has a peak value of approximately 1.1 at a $q*R_g$ of ~1.73 (*Fig. 3.3.3.1c*), which is consistent with a well-folded globular protein.[21] However unlike the bell-shaped curve of the previous globular P(r) plots (*Fig. 3.3.2.1d*), the NTD+MD P(r) function has a subtle bimodal distribution with two overlapping peaks at a radius of ~20 and ~50 Å, characteristic of two independent domains joined by a flexible linker (*Fig. 3.3.3.1e*).[17]

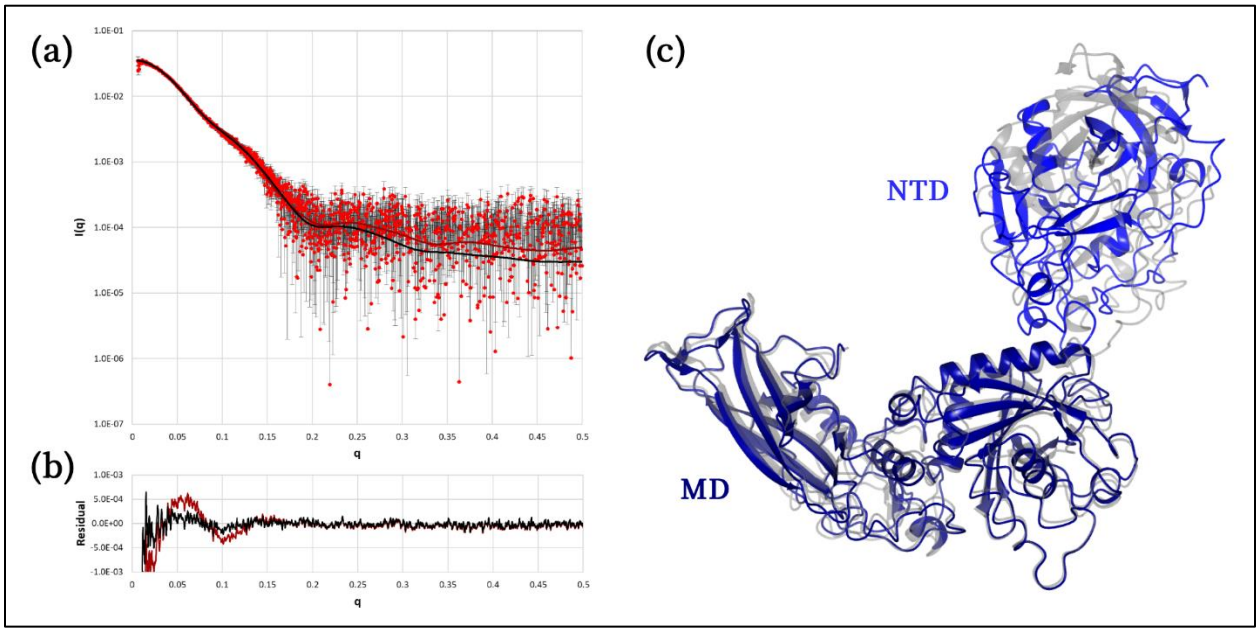


Figure 3.3.3.2 – Simulated Scatter of CORAL-Relaxed and AlphaFold2 Models for the NTD+MD Construct. The FoXS web server[28] was used to generate the simulated solution scattering intensities of the NTD+MD CORAL-relaxed (black) and AlphaFold2-predicted (dark red) models. **(a)** These theoretical scattering curves were compared to real data (red) and **(b)** residuals were calculated for each measured intensity. Data in the scattering curves are presented as mean values ± one standard deviation across 15 azimuthally averaged detector images. These errors were calculated by RAW,[26] which follows standard statistical best practices for error propagation. The simulated profile for the CORAL-relaxed model (FoXS χ^2 value of 1.00) had a better data-to-model fit compared to the AlphaFold2-predicted model (FoXS χ^2 value of 1.60). **(c)** The CORAL-relaxed model (colored) has a similar conformation of NTD relative to MD compared to the AlphaFold2-predicted model (transparent grey).

The NTD+MD AlphaFold2 model was validated by using the FoXS web server[28] to generate a simulated scattering curve that was superimposed over its SAXS data (*Fig. 3.3.3.2a*). Although this model's theoretical scatter generally agreed with the raw data, there were some residual deviations in the low-to-medium q range (q of 0 to 0.15) which suggested that there were differences in global conformation that were unaccounted for in the model (*Fig. 3.3.3.2b*). CORAL

was thus used to optimize the AlphaFold2 model to account for the flexibility of the 15-residue linker between the NTD and MD domains as seen in its P(r) plot (*Fig. 3.3.3.1e*). The better-fitting CORAL-relaxed structure was only slightly different to the AlphaFold2 model, where the NTD was modelled in a rotated conformation relative to its original orientation (*Fig. 3.3.3.2c*). These models also suggest that the linker may not be as disordered as initially predicted as the NTD is found in close proximity to the MD in both models and the linker is not fully extended.

The scattering profiles for the CTD1-4, MD+CTD1-4, and FL constructs are shown in *Fig. 3.3.3.1*. The CTD1-4 Kratky (*Fig. 3.3.3.1d*) and P(r) plots (*Fig. 3.3.3.1f*) are similar to CTD2 and human α-actinin (*Fig. 3.3.2.2*), which indicates that CTD1-4 has a rod-like shape.[30,31] This supports the hypothesis that CTD4 also has a rigid elongated shape due to its structural similarities to CTD2. The entire C-terminal tail thus takes on an extended conformation that may function as a molecular spacer that extends the functional MD out from the surface of the bacterium.[32] This was postulated for the individual CTD2 and CTD4 domains in the previous section but now is also applicable to their role in context of the entire C-terminal tail.

The Kratky (*Fig. 3.3.3.1d*) and P(r) plots (*Fig. 3.3.3.1d*) of the MD+CTD1-4 and FL data have features of both globular and rod-like particles. The MD+CTD1-4 P(r) function has a single bell-shaped curve at low r, corresponding to the well-folded globular MD, as seen in *Fig. 3.3.2.1e*, and has a right-skewed tail as a result of the CTD1-4 domains. Likewise, the FL P(r) function has a bimodal curve at low r, corresponding to the well-folded two-domain NTD+MD, as seen in *Fig. 3.3.3.1d*, and a right-skewed tail as a result of the CTD1-4 domains.

The fact that attributes from the globular domains are seen alongside those from the elongated C-terminal tail provides strong evidence supporting the independent nature of each domain. This becomes more apparent when the AlphaFold2 models of the CTD1-4, MD+CTD1-4, and FL constructs are relaxed with CORAL (*Fig. 3.3.3.3*), where the individual domains are modelled in different conformations relative to each other in each structure. Although this may be due to the inherent limitations of SAXS (*Section 3.1.2*) and/or biases generated from manually designating linkers in CORAL, all three CORAL-relaxed models depicted the C-terminal tail in an extended conformation contrary to the compact conformation of the AlphaFold2 model. The proposed periscope function of the C-terminal tail therefore likely plays an important biological function.

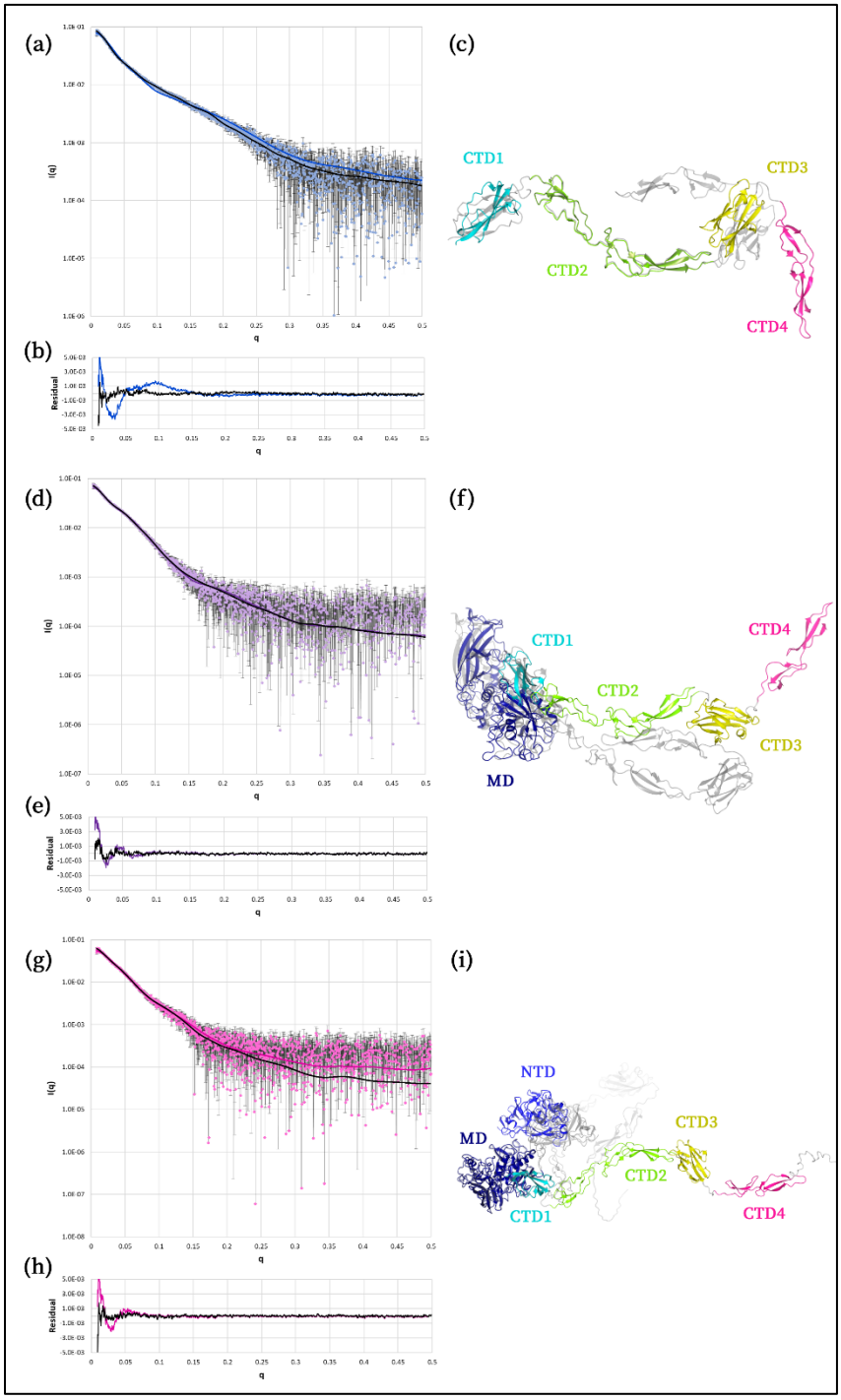


Figure 3.3.3.3 – Simulated Scatter of CORAL-Relaxed and AlphaFold2 Models for Constructs with the C-terminal Domains. The FoXS web server [28] was used to generate the simulated solution scattering intensities of the CORAL-relaxed (black) and AlphaFold2-predicted **(a-c)** CTD1-4 (dark blue), **(d-f)** MD+CTD1-4 (dark purple), and **(g-i)** FL (dark pink) models. **(a,d,g)** Theoretical scattering curves for each protein were compared to real data and **(b,e,h)** residuals were calculated for each measured intensity. Data in the scattering curves are presented as mean values ± one standard deviation across 15 azimuthally averaged detector images. These errors were calculated by RAW,[26] which follows standard statistical best practices for error propagation. The simulated profiles for the CORAL-relaxed models had better data-to-model fits relative to their AlphaFold2-predicted models. The FoXS χ^2 values of the CORAL-relaxed and AlphaFold2-predicted fits, respectively, were 0.79 and 2.98 for CTD1-4, 0.74 and 0.95 for MD+CTD1-4, and 0.65 and 0.93, for FL. **(c,f,i)** The CORAL-relaxed models (colored) were always in an extended conformation relative to the AlphaFold2-predicted model (transparent grey).

3.3.4 – The Full-Length Protease in the Context of Immunoglobulin A

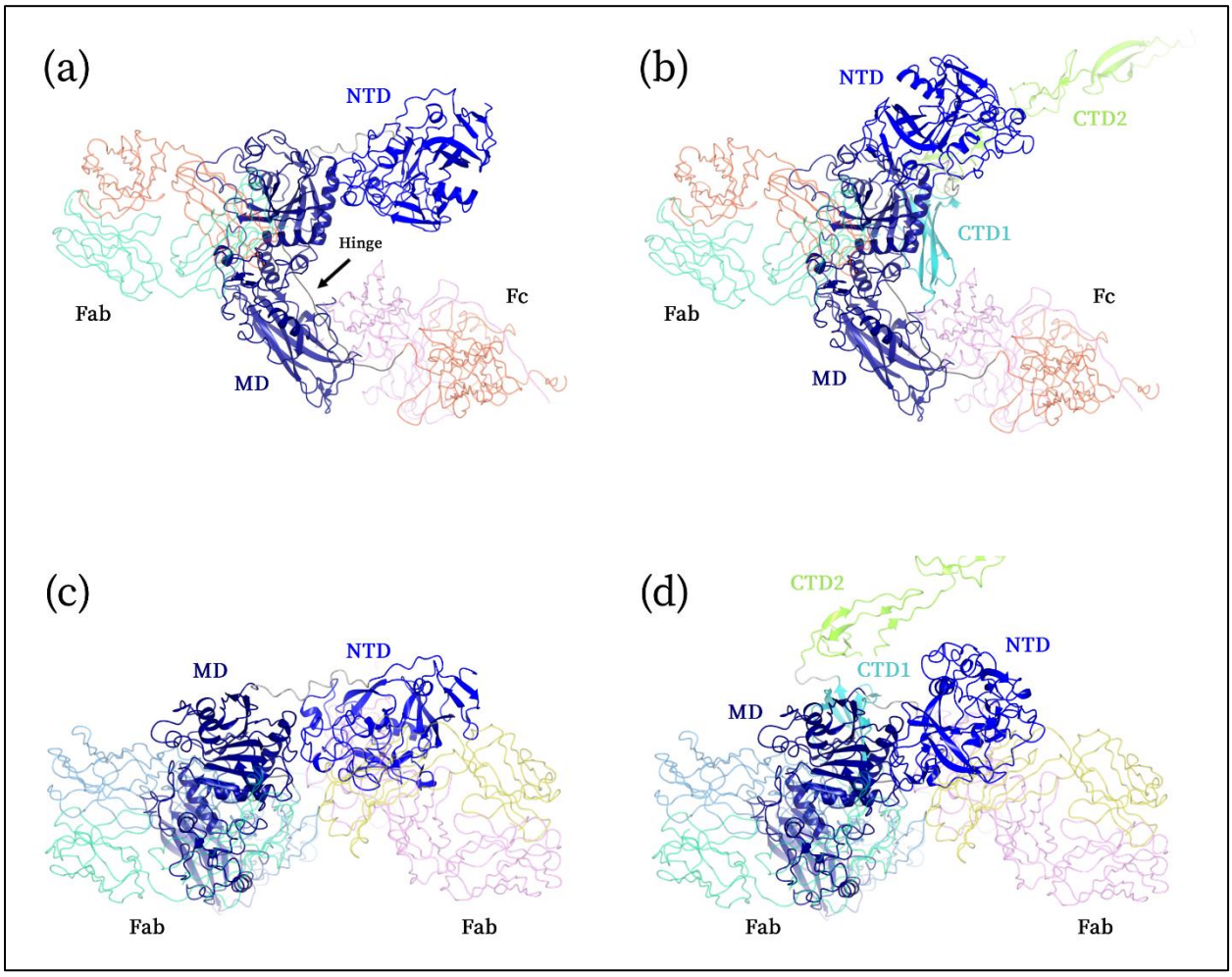


Figure 3.3.4.1 – Working Model of the Full-Length Enzyme-Substrate Complex. The solution-scattering structures of IgA1 (PDB 1IGA)[38] **(a-b)** and IgA2m(1) (PDB 1R70)[39] **(c-d)** were superimposed onto the C_H1-bound protease domain model to generate the substrate-enzyme complex as described in *Fig. 2.3.7.6*. Non-proteolytic domains of the protease were modelled according to the solution structures of the NTD+MD **(a and c)** and FL **(b and d)** constructs. The rest of the C-terminal tail for the FL model extends away from the substrate and is not fully shown in the figure. The heavy chains of IgA1 are depicted in orange and pink and the light chain of the single Fab is in green. The heavy chains of IgA2m(1) are depicted in blue and pink and the light chains are in green and yellow. Both IgA1 and IgA2m(1) are shown with a perspective looking down on the antibody from the top of the Fab domains. Only one Fab domain is shown for IgA1 for simplicity. The Fc domain for IgA2m(1) is below the perspective plane and is hidden to remove visual clutter.

The previous chapter detailed a working model for how the *T. ramosa* IgAP protease domain recognizes IgA (*Section 2.3.7*). It is important to revisit this with what has been revealed about the extended conformation of the full-length enzyme to see how it fits into the working model. Based on the models shown in *Fig. 3.3.4.1ac*, the NTD may be able to recognize and bind to structural elements on the immunoglobulin as a means of substrate capture. This is more likely in the case of IgA2m(1) as the NTD is in very close proximity to its Fab, and the enzyme-substrate

complex with NTD+MD shows severe steric clashes between the Fab and NTD (*Fig. 3.3.4.1c*). Due to the extended hinge of IgA1, the NTD is modelled to be much further away from IgA1 (*Fig. 3.3.4.1a*) compared to IgA2m(1) in its substrate-enzyme complex. It is still possible that NTD plays a role in the capture of IgA1 as the 15-residue linker between NTD and MD may allow for this interaction. However, the fact that the NTD is modelled in close proximity to the MD in the CORAL-relaxed NTD+MD model (*Fig. 3.3.3.2c*) suggests that it may have difficulties in sampling this extended conformation and that linker may not be as flexible as initially predicted.

When the CORAL-relaxed FL model is used to model the substrate-enzyme complexes (*Fig. 3.3.4.1bd*), the placement of the NTD relative to the MD does not significantly shift. The C-terminal tail, however, is oriented away from the substrate in such a way that it cannot make physical contact with IgA, even if the substrate was modelled to be secretory IgA. Under this working hypothesis, the C-terminal domains are not involved in enzyme catalysis or substrate binding, reinforcing the hypothesis that they likely function as a periscope to extend the proteolytic domain away from the bacterium.

3.4 – Chapter Summary and Next Steps

This chapter focused on understanding the *T. ramosa* IgAP from a low-resolution structural biology perspective to compliment the previous structural work done via X-ray crystallography. The low-resolution technique of SAXS was able to reveal key features of the protease that strongly supported the true modular nature of the enzyme. These data confirmed that the NTD, MD, and CTD1 are true domains and that the crystal structures of the NTD and MD+CTD1 are representative of their shape in solution. The CTD2 and CTD4 domains were also shown to be well-folded and have an elongated and rigid shape as hypothesized based on their AlphaFold2-predicted models. All of the domains were put into context of the full-length enzyme, which was shown to have an extended conformation in solution contrary to the AlphaFold2 model. These solution structures were then investigated in terms of the protease's working model of substrate binding which revealed that the NTD may play a role in substrate binding or selection. Under this model, the C-terminal domains likely act purely as periscope domains to allow for the protein to display the proteolytically active domains out further from the surface of the bacterium.

Even though the *T. ramosa* IgAP has a distinct protein architecture relative to other IgAPs, it is similar to the M26 IgA1P family in that this family also contains a variety of domains which have separate non-proteolytic functions to the IgA1P.[35] Unlike the S6 and M26 IgA1Ps, the proteolytically active super-domain (MD+CTD1) of the *T. ramosa* IgAP can be further split to remove the CTD1 without abolishing activity (*Chapter 4*). This results in a minimally active domain that is less than half the size of the full-length enzyme, as opposed to the S6 and M26 IgA1Ps which require the majority of the protein to retain activity.[16,36]

The investigation via crystallographic and solution-scattering techniques detailed in the previous and current chapters is the first account of an in-depth structural characterization of the *T. ramosa* IgAP and the third IgAP family. Unfortunately, this investigation was unable to provide substantial structural evidence to rationalize why the *T. ramosa* IgAP is not very immunogenic unlike the other IgA1Ps.[11] In the absence of concrete evidence, one can suppose that its reduced protease size relative to the S6 and M26 IgA1Ps may play a part in lowering immunogenicity.[16,36] As will be described in *Chapter 4*, there may also be differences in enzymatic activity between the IgAP families which are important to consider in their immunogenicity and biological role. The NTD is the next logical line of questioning due to its unknown function and potential role in the recognition of IgA based on the working model. The CTD3 is also another open avenue of investigation, as it was unable to be produced under the current construct design and purification method. However, since it lies within the C-terminal tail, it is unlikely that it plays a direct role in facilitating IgAP activity and may not contribute to the *T. ramosa* IgAP's lack of immunogenicity.

3.5 – Chapter References

1. Mistry, D. & Stockley, R. A. IgA1 protease. *Int J Biochem Cell Biol* **38**, 1244–1248 (2006).
2. Mulks, M. H., Kornfeld, S. J., Frangione, B. & Plaut, A. G. Relationship Between the Specificity of IgA Proteases and Serotypes in Haemophilus influenzae. *J Infect Dis* **146**, 266–274 (1982).
3. Kilian, M., Mestecky, J. & Schrohenloher, R. E. Pathogenic Species of the Genus Haemophilus and Streptococcus pneumoniae Produce Immunoglobulin A1 Protease. *Infect Immun* **26**, 143–149 (1979).
4. Mulks, M. H. & Plaut, A. G. IgA Protease Production as a Characteristic Distinguishing Pathogenic from Harmless Neisseriaceae. *N Engl J Med* **299**, 973–976 (1978).
5. Lamm, M. E., Emancipator, S. N., Robinson, J. K., Yamashita, M., Fujioka, H., Qiu, J. & Plaut, A. G. Microbial IgA protease removes IgA immune complexes from mouse glomeruli in vivo: Potential therapy for IgA nephropathy. *Am J Pathol* **172**, 31–36 (2008).
6. Xie, L.-S., Huang, J., Qin, W. & Fan, J.-M. Immunoglobulin A1 protease: A new therapeutic candidate for immunoglobulin A nephropathy. *Nephrology* **15**, 584–586 (2010).
7. Male, C. J. Immunoglobulin A1 Protease Production by Haemophilus influenzae and Streptococcus pneumoniae. *Infect Immun* **26**, 254–261 (1979).
8. Woof, J. M. & Mestecky, J. Mucosal immunoglobulins. *Immunol Rev* **206**, 64–82 (2005).
9. Fujiyama, Y., Kobayashi, K., Senda, S., Benno, Y., Bamba, T. & Hosoda, S. A Novel IgA Protease from Clostridium sp. Capable of Cleaving IgA1 and IgA2 A2m(1) but not IgA2 A2m(2) allotype paraproteins. *The Journal of Immunology* **134**, 573–576 (1985).
10. Brandtzaeg, P., Kett, K. & Rognum, T. O. Subclass Distribution of IgG- and IgA-Producing Cells in Secretory Tissues and Alterations Related to Gut Diseases. in *Recent Advances in Mucosal Immunology* (eds. Mestecky, J., McGhee, J. R., Bienenstock, J. & Ogra, P. L.) 321–333 (1987).
11. Kobayashi, K., Hagiwara, K., Kondoh, H. & Fujiyama, Y. Resistance of Normal Serum IgA and Secretory IgA to Bacterial IgA Proteases: Evidence for the Presence of Enzyme-Neutralizing Antibodies in Both Serum and Secretory IgA, and Also in Serum IgG. *Microbiol Immunol* **31**, 1097–1106 (1987).
12. Troeger, C., Blacker, B., Khalil, I. A., Rao, P. C., Cao, J., Zimsen, S. R. M., Albertson, S. B., Deshpande, A., Farag, T., Abebe, Z. *et al.* Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Infect Dis* **18**, 1191–1210 (2018).

13. Brooks, L. R. K. & Mias, G. I. Streptococcus pneumoniae's Virulence and Host Immunity: Aging, Diagnostics, and Prevention. *Front Immunol* **9**, (2018).

14. Senda, S., Fujiyama, Y., Ushijima, T., Hodohara, K., Bamba, T., Hosoda, S. & Kobayashi, K. Clostridium ramosum, an IgA Protease-Producing Species and Its Ecology in the Human Intestinal Tract. *Microbiology and Immunity* **29**, 1019–1028 (1985).

15. Wang, Z., Rahkola, J., Redzic, J. S., Chi, Y. C., Tran, N., Holyoak, T., Zheng, H., Janoff, E. & Eisenmesser, E. Mechanism and inhibition of Streptococcus pneumoniae IgA1 protease. *Nat Commun* **11**, (2020).

16. Johnson, T. A., Qiu, J., Plaut, A. G. & Holyoak, T. Active site gating regulates substrate selectivity in a chymotrypsin-like serine protease. The structure of Haemophilus influenzae IgA1 protease. *J Mol Biol* **389**, 559–574 (2009).

17. Mertens, H. D. T. & Svergun, D. I. Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *J Struct Biol* **172**, 128–141 (2010).

18. Brosey, C. A. & Tainer, J. A. Evolving SAXS versatility: solution X-ray scattering for macromolecular architecture, functional landscapes, and integrative structural biology. *Curr Opin Struct Biol* **58**, 197–213 (2019).

19. Jacques, D. A. & Trewhella, J. Small-angle scattering for structural biology - Expanding the frontier while avoiding the pitfalls. *Protein Science* **19**, 642–657 (2010).

20. Da Vela, S. & Svergun, D. I. Methods, development and applications of small-angle X-ray scattering to characterize biological macromolecules in solution. *Curr Res Struct Biol* **2**, 164–170 (2020).

21. Burger, V. M., Arenas, D. J. & Stultz, C. M. A Structure-free Method for Quantifying Conformational Flexibility in proteins. *Sci Rep* **6**, (2016).

22. Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikhney, A. G., Gajda, M., Gorba, C., Mertens, H. D. T., Konarev, P. V. & Svergun, D. I. New developments in the ATSAS program package for small-angle scattering data analysis. *J Appl Crystallogr* **45**, 342–350 (2012).

23. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

24. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S. & Steinegger, M. ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679–682 (2022).

25. Acerbo, A. S., Cook, M. J. & Gillilan, R. E. Upgrade of MacCHESS facility for X-ray scattering of biological macromolecules in solution. *J Synchrotron Radiat* **22**, 180–186 (2015).

26. Hopkins, J. B., Gillilan, R. E. & Skou, S. BioXTAS RAW: improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis. *J Appl Crystallogr* **50**, 1545–1553 (2017).

27. Svergun, D. I. Determination of the Regularization Parameter in Indirect-Transform Methods Using Perceptual Criteria. *J. Appl. Cryst* **25**, 495–503 (1992).

28. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. Accurate SAXS Profile Computation and its Assessment by Contrast Variation Experiments. *Biophys J* **105**, 962–974 (2013).

29. Piiadov, V., Ares de Araújo, E., Oliveira Neto, M., Craievich, A. F. & Polikarpov, I. SAXSMoW 2.0: Online calculator of the molecular weight of proteins in dilute solution from experimental SAXS data measured on a relative scale. *Protein Science* **28**, 454–463 (2019).

30. Ylänne, J., Scheffzek, K., Young, P. & Saraste, M. Crystal Structure of the α-Actinin Rod Reveals an Extensive Torsional Twist. *Structure* **9**, 597–604 (2001).

31. Sponga, A., Arolas, J. L., Schwarz, T. C., Jeffries, C. M., Rodriguez Chamorro, A., Kostan, J., Ghisleni, A., Drepper, F., Polyansky, A., De Almeida Ribeiro, E. *et al.* Order from disorder in the sarcomere: FATZ forms a fuzzy but tight complex and phase-separated condensates with α-actinin. *Sci. Adv* **7**, 7653–7681 (2021).

32. Whelan, F., Lafita, A., Gilburt, J., Dégut, C., Griffiths, S. C., Jenkins, H. T., St John, A. N., Paci, E., Moir, J. W. B., Plevin, M. J. *et al.* Periscope Proteins are variable-length regulators of bacterial cell surface interactions. *Proc Natl Acad Sci U S A* **118**, (2021).

33. Whelan, F. The long and the short of Periscope Proteins. *Biochem Soc Trans* **50**, 1293–1302 (2022).

34. Monzon, V., Lafita, A. & Bateman, A. Discovery of fibrillar adhesins across bacterial species. *BMC Genomics* **22**, (2021).

35. Paukovich, N., Redzic, J. S., Chi, Y. C., Rahkola, J. T., Issaian, A., Blue, A., Hansen, K. C., Janoff, E. N. & Eisenmesser, E. Z. Streptococcus pneumoniae G5 domains bind different ligands. *Protein Science* **28**, 1797–1805 (2019).

36. Redzic, J. S., Rahkola, J., Tran, N., Holyoak, T., Lee, E., Martín-Galiano, A. J., Meyer, N., Zheng, H. & Eisenmesser, E. A substrate-induced gating mechanism is conserved among Gram-positive IgA1 metalloproteases. *Commun Biol* **5**, (2022).

37. Kikhney, A. G., Borges, C. R., Molodenskiy, D. S., Jeffries, C. M. & Svergun, D. I. SASBDB: Towards an automatically curated and validated repository for biological scattering data. *Protein Science* **29**, 66–75 (2019).

38. Boehm, M. K., Woof, J. M., Kerr, M. A. & Perkins, S. J. The Fab and Fc Fragments of IgA1 Exhibit a Different Arrangement from that in IgG: A Study by X-ray and Neutron Solution Scattering and Homology Modelling. *J Mol Biol* **286**, 1421–1447 (1999).

39. Furtado, P. B., Whitty, P. W., Robertson, A., Eaton, J. T., Almogren, A., Kerr, M. A., Woof, J. M. & Perkins, S. J. Solution Structure Determination of Monomeric Human IgA2 by X-ray and Neutron Scattering, Analytical Ultracentrifugation and Constrained Modelling: A Comparison with Monomeric Human IgA1. *J Mol Biol* **338**, 921–941 (2004).

Chapter 4 – Enzymologically Characterizing the *Thomasclavelia ramosa* Immunoglobulin A Protease

4.1 – Chapter Introduction

4.1.1 – Importance of Kinetic Characterization in Understanding Immunoglobulin A Proteases

Immunoglobulin A proteases (IgAPs) are a group of enzymes secreted from a wide variety of pathogenic and commensal mucosal bacteria that have convergently evolved to cleave human immunoglobulin A (IgA) as a means of reducing the effectiveness of the adaptive immune system.[1] Its role in the virulence of pathogenic mucosal bacteria as well as its high specificity for IgA make them a unique system to study to better understand host-pathogen relationships and enzyme selectivity.[1]

Most of the recent research on IgAPs has focused on determining their structure and hypothesizing about what aspects of their structure are important for catalytic function and substrate recognition.[2–4] There have also been a handful of studies which have detailed methods to assay IgAPs, mainly for the purposes of identifying novel inhibitors to treat bacterial infection.[5–10] There is however a large gap in understanding between these two biochemical approaches. Studies that assay for IgAP activity make little reference to the proteases' structures and structural studies likewise do not use much activity data to support hypotheses about structure.

The importance of kinetic characterization in understanding IgAPs therefore cannot be overstated. These kinetic experiments are the missing piece that bridge the gap between structure and function by providing a means to test hypotheses generated from pure structural analysis on an *in vitro* functional system. This is particularly pertinent to IgAPs as the structural mechanisms for their high substrate specificity have only been described from a purely structural perspective.

4.1.2 – Chapter Objectives

The previous chapters established the domain architecture of the *T. ramosa* IgAP and showed that the domains can function independently. The ability to produce well-folded domains was fortuitous for studying their individual structures but does not provide much insight into their function in the context of the full-length enzyme. This chapter will take an enzymological approach to understanding the *T. ramosa* IgAP by appreciating how the different non-proteolytic domains influence the activity of the protease domain.

A brief overview of the inherent challenges of assaying for IgAP activity will first be summarized before discussing the kinetic assays previously used for characterizing IgAPs. Due to a variety of shortcomings with these methods, a gel-based kinetic assay was developed and validated to obtain Michaelis-Menten kinetic parameters that can be compared amongst all *T. ramosa* IgAP constructs to assess the impact of including non-proteolytic domains on enzymatic activity. Finally, mass spectrometry was used to confirm that the minimized enzyme constructs still retained substrate selectivity despite the loss of the non-proteolytic domains.

4.1.3 – Limitations of Existing Immunoglobulin A Protease Kinetic Assays

As introduced in *Chapter 1*, IgAPs are known to have high substrate specificity as they are unable to cleave general protease substrates and are able to selectively proteolyze IgA over other immunoglobulin classes.[3,4,11–13] This is achieved through mechanisms that have convergently evolved to recognize IgA at sites distal to the active site which trigger structural changes required for proteolysis (*Sections 2.1.3*, *2.1.4*, and *2.3.7*).

This specificity has made it inherently difficult to find suitable substrates to study IgAPs from an enzymological perspective. Choudary *et al.* have designed and characterized a fluorescent peptide based on the autoproteolysis sites of various IgA1Ps that is able to be cleaved by the *Neisseria meningitidis*, *Haemophilus influenzae*, and *Streptococcus sanguinis* IgA1Ps.[8] Although this peptide was successfully used to identify preliminary inhibitors for the *H. influenzae* IgA1P and gives a reliable readout of proteolysis,[9] it is not a suitable substrate for structure-function studies as it is not representative of the proteases' biological substrate and thus would be unable to assess the potential role of non-proteolytic domains on activity. A similar approach was taken by Steinrücke *et al.* by designing a substrate that had the hinge-peptide sequence inserted in between an N-terminal hexahistidine tag and a C-terminal α-helical domain to detect IgAP activity via surface plasmon resonance (SPR).[14] This approach has the same issue as the fluorescent peptide in that it is not biologically relevant and, even though rates can be extracted from SPR data, the kinetics may be confounded by the fact that the substrate is immobilized on a solid surface in the assay. The only suitable substrate for these studies is thus the full IgA molecule. The main concern then becomes what method will be used to generate a readout and if the limitations of this method prevent kinetic comparisons between enzyme constructs.

IgAP activity was historically measured with iodine-125-labelled myeloma IgA (^{125}I-IgA) as the substrate.[15] This was done by quantifying the number of radioactive counts emitted from ^{125}I-labelled Fab fragments following separation by electrophoresis and extracting initial rates from the changes in counts in samples taken over time.[16–18] As preparation of ^{125}I-IgA involved the chemical modification and covalent coupling of ^{125}I atoms to aromatic residues on the protein,[19,20] it is possible that radiolabeling has an effect on IgAP activity.[16] With the availability of larger quantities of IgA and recombinant expression of IgAPs, it is no longer necessary to use ^{125}I-IgA to amplify the signal from a minute quantity of cleaved product to make up for the fact that natively purified IgAPs had low specific activity.[16] Despite this method having been shown to be able to quantify kinetic parameters for the *S. sanguinis* IgA1P, its difficulty of use, particularly in the generation of the ^{125}I-IgA substrate, coupled with its more outdated measurement method makes it generally incompatible with most lab setups.

IgAP activity has also been quantified with an enzyme-linked immunosorbent assay (ELISA).[15,21] IgAP was incubated with unlabeled IgA1 for several hours at 37°C in a separate reaction mixture prior to being quenched by protease inhibitor.[21] This mixture was added directly to an ELISA plate pre-coated with a peroxidase-conjugated anti-IgA heavy chain antibody to quantify the amount of intact IgA remaining after protease treatment.[21] In theory, ELISAs can be suitably used for kinetic analysis. However, due to the experimental methodology used for these studies and their different end goals, their findings can only be interpreted relative to their other results and cannot be extrapolated to understand kinetic differences between IgAPs.

To obtain reliable kinetic information that can be compared between enzyme constructs and IgAPs from different species, I developed a gel-based kinetic assay that can be viewed as a modernized version of the aforementioned ^{125}I-IgA assay. Since the goal of the assay is to derive kinetic parameters that can be meaningfully compared between enzymes, it is important to consider what assumptions are underlying the equations that will be used in the kinetic analysis and whether the experimental methodology fits those assumptions. The Michaelis-Menten equation will first be briefly discussed through this lens to lay a background for understanding the choices made during assay design and validation.

4.1.4 – A Brief Synopsis of Michaelis-Menten Kinetics

The purpose of kinetic analysis is to explain how and why the biophysical behavior of enzymes change as a result of experimental conditions. This type of analysis can be used as a means of quantification to compare the catalytic competencies of different enzymes or as a tool to understand how enzymes function from a structure-function perspective. Due to the multitude of ways an enzyme's activity can be impacted by experimental conditions, it is extremely important to understand how kinetic assays are designed and executed to ensure that the conclusions drawn from kinetic analysis are meaningful; mixing arbitrary amounts of enzyme with arbitrary amounts of substrate will always lead to an experimental outcome, but whether that outcome can be interpreted is another question.

Many enzymes are subject to product inhibition, a phenomenon in which the product of an enzyme-catalyzed reaction binds to the enzyme and prevents it from catalyzing the reaction to its fullest capabilities.[22] This is due to the fact that, in many cases, the product is chemically and/or structurally similar to the substrate. In addition, if a reaction is left for too long prior to its readout, the concentration of substrate will become unknown and the rate may slow down as the substrate becomes depleted.[22] The scientists who founded the field of enzyme kinetics, Victor Henri, Leonor Michaelis, and Maud Menten, controlled for this by performing their assays under initial-rate conditions, or in other words conditions that approximated the rate of product formation at the instant the enzyme was mixed with substrate.[22] These rates would thus not be influenced by the effects of product inhibition, as no product would be formed, nor substrate depletion, as the concentration of substrate would be equal to the starting concentration.[22] As such, initial rates greatly simplify the interpretation of results from both a mathematical and biophysical perspective.

The general rule of thumb for creating initial-rate conditions is to use an enzyme concentration significantly less than the substrate concentration as well as to perform measurements in a timeframe where product inhibition and substrate depletion are considered negligible.[23] What is deemed negligible has historically been 5-10% of the lowest substrate concentration used, however this may be extended to larger percentages if an enzyme is weakly affected by product inhibition.[22] This is empirically validated by assessing the amount of product accumulated over time in what is known as a progress curve and ensuring that it remains linear throughout the rate-determining region.[22]

The gold-standard method for assessing and comparing enzyme activity under initial-rate conditions is described by the Michaelis-Menten equation:[24]

$$v([S]) = \frac{V_{max}[S]}{K_M + [S]} = \frac{k_{cat}[E][S]}{K_M + [S]}$$

This equation describes the initial rate of product formation, ν, at varying concentrations of substrate, [S], and a known concentration of enzyme, [E].[24] It is important to note that this equation is used as an empirical and mathematical model that is fit to real data;[22] it is fortuitous that most enzymes, when assayed under Michaelis-Menten conditions, have rates that can be fit to this curve.[22] This curve is fit by two parameters: the maximal initial rate, V_{max}, and the Michaelis-Menten constant, K_M.[22] By definition, V_{max} is the asymptotic limit of the curve, meaning that it is the initial rate that is theoretically measured at an infinite substrate concentration for a given enzyme concentration.[22] K_M is the substrate concentration at which the enzyme has an initial rate that is half of V_{max}.[22] Although K_M is often referred to as the pseudo-equilibrium constant for forming the enzyme-substrate complex, it has little bearing on how tightly the enzyme binds substrate and should not be treated as such.[25]

The biophysical meaning of these parameters is much more complicated and nuanced than it might seem. Unlike in simpler small-molecule systems, it is virtually impossible to convert the values of these parameters to the elementary rate constants of steps involved in substrate binding, conformational change, catalysis, and product release. In most cases, we can only understand these parameters in terms of the general steps of enzyme catalysis that they represent:

$$E + S \leftrightharpoons ES \leftrightharpoons EP \rightarrow E + P$$

Here, a single enzymatic cycle is presented and starts with a free enzyme (E) binding to a substrate (S) to form an enzyme-substrate complex (ES). ES then undergoes catalysis to form an enzyme-product complex (EP) before releasing the product (P) to regenerate the free enzyme. Product release is shown as a unidirectional arrow as it is assumed to be irreversible due to the low amount of product formed. Because all steps must occur to regenerate the enzyme, kinetic measurements encapsulate all steps of an enzymatic cycle.[26] In other words, the time it takes for each step of the cycle to occur is reflected in the overall kinetic measurement and the amount of time taken in each step can change depending on the concentration of substrate in the reaction.[26]

The two fundamental parameters that biophysically represent the Michaelis-Menten curve are k_{cat}/K_M and k_{cat} (calculated by $V_{max}/[E]$), not k_{cat} and K_M.[25] These parameters represent the initial-rate behaviors of the enzyme at near-zero and saturating concentrations of substrate, respectively.[25,27] Under substrate-limited conditions (near-zero concentrations), the step that has the greatest influence on the measured rate is that which has the highest transition-state energy, often referred to as the rate-determining step.[25,27] For any given enzyme, the rate-determining step can be substrate binding, chemical catalysis, or product release.[27] This is not to say that non-rate-determining steps do not affect k_{cat}/K_M as its value is influenced by all steps prior to and including the first irreversible step in the enzymatic reaction.[25,27] Another way to look at k_{cat}/K_M is that it represents the rate at which productive complexes that are "destined" to reach product are formed.[25] k_{cat} can be complementarily viewed as the other half of this process, where it represents the rate at which these productive complexes dissociate and free up the enzyme.[25] In this scenario, every enzyme-substrate complex is considered to be productive and thus the value of k_{cat} is influenced by all steps after the formation of ES all the way to the release of product.[25] This is because at infinite substrate concentration, the time it takes to form ES becomes zero.

The Michaelis-Menten equation is built on two related but independent underlying assumptions.[23] The steady-state assumption states that concentration of all intermediates prior to the rate-determining step do not change during the reaction as the rate of forming and dissolving these intermediates is equivalent.[22,23] This occurs after an initial transient "burst" phase that generates enough intermediate species to reach this state and it is assumed that this timescale occurs much faster than the timescale for enzyme turnover.[22,23] The second assumption is the reactant stationary assumption, which states that the concentration of substrate does not appreciably change when steady state is reached, even after some substrate is consumed in the initial transient phase that forms all required steady-state intermediates.[23] This assumption may break down if the substrate concentration begins to come close to the enzyme concentration but is also dependent on the value of K_M.[23] Regardless, these assumptions are generally valid if the enzyme concentration is significantly lower than the substrate concentration and K_M.[22,23]

In summary, an assay that follows the assumptions of the Michaelis-Menten equation must be designed in such a way that:

(1) The enzyme concentration is several magnitudes less than all substrate concentrations assayed.
(2) It is sensitive enough to detect product in the concentration range where issues of product inhibition and substrate depletion are negligible. What is deemed negligible must be empirically determined.
(3) Along the same line as (2), the signal is strongly correlated with the concentration of product, particularly in the regions described in (2).
(4) It can provide enough data to perform a reasonable regression of the linear region of the progress curve to extract initial rates. This means that it can discern between the linear and non-linear regions of the progress curve.
(5) And particular to the case of IgAPs, it uses and can detect its biologically relevant substrate, IgA.

4.2 – Materials and Methods

4.2.1 – Preparation of the Monomeric Immunoglobulin A1 Substrate

Monomeric myeloma IgA1 (Athens Research and Technology, Athens, GA, USA) was loaded onto a pre-packed Superdex 200 Increase 10/300 GL column (Cytiva; Marlborough, MA, USA), equilibrated in 25 mM HEPES, pH 7.5, and run at 0.5 mL/min. All buffer-exchanged fractions containing IgA1 were pooled, concentrated with a 50 kDa molecular-weight-cutoff (MWCO) centrifugal concentrator (Millipore Sigma; Burlington, MA, USA), frozen in 20 μL pellets via direct immersion in liquid nitrogen, and stored at -80 ^{o}C. Protein concentration was measured using the 1% mass extinction coefficient of 13.2, as provided by the manufacturer. IgA1 was used at the frozen concentration (for a 38 μM IgA1 reaction) and diluted accordingly for reactions containing 1, 3, 5, 10, and 20 μM IgA1.

4.2.2 – Preparation of the Thomasclavelia ramosa Immunoglobulin A Protease Enzymes

30 μL of each *T. ramosa* IgAP enzyme construct was diluted to 500 μL with 25 mM HEPES, pH 7.5 + 1 mM TCEP in the presence of 25% (v/v) Chelex 100 resin (BioRad) and was mixed overnight at room temperature. The Chelex 100 resin was prepared and equilibrated as previously described in the Methods section in *Chapter 2*. Chelex 100 resin was pelleted by light centrifugation at 1,000xg for 5 minutes and the supernatant was incubated with 1 mM $ZnCl_2$ for two hours at room temperature. The zinc-supplemented supernatant was then desalted back into 25 mM HEPES, pH 7.5 + 1 mM TCEP with a 0.5 mL 7 kDa MWCO Zeba Spin Desalting Column (Thermo Scientific, Waltham, MA, USA) and the enzyme concentration was remeasured using 1% mass extinction coefficients theoretically determined from their primary sequence using the ProtParam program in Expasy (*Appendix A*)[28] prior to use in the reaction mixture.

4.2.3 – Preparation of the Fully Cleaved Immunoglobulin A1 Internal Standards

250 μL of 1 μM IgA1 and 500 μL of 2 μM IgA1 (Athens Research and Technology) were incubated with 40 nM *T. ramosa* IgAP MD overnight at 37 °C in 25 mM HEPES, pH 7.5. 25 μL samples were prepared by mixing 20 μL reaction volume with 5 μL 4X SDS-PAGE loading dye (0.2 M Tris, pH 6.8 + 8% (w/v) SDS + 0.4% (w/v) bromophenol blue + 40% (v/v) glycerol + 4% (v/v) β-mercaptoethanol + 50 mM EDTA). Samples were boiled for 5 min at 95 °C in a dry-heat block and spun down for 1 min at 17,000xg prior to being frozen at -20 °C. Aliquots were thawed as needed for each gel. 1 μM internal standards were only used for gels that contained 20 pmol IgA1 and 2 μM internal standards were used for all other gels.

4.2.4 – Standardized Protocol for the Gel-Based Kinetic Assay

Various enzyme constructs were incubated with 1, 3, 5, 10, 20, and 38 μM IgA1 (Athens Research and Technology) at 37 °C in 25 mM HEPES, pH 7.5. 0.15, 0.5, 0.75, or 1.5 nM enzyme was used for all MD, MD+CTD1, NTD+MD, and FL reactions, respectively, with the exception of the 38 μM IgA1 reaction for MD where 0.5 nM enzyme was used. Five samples were taken from the same on-going reaction at 30-minute time points over the course of two and a half hours. The reaction was quenched with the addition of 4X SDS-PAGE loading dye and samples were

boiled for 5 min at 95 ^{o}C in a dry-heat block. All samples were spun down for 1 min at 17,000xg prior to being completely loaded onto the gel with gel-loading tips.

All samples were prepared in 25 μL containing 20 μL reaction volume and 5 μL 4X SDS-PAGE loading dye (0.2 M Tris, pH 6.8 + 8% (w/v) SDS + 0.4% (w/v) bromophenol blue + 40% (v/v) glycerol + 4% (v/v) β-mercaptoethanol + 50 mM EDTA). Reactions that contained more than 2 μM IgA1 were diluted with buffer such that the final sample contained 40 pmol of IgA1. Samples for the 1 μM IgA1 reactions were not diluted to make the sample (*i.e.*, 20 μL of the reaction was added directly to the sample).

Gels were cast with a 12% (w/v) resolving and 4% (w/v) stacking layer using a 30% (w/v) stock 37.5:1 acrylamide:bis-acrylamide solution (BioRad). The BioRad 1.0 mm-thickness 10-well Mini-PROTEAN and Tetra Cell electrophoresis systems were used for casting and running the gels, respectively. The resolving layer was cast such that it stopped 2 cm below the top of the glass short plate. Gels were initially run at 80 V before being ramped up to 160 V when the dye front encountered the separating gel (approximately 30 min). The gels were stopped when the dye front ran off the bottom of the gel (approximately 60 min from the start of the 160V phase).

100 mL of Whatman Grade 1-filtered (Cytiva) SDS-PAGE stain (0.25% (w/v) Coomassie R-250 + 50% (v/v) methanol + 10% (v/v) glacial acetic acid) was added to the gels, microwaved for 1 min, and left to stain on a rocker for 5 min. 200 mL of SDS-PAGE destain (20% (v/v) methanol + 10% (v/v) glacial acetic acid) was then added, microwaved for 1 min, and the gels were left to destain on a rocker for two hours at room temperature. The gels were then immediately imaged on a Gel Doc XR+ gel imager (BioRad) with a white-light transilluminator. All SDS-PAGE stain and destain reagents were purchased from BioShop (Burlington, ON, Canada).

4.2.5 – Gel Image Processing and Michaelis-Menten Curve Fitting

Gel images were processed with the BioRad Image Lab (version 6.1.0, build 7) program. The band intensities for the heavy-chain fragment associated with the IgA1 Fc (H1) were integrated using the program's built-in densitometry features. The background-corrected H1 band intensity was divided by that of the H1 band in the internal standard to calculate the percentage of IgA1 cleaved at every time point. This percentage was plotted with respect to time and a simple

linear regression was used to calculate the rate of the reaction (in units of molarity of IgA1 cleaved per minute) from multiplying the slope of the regression (in units of percentage of IgA1 cleaved per minute) by the concentration of substrate in the reaction. The rate was then normalized by the enzyme concentration, plotted as a function of IgA1 concentration, and fit to the Michaelis-Menten equation to extract kinetic parameters. k_{cat} and K_M parameters were fit and their associated standard error were calculated using the Enzyme Kinetics module (version 1.3) in SigmaPlot (version 11.0). Standard errors for k_{cat}/K_M were calculated using statistical equations for the propagation of error.

4.2.6 –Peptide Cleavage Analysis by MALDI-ToF Mass Spectrometry

A synthetic peptide (N-VPCPVPST-C; GenScript, Piscataway, NJ, USA) based on the *T. ramosa* IgAP IgA1 cleavage site was dissolved in 25 mM HEPES, pH 7.5 and frozen in small-volume aliquots at -20 ºC for storage. 10 nM FL or MD *T. ramosa* IgAP was incubated with 15 μM peptide in 25 mM HEPES, pH 7.5 in a reaction volume of 20 μL for 18 hours at 37 ºC. Peptide without enzyme was also subjected to the same conditions as the negative control.

The entire reaction volume was subjected to a C_{18} ZipTip (Millipore Sigma) to prepare each sample for mass spectrometry. The sample was first acidified with trifluoroacetic acid (TFA; Millipore Sigma) to a final concentration of 0.5% (v/v). The C_{18} resin was wetted with 100% (v/v) acetonitrile and then equilibrated in 0.1% (v/v) TFA, before allowing the peptides in the sample to bind. After washing the resin with 0.1% (v/v) TFA, the bound peptides were desorbed with a 5 μL mixture of 1% (v/v) formic acid and 50% (v/v) methanol. 2 μL sample was mixed with a 2 μL of α-cyano-4-hydroxycinnamic acid (HCCA; made in-house at the University of Waterloo Mass Spectrometry Facility) as the matrix, dried onto the target plate, and subjected to matrix-assisted laser desorption/ionization time-of-flight (MALDI-ToF) mass spectrometry. This was done at the University of Waterloo Mass Spectrometry Facility on a Bruker Autoflex Speed MALDI instrument.

4.3 – Results and Discussion

4.3.1 – Developing a Versatile Gel-Based Kinetic Assay for Immunoglobulin A Proteases

Although the ^{125}I-IgA assay has a number of drawbacks, the idea of using electrophoresis to separate cleavage products followed by the identification of those products using a gel-based technique is a relatively simple and accessible way to determine IgAP activity. Taking inspiration from this ^{125}I-IgA assay, I designed and validated an IgAP kinetic assay that can quantify the amount of cleavage product by densitometric analysis on a reducing and denaturing sodium-dodecyl-sulfate polyacrylamide gel-electrophoresis (SDS-PAGE) gel.

An SDS-PAGE-based assay has been previously used to analyze the activity of the *S. pneumoniae* IgA1P.[29] These researchers, however, were only interested in testing the protease against putative small-molecule inhibitors and hence let the reaction run to near completion (75-90%) to amplify the effects of the inhibitors.[29] The assay was thus not run under initial-rate conditions and is not suitable for kinetic analysis in its described state. The following discussion describes the various stages of method validation that were done to ensure that the newly developed IgAP assay was suitable for Michaelis-Menten kinetic analysis.

It is important to note that the true biological substrate for IgAPs is secretory IgA (*Section 1.3*). However, for the purposes of simplicity and as the starting point for assay development, monomeric IgA1 was used in all of the experiments described below. Monomeric IgA also contains all of the structural features that are hypothesized to play a role in substrate recognition in the *H. influenzae* (*Section 2.1.3*) IgA1P, *S. pneumoniae* (*Section 2.1.4*) IgA1P, and *T. ramosa* IgAP protease domain (*Section 2.3.7*). It should thus be a suitable starting substrate to understand the role of other domains in the *T. ramosa* IgAP.

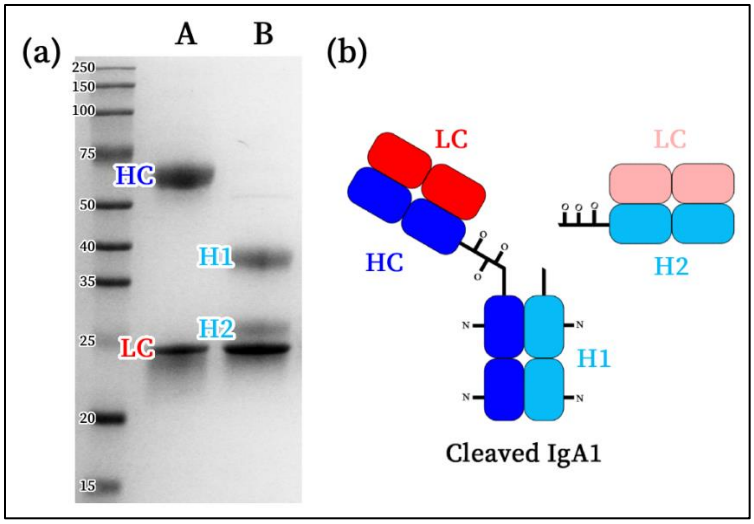


Figure 4.3.1.1 – Cleavage Pattern of IgA1 on an SDS-PAGE Gel. (a) The SDS-PAGE banding pattern of intact IgA1 (A) and IgA1 cleaved by the *T. ramosa* IgAP (B) are distinct and can be separately analyzed via densitometry. **(B)** The intact heavy chain (HC; blue) is split into two fragments following proteolysis while the light chain (LC; red and light red) remains unchanged. The HC fragment associated with the Fc (H1; light blue) is larger compared to the fragment associated with the Fab (H2; light blue) due the N-linked glycosylations on the Fc.

The substrate (intact IgA1) and products (IgA1 Fc and Fab domains) of the IgAP-catalyzed reaction are easily visualized and separated on an SDS-PAGE gel (*Fig. 4.3.1.1*). As introduced in *Section 1.3*, monomeric IgA1 has a quaternary structure consisting of two heavy chains (HC) and two light chains (LC).[30] Under the reducing and denaturing conditions of the SDS-PAGE gel, the covalent and non-covalent bonds that hold the quaternary structure break apart and lead to two discrete bands on the gel. When monomeric IgA1 is treated with IgAP, the HC is cleaved into two products associated with the Fab or Fc.[1] The HC region associated with the Fc is seen at ~37.5 kDa (HC fragment #1; H1) on an SDS-PAGE gel and the HC region associated with the Fab is seen at ~30 kDa (HC fragment #2; H2). Although H1 and H2 have protein components that are very similar in size, they migrate differently on the gel due to the glycosylations on each fragment.[17] H1 thus typically runs higher than H2 due to the N-linked glycosylations on the Fc.[31]

The gel-based IgAP assay capitalizes on the distinct banding patterns of the substrate and product to determine the initial rates of IgA1 cleavage at various substrate concentrations. This was done by quantifying the intensity of these bands using densitometry and tracking how this intensity changed over time to extract initial rates. It was important to validate the assay prior to making kinetic measurements of *T. ramosa* IgAP constructs to ensure that the experimental conditions followed Michaelis-Menten assumptions and could be meaningfully discussed.

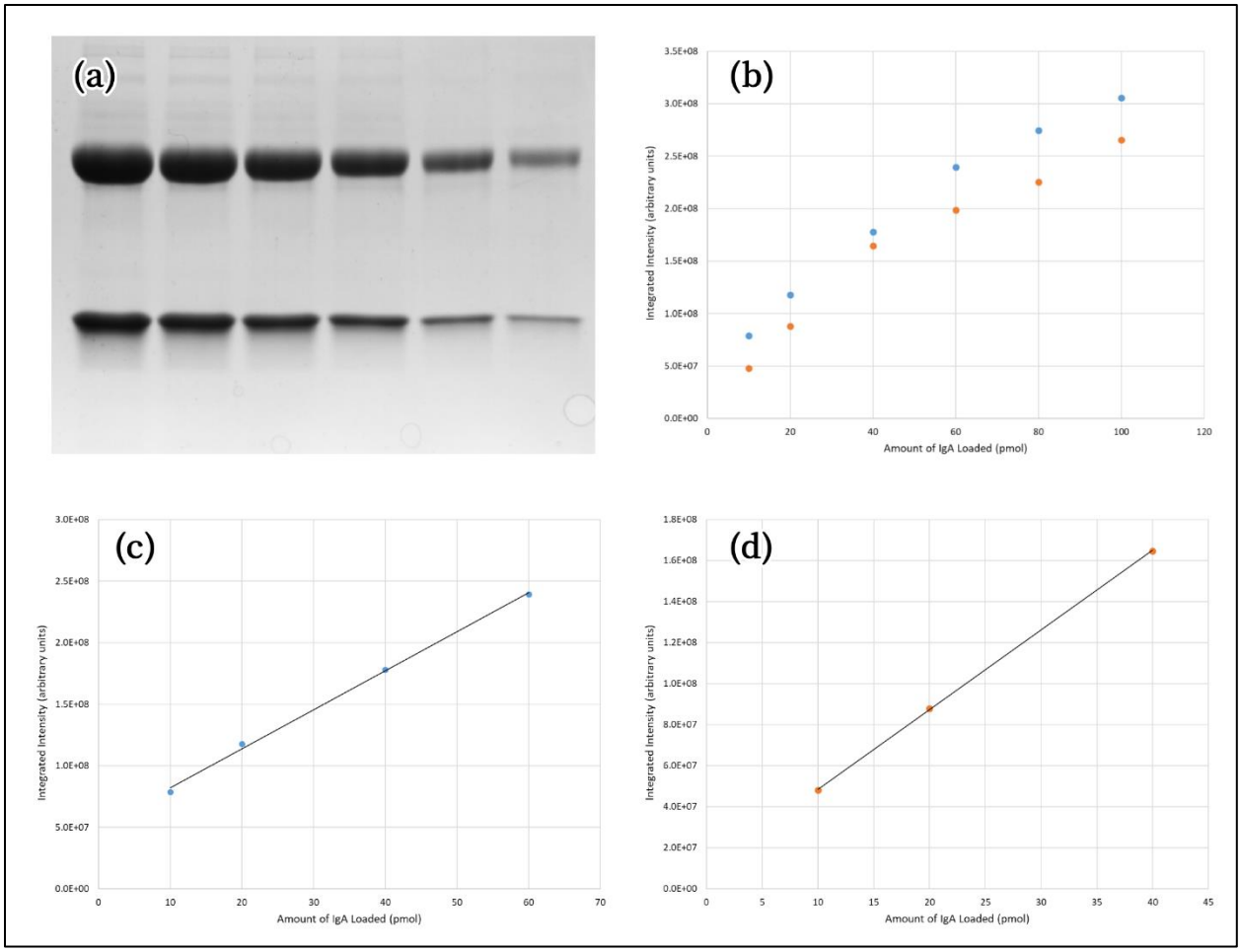


Figure 4.3.1.2 – Testing Linearity of the Gel-Based Kinetic Assay Using a Substrate Standard Curve. (a) Intact IgA1 was titrated from 10-100 pmol and loaded on an SDS-PAGE gel. **(b)** The bands for the IgA1 heavy chain (blue) and light chain (orange) were integrated and their intensities were plotted as a function of amount of IgA loaded. **(c)** The IgA1 heavy chain intensity remains linear (R^2 = 0.998) up to 60 pmol while **(d)** the intensity for the light chain remains linear (R^2 = 0.999) up to 40 pmol.

The first step of validation was to determine the functional range of the assay by fitting measurements within that range to a standard curve. Various amounts of monomeric IgA1 ranging from 10 to 100 pmol were visualized on a gel and the intensities of the HC and LC were digitally integrated for quantification (*Fig. 4.3.1.2a*). It was apparent that the intensity measurements at higher amounts of IgA1 deviated from the lower end of the curve (*Fig. 4.3.1.2b*). However, the integrated signal was linear (R^2 > 0.99) up to and including 60 pmol IgA1 (~3 μM) for the HC band and 40 pmol IgA1 (~2 μM) for the LC band (*Fig. 4.3.1.2cd*). 60 pmol (~3 μM) was thus set as the maximum amount of IgA1 that can be loaded onto a single gel lane while maintaining linearity with respect to the amount of IgA1 loaded.

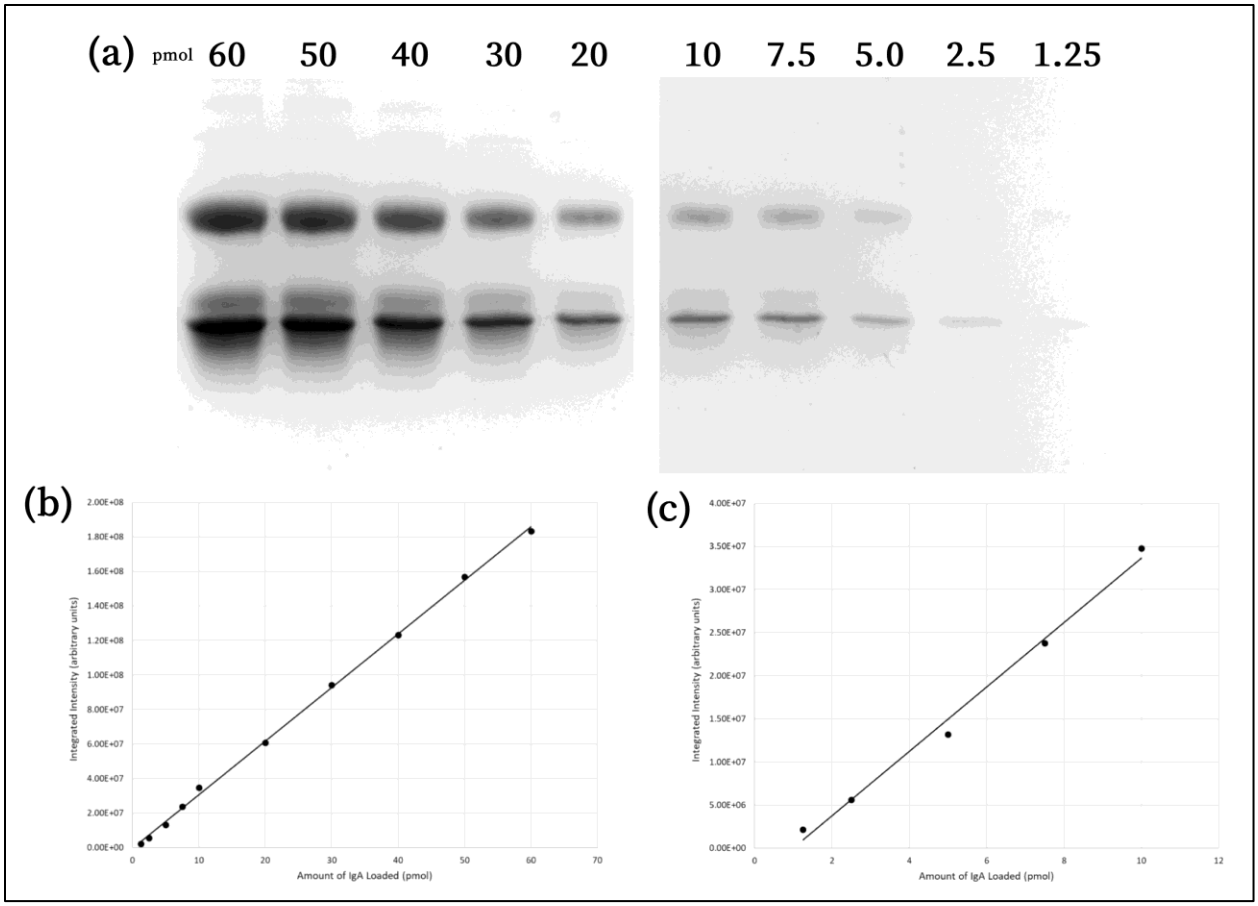


Figure 4.3.1.3 – Testing Linearity of the Gel-Based Kinetic Assay Using a Product Standard Curve. (a) IgA1 that was fully cleaved with *T. ramosa* IgAP MD was titrated from 1.25-60 pmol and loaded on an SDS-PAGE gel. **(b)** The band for the IgA1 H1 fragment was integrated and its intensities were plotted as a function of amount of IgA loaded. This standard curve shows a strong linear correlation (R^2 = 0.999) for all amounts loaded. **(c)** The same data for 1.25-10 pmol cleaved IgA1 was replotted in a separate standard curve that is more representative of intensities that will be measured within the initial-rate region of the progress curve. This curve also shows a strong linear correlation (R^2 = 0.992) for all amounts loaded.

The functional range must also account for the fact that the assay would need to be able to reliably distinguish between intensities in a reaction that progresses only 5-10%, as this is typically assumed to fulfill initial-rate conditions. Under the restrictions imposed above of a maximum loading capacity of 60 pmol IgA1 per lane, the assay would need to distinguish between intensities that fall within 3-6 pmol of cleaved IgA1. A subsequent standard curve was done where fully cleaved IgA1 was titrated from 1.25 to 60 pmol per lane (*Fig. 4.3.1.3a*). As was seen in *Fig. 4.3.1.2* for the HC, the intensity of the H1 band was linear for the entire titration range up to and including 60 pmol (*Fig. 4.3.1.3b*). This held true when the data were replotted with only the lower end of the standard curve (*Fig. 4.3.1.3c*), which was more representative of the intensity measurements that would be seen in real data under initial-rate conditions. These two standard curves suggest

that the assay can detect bands reliably within the range of 1.25 to 60 pmol of IgA loaded and that these intensity measures are strongly correlated with the corresponding amount of IgA loaded per lane. Based on these results, all SDS-PAGE samples were prepared such that they contained 40 pmol of IgA regardless of the concentration of IgA in the reaction mixture to ensure that all band intensities would fall within the functional range of the assay. This meant that reactions with substrate concentrations greater than 2 μM were diluted down to 2 μM in the SDS-PAGE sample and substrate concentrations equal to or less than 2 μM were sampled undiluted.

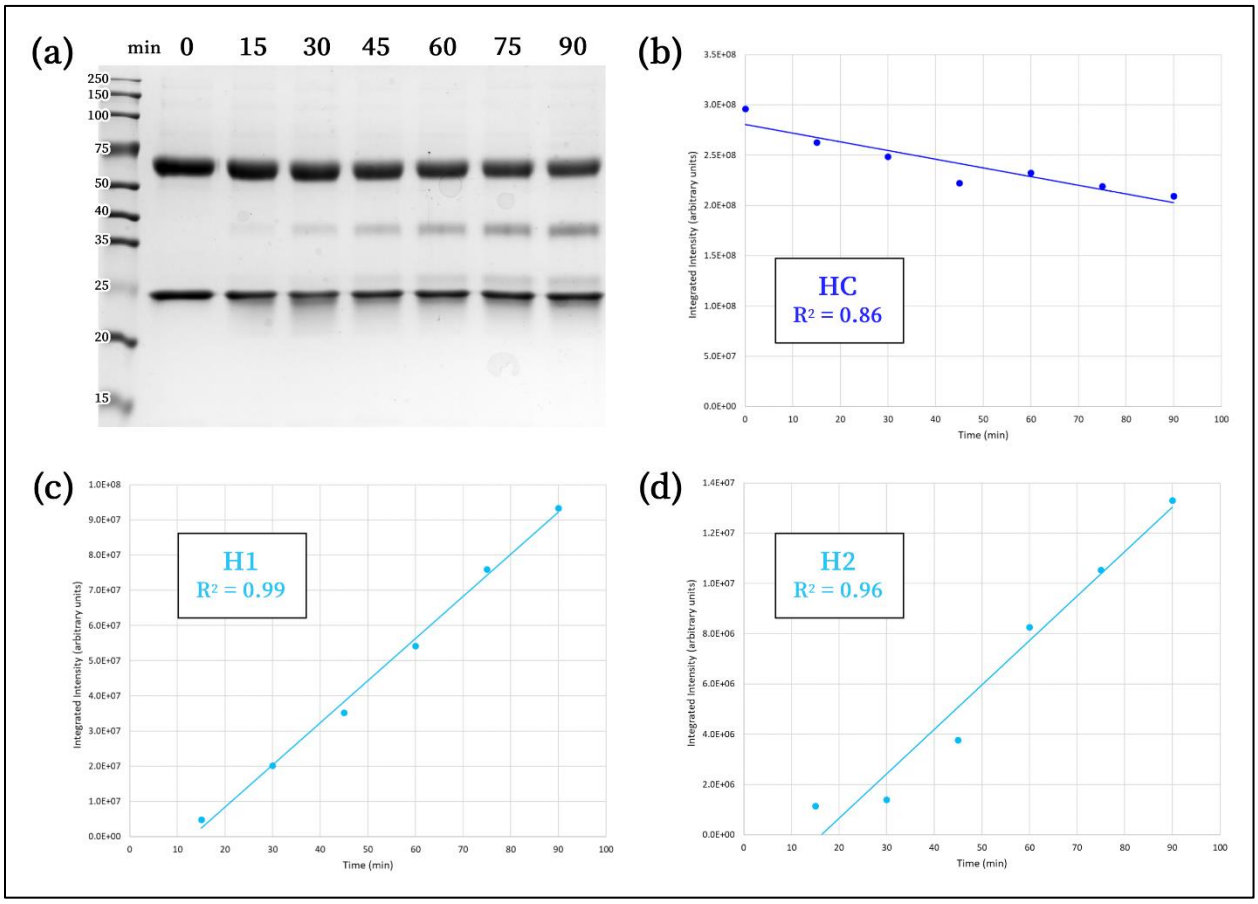


Figure 4.3.1.4 – The Cleaved Heavy-Chain Fragment Associated with the Fc Provided the Most Reliable Readout for the Gel-Based Assay. (a) 0.5 nM *T. ramosa* IgAP MD was incubated with 3 μM IgA1 at 37 ºC and sampled at 15-minute intervals. The integrated band intensities for the **(b)** heavy chain, **(c)** H1 fragment, and **(c)** H2 fragment were plotted with respect to time.

The second step of validation was to determine the best SDS-PAGE band to use for densitometric analysis. This is because the disappearance of the HC, appearance of the H1, and appearance of the H2 bands can all be theoretically used to measure the progress of the reaction. IgA1 was incubated with the *T. ramosa* IgAP MD and sampled over time to test the suitability of

all three bands (*Fig. 4.3.1.4a*). Although all bands behaved as expected over the course of the reaction, the H1 band (*Fig. 4.3.1.4c*) had the strongest linear correlation of its band intensity with respect to time and thus was the best band to use for densitometric analysis.

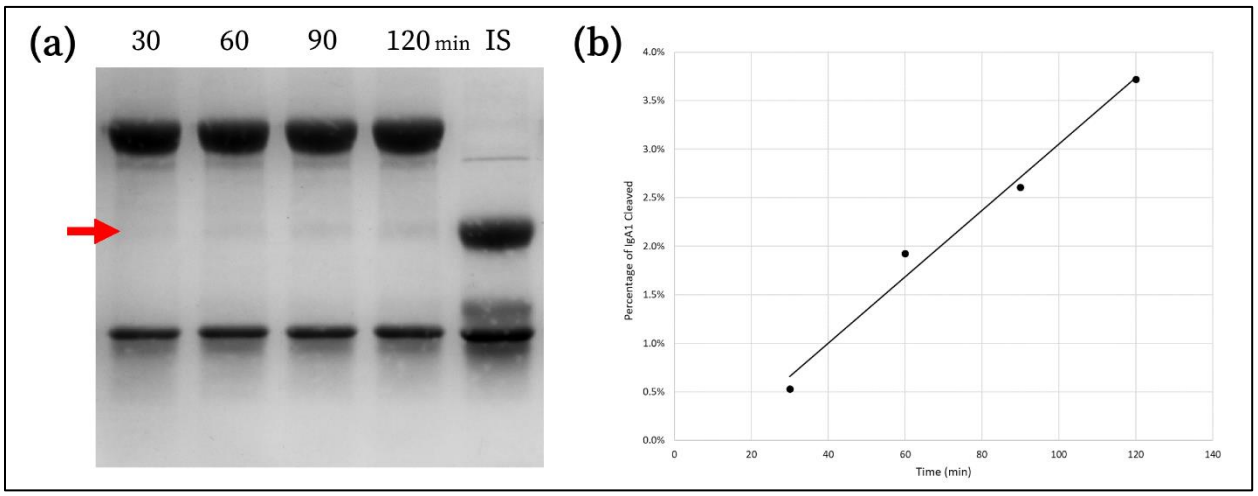


Figure 4.3.1.5 – The Gel-Based Assay is Capable of Determining Initial Rates Under Michaelis-Menten Conditions. (a) 0.75 nM *T. ramosa* IgAP NTD+MD was incubated with 50 μM IgA1 at 37 °C and sampled at 30-minute intervals. These samples were run alongside a fully cleaved 2 μM IgA1 internal standard (IS) to determine the percentage of IgA1 cleaved in each lane. **(b)** The integrated band intensities for the H1 fragment were converted to the percentage of IgA1 cleaved and plotted with respect to time. Under initial-rate conditions (<5% substrate consumed), the H1 band intensity shows a strong linear correlation with time ($R^2 = 0.98$).

It became apparent during the course of validation that an external calibration curve (*Fig. 4.3.1.3*) was not able to reproducibly determine the amount of product present during a reaction due to variations in band intensities between SDS-PAGE gels for the same sample. To combat this, an internal standard was used to normalize the H1 band to a constant band in the same gel so that results could be compared between gels.

The light chain (LC) of IgA1 was the first internal standard tested as it is theoretically proportional to the amount of IgA1 loaded into each lane and hence would account for differences in sample preparation and gel loading. It would also be subject to the same gel workup as the H1 band and thus account for differences in gel staining, destaining, and imaging. Unfortunately, the LC was an unreliable internal control due to the diffuse nature of the LC band, which was prone to error in processing because of inaccurate background subtraction. This was made worse by the fact that the H2 band overlapped with the LC in samples with a larger proportion of IgA1 cleaved.

Most of the variation in the measured band intensities between gels seemed to be from the gel workup rather than sample preparation and loading. This was supported by the fact that bands

within a gel seemed to be internally consistent without normalization (*Fig. 4.3.1.3* and *4.3.1.4c*). This is in contrast to band intensities across gels, which were susceptible to the batch-to-batch variation of SDS-PAGE stain and destain used as well as the exposure time of the image. The SDS-PAGE staining step was particularly variable due to the differing amounts of Coomassie R-250 dye that were lost to the filter paper when the stain was made.

As an alternative approach to the LC, a batch of IgA1 was fully cleaved with the *T. ramosa* IgAP, diluted to 40 pmol per sample, and frozen in 25 μL aliquots with SDS-PAGE loading dye to serve as internal standards for each gel. The intensity of the H1 band for the internal standard would be used as the reference to determine the percentage of IgA1 cleaved at a given time point. This would get around the need to use an external standard curve to convert changes in intensity to rate to avoid errors associated the gel workup. *Fig. 4.3.1.5* is a representative example of data with this new internal standard that was obtained under initial-rate conditions. The following steps were used to convert densitometry measurements into the initial rate. The H1 band intensities (*Fig. 4.3.1.5a*) were first normalized to the intensity of the internal standard to convert them into the percentage of IgA1 cleaved at each time point. This was done with the equation:

$$\%\ IgA1\ Cleaved = \frac{Integrated\ intensity\ of\ H1\ band\ in\ sample}{Integrated\ intensity\ of\ H1\ band\ in\ internal\ standard} * 100\%$$

The percentage of IgA1 cleaved was plotted with respect to time and used to calculate the slope from a simple linear regression (*Fig. 4.3.1.5b*). The unit of this slope was percentage of IgA1 cleaved per minute. Although samples from reactions with substrate concentrations greater than 2 μM were diluted to 2 μM to fall within the functional range of the assay, the value of the slope is independent of the substrate concentration as it represents the proportion of IgA1 cleaved relative to the total amount of substrate. The initial rate is therefore simply calculated with the following equation and takes on the units of molarity per minute:

$$Initial\ rate\ (v) = Slope\ of\ linear\ regression * [S]$$

It was important to verify that the measurements used to calculate the linear regression were in the initial-rate region of the progress curve. This was done by assessing the correlation coefficient between the time points to ensure strong linearity ($R^2 > 0.98$) and by making sure that the y-intercept of the linear regression was close to zero. Non-zero y-intercepts resulted from

fitting the regression to the non-linear portion of the progress curve and run contrary to the assumptions of underlying initial rates as they imply that there is some amount of product that exists at time zero.

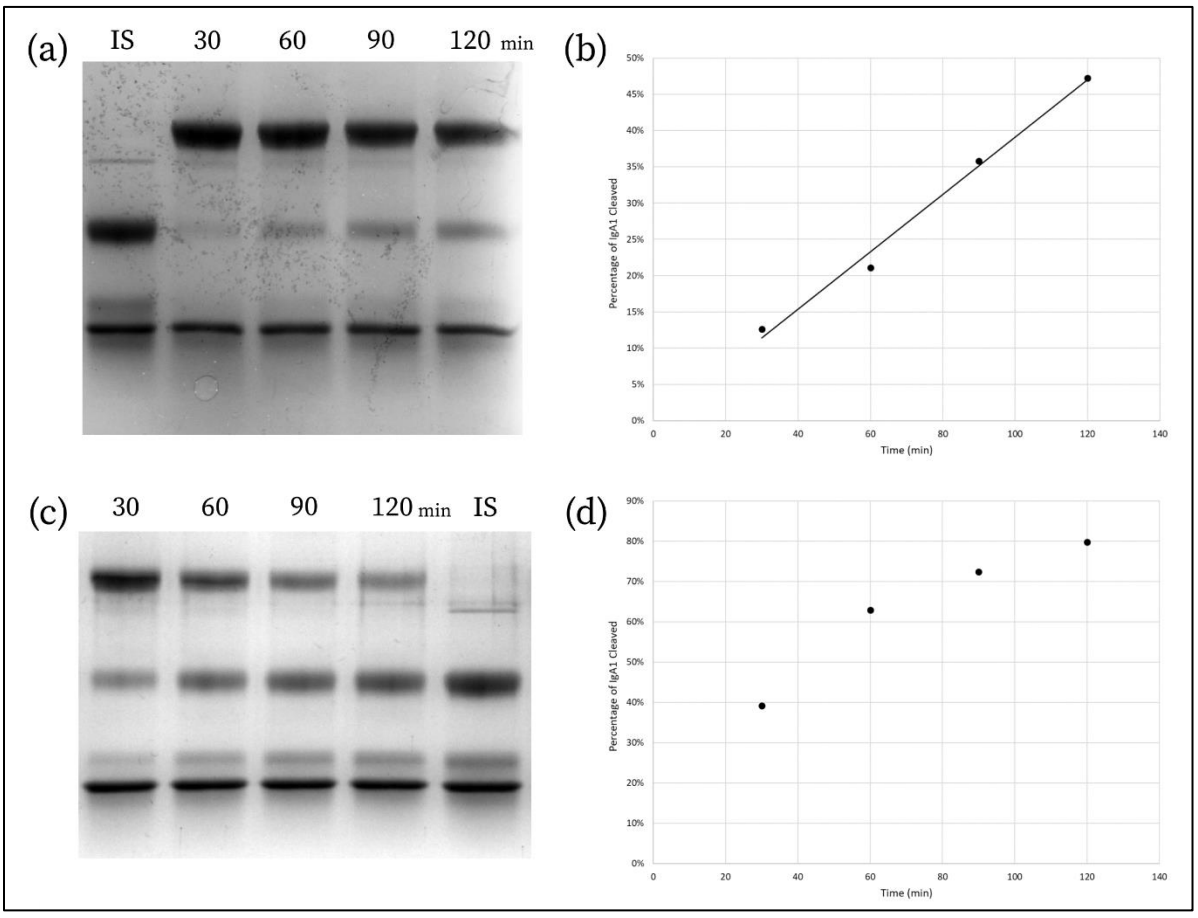


Figure 4.3.1.6 – The *Thomasclavelia ramosa* Immunoglobulin A Protease is Robust Against Product Inhibition. (a) 1.5 nM full-length *T. ramosa* IgAP was incubated with 1 μM IgA1 at 37 ºC and sampled at 30-minute intervals. These samples were run alongside a fully cleaved 2 μM IgA1 internal standard (IS) to determine the percentage of IgA1 cleaved in each lane. **(b)** The integrated band intensities for the H1 fragment were converted to the percentage of IgA1 cleaved and plotted with respect to time. The H1 band intensity shows a strong linear correlation with time (R^2 = 0.99) despite the reaction progressing significantly past what is normally considered to be the initial-rate range. **(c)** 0.5 nM *T. ramosa* IgAP MD was incubated with 1 μM IgA1 at 37 ºC and sampled at 30-minute intervals. **(d)** Unlike (a) and (b), the reaction progressed too far and is slowing down due to effects of product inhibition and/or substrate depletion.

All *T. ramosa* IgAP enzyme constructs were surprisingly robust in fulfilling these initial-rate conditions as they were not significantly affected by product inhibition (*Fig. 4.3.1.6*). This made these enzymes relatively easy to characterize as experimental conditions did not need to be fully optimized for each enzyme and substrate concentration to ensure linearity. Initial-rate conditions were still valid for a reaction that progressed to nearly 50% completion (*Fig. 4.3.1.6ab*),

which is much more than what is generally accepted for initial-rate kinetics.[22] Product inhibition and substrate depletion only began to have an effect when the reaction proceeded to near completion (*Fig. 4.3.1.6cd*).

In summary, I have presented experimental conditions where the gel-based IgAP assay does not violate the assumptions inherent to the Michaelis-Menten equation. The assay is therefore suitable for understanding the potential functions of the *T. ramosa* IgAP non-proteolytic domains in enzymatic activity. The experiments discussed in this section provide data to support the validity of the assay with respect to the criteria set out in the end of *Section 4.1.4*. These criteria have been addressed as follows, in the same order that they are introduced in that section:

(1) The enzyme concentration (0.15-1.5 nM) was over 650 times less than the lowest concentration of substrate (1 μM) used in the reaction.
(2) The assay can detect small amounts of product from reactions that have progressed as little as ~3% (1.25 pmol in a 40 pmol sample). This falls in the rule-of-thumb range where issues of product inhibition and substrate depletion are assumed to be negligible.
(3) The integrated intensity of the H1 band has a strong linear correlation ($R^2 > 0.99$) with the amount of IgA1 product loaded over a large functional range (1.25 to 60 pmol) that covers the initial-rate region as described in (2).
(4) The assay can distinguish between reactions that are in the linear and non-linear regions of the progress curve by assessing the correlation coefficient and y-intercept of the linear regression (*Fig. 4.3.1.6*).
(5) Finally, this assay uses the proteases' unlabeled biologically relevant substrate, IgA1, to make kinetic conclusions.

4.3.2 – Limitations to the Gel-Based Immunoglobulin A Protease Kinetic Assay

There are five aspects of the gel-based IgAP assay that require further discussion:

Firstly, there is no practical way to quantify experimental error in the measurements or the derived kinetic parameters. Due to the way the experiment is set up and how the reaction progress is monitored, the assay only yields one experimental measurement for each time point and substrate concentration. True replication would involve setting up multiple identical assays on different days to obtain replicates of the full Michaelis-Menten curve. Although this can be done in theory, it is not practical given the cost and amount of substrate required as well as the amount of time it takes to carry out the assay. One of the reasons why the reaction was monitored over time was to provide some confidence in the derived initial rates, as linear regression can detect the level of variation, despite the lack of a real measure of error.

Secondly, if this assay is to be used for other IgAPs or Ig proteases, the functional range of the assay may not be sufficient for proteases that are more sensitive to product inhibition. Qualitative results for the *Streptococcus pneumoniae* and *Haemophilus influenzae* IgA1Ps suggest that they bind much tighter to IgA1 as the inactive mutants of the protein can form complexes with IgA1 that can be isolated and characterized separately (data not shown). This is the basis for how the structure of the *S. pneumoniae* IgA1P-IgA1 complex was able to be determined by cryo-electron microscopy.[2] It is thus reasonable to assume that the enzymes also have some affinity for the IgA1 cleavage products based on their structural similarities to the intact IgA1. If this is the case, the assay may need to reliably detect product from reactions that progress less than 5%, which begins to fall outside of the tested range of the assay.

Preliminary semi-quantitative testing with the gel-based assay suggests that the *S. pneumoniae* and *H. influenzae* IgA1Ps are also significantly faster than the *T. ramosa* IgAP (data not shown). In the event that their k_{cat}/K_M values are much larger than the *T. ramosa* IgAP, the assay may need to use substrate concentrations much lower than 1 μM to fully define the curve, which may fall below the lower limit of detection for an SDS-PAGE gel. Initial testing using silver stain instead of Coomassie R-250 suggested that silver staining was not suitable for kinetic analysis (data not shown). This was because the gels were easily overdeveloped in a matter of seconds and thus it was not a reproducible way to ensure that the gel bands remained in the functional range of the assay.

In addition, the *H. influenzae* IgA1P cleaves IgA1 in the middle of the hinge (*Fig. 1.4.1b*) and splits the HC into two nearly-equivalent fragments. This means that H1 and H2 are no longer separable on an SDS-PAGE gel and cannot be quantified individually. This can be alleviated by pre-treating IgA1 with neuraminidase to remove negatively charged sialic acids from the IgA1 glycosylations, which changes the migration behavior of H1 and H2 such that they can be separated on a gel.[17] The effect of removing sialic acid from the substrate may impact its ability to be cleaved and hence be semi-artifactual, however, initial studies on the *Neisseria gonorrhoeae* and *N. meningitidis* IgA1Ps suggest that desialyation does not impact activity.[17] Regardless, this may limit its use towards characterizing certain IgAPs and is an additional point of consideration for the generalizability of the assay.

Finally, this gel-based kinetic assay assumes that the rate of cleavage and the enzyme's specificity are identical for a fully intact IgA1 molecule (two Fab domains) and a half-cleaved IgA1 molecule (only one Fab). This is important to note as there is only one active site per enzyme and, as such, the complete cleavage of IgA into one Fc and two Fab domains must occur in two sequential reactions. Due to the reducing and denaturing conditions of the SDS-PAGE gel, it is impossible to distinguish whether the product bands (H1 and H2) come from a half-cleaved or fully cleaved IgA1 molecule. However, the fact that the rate remains linear up to ~50% of the reaction (*Fig. 4.3.1.6b*) suggests that there may not be significant kinetic differences between the two substrates as the rate does not appreciably change under conditions where a non-negligible amount of half-cleaved IgA1 is present in solution.

4.3.3 – Kinetic Comparisons of Thomasclavelia ramosa Immunoglobulin A Protease Constructs

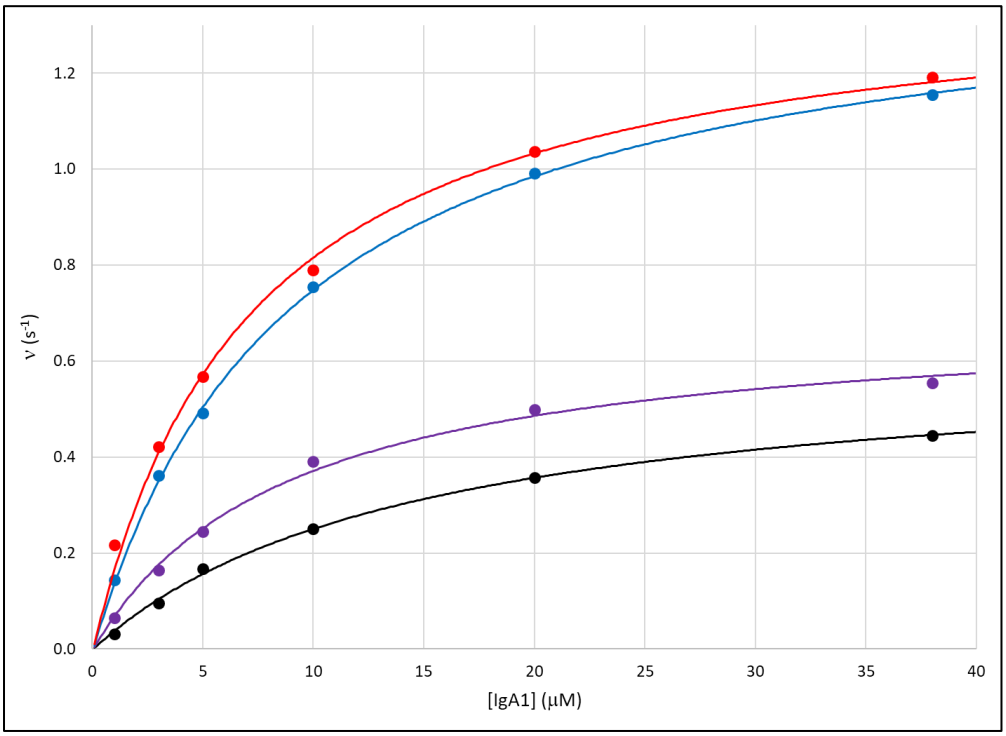


Figure 4.3.3.1 – Michaelis-Menten Curves for Various Domain Constructs. The enzyme-normalized initial rates (ν) for the FL (black), NTD+MD (purple), MD+CTD1 (blue), and MD (red) enzymes were plotted against IgA1 concentration and fit to the Michaelis-Menten equation. The Michaelis-Menten parameters that best fit these data are summarized in *Table 4.3.3.1*.

Table 4.3.3.1 – Michaelis-Menten Parameters for Various Domain Constructs. The following is a tabulation of the best-fitting Michaelis-Menten constants for the curves shown in *Fig. 4.3.3.1*. These parameters are tabulated with their associated standard error values. These error values only represent variation in the mathematical fit and not in the experimental methodology itself, due to the impracticalities of repeating the assay in replicates. Thus, the following error values likely underestimate the true error of the parameters.

Protein construct	*MD*	*MD+CTD1*	*NTD+MD*	*FL*
k_{cat}/K_M ($M^{-1}s^{-1}$)	$2x10^{-7} \pm 2x10^{-8}$	$2x10^{-7} \pm 5x10^{-9}$	$8x10^{-8} \pm 9x10^{-9}$	$4x10^{-8} \pm 3x10^{-9}$
k_{cat} (s^{-1})	1.4 ± 0.04	1.4 ± 0.02	0.70 ± 0.03	0.62 ± 0.02
K_M (μM)	7.2 ± 0.6	9.2 ± 0.3	8.9 ± 0.9	15 ± 1

The Michaelis-Menten curves and parameters for various *T. ramosa* IgAP domain constructs are shown in *Fig. 4.3.3.1* and *Table 4.3.3.1*, respectively. These kinetic data show a clear increase in both fundamental parameters as the enzyme is truncated. While large increases in these parameters were seen when the C-terminal tail (comparing NTD+MD vs. FL) and NTD (comparing NTD+MD vs. MD) were removed, only small changes were seen when CTD1 was truncated (comparing MD+CTD1 vs. MD). Whether these values are truly statistically different is not known due to practical limitations of the assay in quantifying error (*Section 4.3.2*). However, this suggests that the CTD1 does not have as much of an impact on the enzymatic cycle compared to the NTD and C-terminal tail. Removing the C-terminal tail seemed to have a pronounced effect on k_{cat}/K_M while removing the NTD impacted both k_{cat} and k_{cat}/K_M.

As described in *Section 4.1.4*, k_{cat}/K_M represents the effective rate of substrate capture into a catalytically competent state.[25] It is mainly impacted by the rate-determining step, which varies from enzyme to enzyme and can be the step of substrate binding, chemical catalysis, or product release.[27] It is hypothesized that the rate-determining step for the *T. ramosa* IgAP is substrate binding as the disorder-to-order transition that is proposed in the aforementioned working hypothesis (*Section 2.3.7*) suggests that conformational changes in and distal to the active site are necessary for catalysis. These conformational changes require large rearrangements of the substrate and enzyme which may lead to a greater entropic penalty that increases the energetic barrier of the substrate-binding transition states.

Assuming the rate-determining step does not change when the enzyme is minimized, the fact that k_{cat}/K_M increases as the enzyme is truncated suggests that the non-proteolytic domains do not contribute to substrate binding. This corroborates with the working hypothesis as the disorder-to-order transition is seen in the MD+CTD1 protein in isolation, independent of the other domains.

If the NTD and/or C-terminal tail significantly contributed to substrate binding, k_{cat}/K_M would be expected to decrease as the enzyme was truncated. The changes in catalytic efficiency are therefore postulated to be due to steric effects from the non-proteolytic domains, which lowers the frequency of forming a catalytically competent complex with the protease. If there were any energetic benefits involved in substrate binding from the non-proteolytic domains, the experimental results indicate that these may be outweighed by this steric factor. The differences in k_{cat}/K_M between MD and MD+CTD1 also support this steric hypothesis, assuming that they are statistically significant when experimental error is measured.

k_{cat} represents the rate of regenerating free enzyme after it is captured into the enzyme-substrate complex (*Section 4.1.4*).[25] Because k_{cat} is calculated from a fit of the Michaelis-Menten equation at infinite substrate concentration, under these conditions, the time it takes to form the enzyme-substrate complex is negligible and thus the substrate-binding step cannot be measured using k_{cat}.[25] The main bottlenecks that contribute to k_{cat} are thus chemical catalysis and product release.[25] There is no experimental evidence to suggest that one step contributes more than the other, however as the same trend is seen in k_{cat} as k_{cat}/K_M, one can postulate that the major contributor to k_{cat} is product release for the same reasons of steric interference preventing the product from leaving the enzyme-product complex. Because a large increase in k_{cat} was seen when the NTD was removed, the NTD may inhibit product release as it may have some affinity to the cleaved IgA products. The working model of the FL-IgA complex (*Section 3.3.4*) depicts the NTD near the substrate such that it may be able to interact with the products during the enzymatic cycle.

Unfortunately, these data in isolation are unable to make confident conclusions to support the role of the non-proteolytic domains in catalysis, if any. These data however are consistent with the working hypothesis and provide a starting point for future work on testing this hypothesis. These kinetic data also conclusively show that the *T. ramosa* IgAP has a truly modular domain architecture as it can be minimized to just the MD (~40% the size of the FL enzyme) while retaining activity. In addition, the truncated MD construct is more active *in vitro* at all assayed concentrations of IgA1 relative to the FL construct.

4.3.4 – Verifying Substrate Specificity for the Reduced Thomasclavelia ramosa Construct

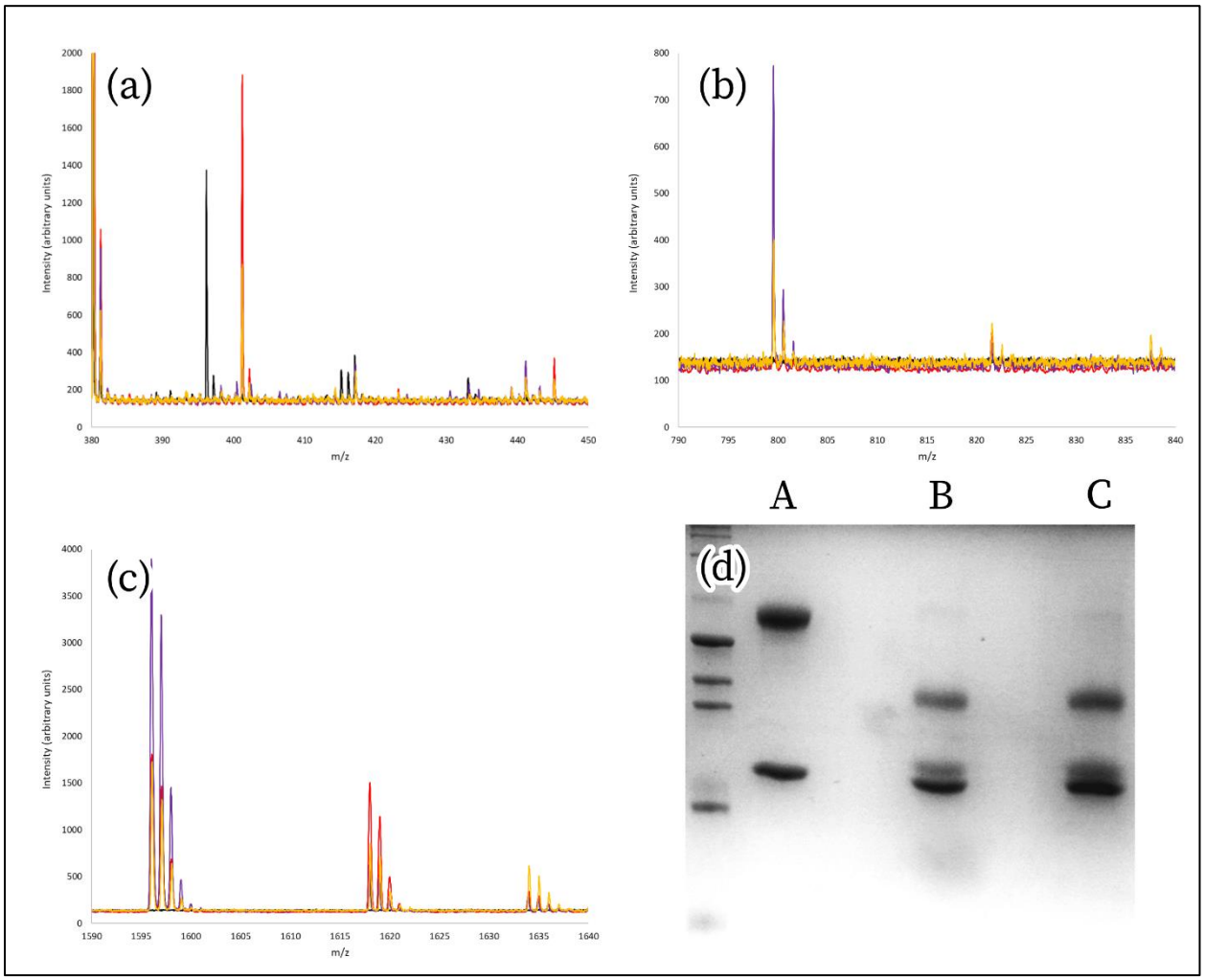


Figure 4.3.4.1 – The *Thomasclavelia ramosa* Immunoglobulin A Protease Middle Domain Does Not Nonspecifically Cleave the IgA1 Hinge Peptide. MALDI-ToF mass spectrometry was used to detect the masses of intact (798.95 g/mol) and cleaved peptide (402.5 and 414.5 g/mol) after incubation with various *T. ramosa* IgAP enzyme constructs. **(a)** No peaks corresponding to the cleaved peptide were seen in the samples when the peptide was mixed with the full-length enzyme (orange) or the MD (purple). All other peaks could be accounted for in the matrix control (black) or peptide-only control (red). **(b)** Peaks corresponding to the intact peptide were seen in all samples but the matrix control. The intact peptide also showed peaks in complex with Na^+ (~23 g/mol) or K^+ (~39 g/mol). **(c)** Peaks corresponding to an induced intact peptide dimer were seen in all samples but the matrix control alongside their complexes with Na^+ and K^+. **(d)** 2 μM IgA1 (A) was incubated with ~1.3 nM MD (B) or full-length enzyme (C) at the same molar ratio of substrate to enzyme and identical reaction conditions as the mass-spectrometry samples. All of the IgA1 was cleaved under these conditions.

Structural and functional characterization of the various *T. ramosa* IgAP enzyme constructs confirm that the non-proteolytic domains are not necessary for activity of the protease domain. Despite this, it was important to verify that the minimized MD construct still retained high substrate specificity in the event that removing all other non-proteolytic domains led to a more

promiscuous enzyme without impacting its ability to cleave IgA. An octapeptide based on the IgA1 hinge sequence surrounding the *T. ramosa* IgAP cleavage site (N-VPCPVPST-C) was used to probe for this. This octapeptide was incubated with the full-length and MD enzyme constructs and MALDI-ToF mass spectrometry was used to detect the mass of peptides in the reaction mixture (*Fig. 4.3.4.1*).

There was no mass-spectrometric evidence that any peptide was cleaved by either construct as no peaks were seen at the expected m/z range of the cleavage products that were not in the controls (*Fig. 4.3.4.1a*). Furthermore, the intact peptide showed strong signals at m/z ranges corresponding to a monomer and matrix-induced dimer (*Fig. 4.3.4.1bc*),[32] suggesting that the lack of detection was not due to a problem with the sample workup. In contrast, all of the IgA1 was cleaved when the reaction was scaled up and reproduced with the same incubation conditions and molar ratio of substrate to enzyme as the peptide reactions (*Fig. 4.3.4.1d*).

4.4 – Chapter Summary and Next Steps

This chapter focused on understanding the *T. ramosa* IgAP from an enzymological perspective to bridge the gap between structure and function. *Chapter 2* and *Chapter 3* described the crystal and solution structures of individual and multiple domains and provided a working hypothesis for how the enzyme recognizes and cleaves IgA. This working hypothesis was tested using a discontinuous gel-based kinetic assay that was inspired by a similar assay using radiolabeled IgA.[15] Through many iterations in method development, the gel-based assay was validated by demonstrating that it followed the assumptions inherent to the Michaelis-Menten equation. This allowed for kinetic comparisons of Michaelis-Menten parameters between enzyme constructs to begin to understand the functional roles of each *T. ramosa* IgAP domain in the enzymatic cycle.

These kinetic data suggest that the NTD and CTDs sterically hinder the enzyme from forming the enzyme-substrate complex. The NTD is also hypothesized to have some affinity towards the IgA cleavage products as it may affect the rate of product release. However, more direct experimental evidence is needed to understand the biological function of these domains. This can be done through mutation of the substrate or enzyme at sites postulated to be involved in substrate binding and/or product release and probing for their functional consequences via this gel-

based kinetic assay. Regardless, these data highlight the modular nature of the *T. ramosa* IgAP as its activity is not only retained but is improved upon truncation to the smallest enzyme construct.

4.5 – Chapter References

1. Mistry, D. & Stockley, R. A. IgA1 protease. *Int J Biochem Cell Biol* **38**, 1244–1248 (2006).
2. Wang, Z., Rahkola, J., Redzic, J. S., Chi, Y. C., Tran, N., Holyoak, T., Zheng, H., Janoff, E. & Eisenmesser, E. Mechanism and inhibition of Streptococcus pneumoniae IgA1 protease. *Nat Commun* **11**, (2020).
3. Redzic, J. S., Rahkola, J., Tran, N., Holyoak, T., Lee, E., Martín-Galiano, A. J., Meyer, N., Zheng, H. & Eisenmesser, E. A substrate-induced gating mechanism is conserved among Gram-positive IgA1 metalloproteases. *Commun Biol* **5**, (2022).
4. Johnson, T. A., Qiu, J., Plaut, A. G. & Holyoak, T. Active site gating regulates substrate selectivity in a chymotrypsin-like serine protease. The structure of Haemophilus influenzae IgA1 protease. *J Mol Biol* **389**, 559–574 (2009).
5. Burton, J., Wood, S. G., Lynch, M. & Plaut, A. G. Substrate Analogue Inhibitors of the IgA1 Proteinases from Neisseria Gonorrhoeae. *J Med Chem* **31**, 1647–1651 (1988).
6. Wood, S. G., Lynch, M., Plaut, A. G. & Burton, J. Tetrapeptide Inhibitors of the IgA1 Proteinases from Type I Neisseria gonorrhoeae. *J Med Chem* **32**, 2407–2411 (1989).
7. Garner, A. L., Fullagar, J. L., Day, J. A., Cohen, S. M. & Janda, K. D. Development of a high-throughput screen and its use in the discovery of Streptococcus pneumoniae immunoglobulin A1 protease inhibitors. *J Am Chem Soc* **135**, 10014–10017 (2013).
8. Choudary, S. K., Qiu, J., Plaut, A. G. & Kritzer, J. A. Versatile Substrates and Probes for IgA1 Protease Activity. *ChemBioChem* **14**, 2007–2012 (2013).
9. Shehaj, L., Choudary, S. K., Makwana, K. M., Gallo, M. C., Murphy, T. F. & Kritzer, J. A. Small-Molecule Inhibitors of Haemophilus influenzae IgA1 Protease. *ACS Infect Dis* **5**, 1129–1138 (2019).
10. Bachovchin, W. W., Plaut, A. G., Flentke, G. R., Lynch, M. & Kettner, C. A. Inhibition of IgA1 proteinases from Neisseria gonorrhoeae and Hemophilus influenzae by peptide prolyl boronic acids. *Journal of Biological Chemistry* **265**, 3738–3743 (1990).
11. Plaut, A. G., Genco, R. J. & Tomasi, T. B. Isolation of an Enzyme from Streptococcus sanguis Which Specifically Cleaves IgA. *J Immunol* **113**, 289–291 (1974).
12. Fujiyama, Y., Kobayashi, K., Senda, S., Benno, Y., Bamba, T. & Hosoda, S. A Novel IgA Protease from Clostridium sp. Capable of Cleaving IgA1 and IgA2 A2m(1) but not IgA2 A2m(2) allotype paraproteins. *The Journal of Immunology* **134**, 573–576 (1985).
13. Kosowska, K., Reinholdt, J., Rasmussen, L. K., Sabat, A., Potempa, J., Kilian, M. & Poulsen, K. The Clostridium ramosum IgA Proteinase Represents a Novel Type of Metalloendopeptidase. *J Biol Chem* **277**, 11987–11994 (2002).

14. Steinrücke, P., Aldinger, U., Hill, O., Hillisch, A., Basch, R. & Diekmann, S. Design of helical proteins for real-time endoprotease assays. *Anal Biochem* **286**, 26–34 (2000).

15. Reinholdt, J. & Kilian, M. A Sensitive Enzyme-Linked Immunosorbent Assay for IgA Protease Activity. *J Immunol Methods* **63**, 367–376 (1983).

16. Plaut, A. G., Gilbert, J. V. & Heller, I. Assay and Properties of IgA Protease of Streptococcus sanguis. in *Secretory Immunity and Infection* (eds. McGhee, J. R., Mestecky, J. & Babb, J. L.) 489–501 (Plenum Press, 1978).

17. Gilbert, J. V, Plaut, A. G., Longmaid, B. & Lamm, M. E. Inhibition of Microbial IgA Proteases by Human Secretory IgA and Serum. *Mol Immunol* **20**, 1039–1049 (1983).

18. Bachovchin, W. W., Plaut, A. G., Flentke, G. R., Lynch, M. & Kettner, C. A. Inhibition of IgA1 proteinases from Neisseria gonorrhoeae and Hemophilus influenzae by peptide prolyl boronic acids. *Journal of Biological Chemistry* **265**, 3738–3743 (1990).

19. Greenwood, F. C., Hunter, W. M. & Glover, J. S. The Preparation of 131I-Labelled Human Growth Hormone of High Specific Radioactivity. *J Biochem* **89**, 114–123 (1963).

20. Gupta, S., Batra, S. & Jain, M. Antibody Labeling with Radioiodine and Radiometals. *Methods of Molecular Biology* **1141**, 147–157 (2014).

21. Reinholdt, J. A method for titration of inhibiting antibodies to bacterial immunoglobulin A1 proteases in human serum and secretions. *J Immunol Methods* **191**, 39–48 (1996).

22. Srinivasan, B. A guide to the Michaelis–Menten equation: steady state and beyond. *FEBS Journal* **289**, 6086–6098 (2022).

23. Schnell, S. Validity of the Michaelis-Menten equation - steady-state or reactant stationary assumption: that is the question. *FEBS Journal* **281**, 464–472 (2014).

24. Deichmann, U., Schuster, S., Mazat, J. P. & Cornish-Bowden, A. Commemorating the 1913 Michaelis-Menten paper Die Kinetik der Invertinwirkung: three perspectives. *FEBS Journal* **281**, 435–463 (2014).

25. Northrop, D. B. On the Meaning of Km and V/K in Enzyme Kinetics. *J Chem Educ* **75**, 1153 (1998).

26. Dalziel, K. Initial Steady State Velocities in the Evaluation of Enzyme-Coenzyme-Substrate Reaction Mechanisms. *Acta Chem Scand* **11**, 1706–1723 (1957).

27. Park, C. Visual Interpretation of the Meaning of kcat/Km in Enzyme Kinetics. *J Chem Educ* **99**, 2556–2562 (2022).

28. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D. & Bairoch, A. Protein Analysis Tools on the ExPASy Server. in *The Proteomics Protocols Handbook* (ed. Walker, J. M.) 571–607 (Humana Press Inc., Totowa, 2005).

29. Walker, S. G., Carnu, O. I., Tüter, G. & Ryan, M. E. The immunoglobulin A1 proteinase from Streptococcus pneumoniae is inhibited by tetracycline compounds. *FEMS Immunol Med Microbiol* **48**, 218–222 (2006).

30. Woof, J. M. & Kerr, M. A. The function of immunoglobulin A in immunity. *J Pathol* **208**, 270–282 (2006).

31. Yoo, E. M. & Morrison, S. L. IgA: An immune glycoprotein. *Clinical Immunology* **116**, 3–10 (2005).

32. Leopold, J., Popkova, Y., Engel, K. M. & Schiller, J. Recent Developments of Useful MALDI Matrices for the Mass Spectrometric Characterization of Lipids. *Biomolecules* **8**, (2018).

Chapter 5 – Concluding Remarks and Future Directions

5.1 – Role of Structural Biology and Enzymology in Understanding Immunoglobulin A Proteases

Ever since their discovery in the early 1970's,[1] the study of immunoglobulin A proteases (IgAPs) continues to provide multifaceted insight into a variety of distally related subjects, ranging from adaptive mucosal immunity[2] and host-microorganism relationships[3–5] to the convergent evolution of enzyme structure[6–8] and substrate specificity.[9] In addition to furthering the academic understanding of these fields, this research has served as the foundation for applied IgAP research that aims to find new drugs to treat mucosal pathogens,[10–12] in order to overcome antibiotic resistance,[13–19] and to use IgAPs themselves as protein-based therapeutics for antibody-deposition diseases.[20–22] The importance of this basic research can therefore not be overstated as the success of these applied studies is reliant on the quality and rigorousness of this fundamental research.

The advent of improved protein structural-prediction software has significantly lowered the barrier of entry for investigating new avenues of research into a countless number of protein systems.[23] IgAPs are no exception to this. This thesis focused furthering our basic understanding of the IgAP from *Thomasclavelia ramosa* using structural-prediction models as a starting point as it had yet to be biochemically characterized since its initial discovery.[24] This protease is a representative member of the last of three known IgAP families that has yet to be investigated under an in-depth structural biology and structural enzymology research scope. This thesis thus aimed to methodically determine the structures of the *T. ramosa* IgAP and postulate on their structural mechanisms in order to deepen our fundamental understanding of IgAPs and further the goals of clinically relevant IgAP research. This was done using the three major techniques of X-ray crystallography, small-angle X-ray scattering (SAXS), and Michaelis-Menten kinetic assays, which were the focus of their respective chapters.

5.2 – Summary of Experimental Findings

The *T. ramosa* IgAP was first probed using X-ray crystallography in *Chapter 2* to understand the structure and function of each of its independently folded domains. The enzyme was first split into six domains (*Section 2.3.1*), which were recombinantly expressed and purified from a variety of single- and multi-domain expression constructs to sufficient purity for downstream biochemical analyses.

The crystal structure of the NTD was determined and discussed in the context of structurally homologous viral proteins to postulate on its unknown function (*Section 2.3.3*).[25] Its potential viral origin highlighted a beta-hairpin structure that may be an important surface for ligand interaction.[25] Coupled with the kinetic (*Section 4.3.3*) and SAXS (*Section 3.3.3* and *3.3.4*) data on the NTD+MD construct, the NTD was hypothesized to play a role in substrate and/or product recognition due to its proximity to IgA, IgA2m(1) in particular, in the working model of the enzyme-substrate complex.

The crystal structure of the MD+CTD1 was determined and discussed with regards to its ability to bind zinc at two positions, its enzymatic mechanism, and its unusual locally disordered active site (*Section 2.3.4* to *2.3.6*). These data led to a working model for how the *T. ramosa* IgAP recognizes and is specific for IgA (*Section 2.3.7*). Briefly, the working model suggests that IgA is first recognized by structural features distally located on the non-prime surface of the active site. This interaction stabilizes the naturally malformed active-site zinc, residues, and peptide-binding surface such that it can now accept and proteolyze the IgA hinge. The enzyme's specificity for IgA is reliant on this active-site disorder to prevent non-specific substrates from binding in the absence of these distal interactions.

This working model was used to compare the structures of the *T. ramosa* IgAP MD to other members of the bioinformatically classified M64 protease family. This analysis strongly suggested that these other M64 proteases, although annotated as IgAPs,[26] do not have IgAP activity due to steric clashes with how IgA is proposed to bind to the enzyme (*Section 2.3.8*). This served as a general warning against using bioinformatic classifications to make biochemical conclusions without accompanying experimental evidence.

SAXS was then used as a complimentary structural method to determine the solution structures of *T. ramosa* IgAP constructs that could not be analyzed using crystallography due to their inherent flexibility or inability to form protein crystals (*Chapter 3*). SAXS is a powerful tool for understanding the *T. ramosa* IgAP in particular due to its modular domain architecture, which is distinct from the structures of the other two IgAP families.[27,28] Solution-scattering data confirmed that the NTD, MD, and CTD2 are true domains that can fold when isolated from the rest of the enzyme (*Section 3.3.2* and *3.3.3*). The solution structures of CTD2 and CTD4, by inference using the CTD1-4 solution structure, were unique amongst all of the domains due to their elongated shape, which were likely stabilized by binding zinc to residues that were found between their beta sheets (*Section 2.3.1*). This elongated shape was also seen in the solution conformations of the entire C-terminal tail (CTD1-4) and consequently the full-length enzyme (*Section 3.3.3*). Its shape was hypothesized to be a biological mechanism that helped distance and display the protease domain away from the surface of the bacterium.[29] The working model of the substrate-enzyme complex was amended with these new SAXS data and revealed that the C-terminal tail extended away from the substrate in such a way that precluded it from being involved in substrate recognition (*Section 3.3.4*).

Finally, an SDS-PAGE-based discontinuous kinetic assay was developed and validated from the ground up to bridge the gap between IgAP structure and function (*Chapter 4*). This assay was designed to ensure that the assumptions underlying the Michaelis-Menten equation remained valid in order to extract kinetic parameters that could be meaningfully compared between enzyme constructs (*Section 4.3.1*). The fundamental Michaelis-Menten kinetic constants, k_{cat} and k_{cat}/K_M,[30] were experimentally determined for the NTD+MD, MD, MD+CTD1, and FL enzymes and were compared to look for functional trends as the enzyme was truncated to its minimally sized active domain (*Section 4.3.3*). These results corroborated with the crystallography and solution-scattering data to suggest that not only can the structure of the *T. ramosa* IgAP be minimized to the MD to produce a well-folded protein, but the proteolytic activity of said domain is retained and improves upon minimization. This minimized MD construct also does not lose its high substrate specificity as it, like the full-length enzyme, does not cut the IgA1 hinge peptide in isolation (*Section 4.3.4*).

5.3 – Impact of Work and Next Steps

This thesis provides the first in-depth structural and functional analysis of the biochemical workings of the *T. ramosa* IgAP, a representative enzyme of the third and least-understood IgAP family. Its structure has been experimentally validated from both a crystallographic and solution-scattering approach to generate a working model on the structure and conformation of the protease in solution, how it binds substrate with high specificity, and what each domain potentially does.

In light of recent breakthroughs in protein structural prediction, there has been a tendency and sentiment to over-rely on these predictive models as a pseudo-replacement to experimental work. However, the importance of experimental model validation is evident in the findings of this thesis. The working model of the enzyme-substrate complex for the *T. ramosa* IgAP would not have been hypothesized without observing the disordered active site in the MD+CTD1 crystal structures. The potential misclassification of all M64 proteases as IgAPs is also a symptom of a lack of experimental validation. Finally, the conformation of the C-terminal tail for the full-length enzyme was incorrectly predicted in a collapsed state whereas the SAXS data showed that it was in fact in an extended conformation, which likely has biological implications.

The kinetic assay developed in *Chapter 4* is one of the very few IgAP assays, with the exception of the original radioactive IgA assay,[31] that is able to derive meaningful Michaelis-Menten parameters that can be used to test structural hypotheses and compare between IgAP constructs. Although monomeric IgA1 was used for initial validation and enzyme characterization, this assay can easily be adapted to IgA2m(1), secretory IgA, and even other Igs to assay proteases that cleave other Ig classes.[32] IgA2m(1) and secretory IgA are pertinent biological substrates that were not investigated in this thesis. There may be kinetic differences between the various truncated constructs in their ability to cleave IgA2m(1) versus IgA1, which may reveal information on the role of the other domains, particularly the NTD, in proteolysis.

There are, of course, many other open questions that require further investigation. The following is lists some of these questions in no particular order:

As CTD3 was unable to be recombinantly produced in isolation with the current construct and purification protocol, it was not investigated in this thesis. The fact that it is bioinformatically annotated as an X2 carbohydrate-binding domain (pfam03442)[33] suggests that it may be involved

in anchoring to the cell surface after secretion.[34] This, however, needs to be experimentally validated, after it is successfully recombinantly produced and purified, by identifying ligands that it interacts with. Along the same lines, potential ligands for the NTD can also be hypothesized through a more thorough investigation of the beta-hairpin loop that may be involved in substrate interaction. As CTD3 is found in the C-terminal tail flanked by CTD2 and CTD4, it is also possible that CTD3 functions as a periscope domain due to its structural homology to a group of Ig-like folds often found in tandem with other periscope domains.[35]

The role of the second zinc-binding site in the *T. ramosa* IgAP MD still remains a mystery. Now that crystallization conditions for the MD+CTD1 have been determined and optimized, future work can focus on understanding the interplay between the two sites by collecting crystal datasets at various soaked concentrations of zinc to look at the effect of metal titration. This experimental method can also be used in the context of other M64 proteases to understand the role of the second zinc as a general biochemical phenomenon.

To further test the working model, the IgA1 hinge peptide and/or the IgA1 Fab with the hinge can be used to form a complex with the MD in an attempt to determine a crystal structure of the substrate-enzyme complex. This was previously attempted by mixing the wild type MD+CTD1 with the IgA1 hinge peptide, as the MD does not cut the peptide in isolation, however there was no density to support peptide binding in the crystal structure. Given what is now known about the disordered active site in the wild type enzyme, it is reasonable to try this again using the MD+CTD1 E540A mutant to see if a complex can be formed.

To provide more evidence to support the hypothesized role of zinc in the rigid CTD2 and CTD4 domains, SAXS data can be collected on metal-chelated samples of these constructs to test if the protein begins to unfold or gain flexibility as a result of the chelation. Although attempts were unsuccessful in this thesis, these domains can also be subject to crystallization screening to confirm their biophysical and biochemical properties using crystallography.

Overall, this thesis greatly expands upon what is known about the structure, function, and activity of IgAPs. My work will provide a concrete experimental foundation for future research on IgAP inhibitors, IgAPs as a protein-based therapeutic, and the M64 family of proteases, among many other applications.

5.4 – Chapter References

1. Mehta, S. K., Plaut, A. G., Calvanico, N. J. & Tomasi, T. B. Human Immunoglobulin A: Production of an Fc Fragment by an Enteric Microbial Proteolytic Enzyme. *J Immunol* **111**, 1274–1276 (1973).
2. Chaplin, D. D. Overview of the Immune Response. *J Allergy Clin Immunol* **125**, S3-23 (2010).
3. Mulks, M. H., Kornfeld, S. J. & Plaut, a G. Specific proteolysis of human IgA by Streptococcus pneumoniae and Haemophilus influenzae. *J Infect Dis* (1980).
4. Kilian, M., Mestecky, J., Kulhavy, R., Tomana, M. & Butler, W. T. IgA1 proteases from Haemophilus influenzae, Streptococcus pneumoniae, Neisseria meningitidis, and Streptococcus sanguis: comparative immunochemical studies. *The Journal of Immunology* **124**, 2596–2600 (1980).
5. Casadevall, A. & Pirofski, L.-A. Host-Pathogen Interactions: Basic Concepts of Microbial Commensalism, Colonization, Infection, and Disease. *Infect Immun* **68**, 6511–6518 (2000).
6. Rawlings, N. D., Barrett, A. J. & Bateman, A. MEROPS: The database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res* **40**, (2012).
7. Rawlings, N. D. & Bateman, A. How to use the MEROPS database and website to help understand peptidase specificity. *Protein Science* **30**, 83–92 (2021).
8. Mistry, D. & Stockley, R. A. IgA1 protease. *Int J Biochem Cell Biol* **38**, 1244–1248 (2006).
9. Plaut, A. G., Genco, R. J. & Tomasi, T. B. Isolation of an Enzyme from Streptococcus sanguis Which Specifically Cleaves IgA. *J Immunol* **113**, 289–291 (1974).
10. Male, C. J. Immunoglobulin A1 Protease Production by Haemophilus influenzae and Streptococcus pneumoniae. *Infect Immun* **26**, 254–261 (1979).
11. Troeger, C., Blacker, B., Khalil, I. A., Rao, P. C., Cao, J., Zimsen, S. R. M., Albertson, S. B., Deshpande, A., Farag, T., Abebe, Z. *et al.* Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Infect Dis* **18**, 1191–1210 (2018).
12. Brooks, L. R. K. & Mias, G. I. Streptococcus pneumoniae's Virulence and Host Immunity: Aging, Diagnostics, and Prevention. *Front Immunol* **9**, (2018).
13. Bachovchin, W. W., Plaut, A. G., Flentke, G. R., Lynch, M. & Kettner, C. A. Inhibition of IgA1 Proteinases from Neisseria gonorrhoeae and Haemophilus influenzae by Peptide Prolyl Boronic Acids. *Journal of Biological Chemistry* **265**, 3738–3743 (1990).
14. Burton, J., Wood, S. G., Lynch, M. & Plaut, A. G. Substrate Analogue Inhibitors of the IgA1 Proteinases from Neisseria gonorrhoeae. *J Med Chem* **31**, 1647–1651 (1988).

15. Wood, S. G., Lynch, M., Plaut, A. G. & Burton, J. Tetrapeptide Inhibitors of the IgA1 Proteinases from Type I Neisseria gonorrhoeae. *J Med Chem* **32**, 2407–2411 (1989).

16. Garner, A. L., Fullagar, J. L., Day, J. A., Cohen, S. M. & Janda, K. D. Development of a high-throughput screen and its use in the discovery of Streptococcus pneumoniae immunoglobulin A1 protease inhibitors. *J Am Chem Soc* **135**, 10014–10017 (2013).

17. Choudary, S. K., Qiu, J., Plaut, A. G. & Kritzer, J. A. Versatile Substrates and Probes for IgA1 Protease Activity. *ChemBioChem* **14**, 2007–2012 (2013).

18. Shehaj, L., Choudary, S. K., Makwana, K. M., Gallo, M. C., Murphy, T. F. & Kritzer, J. A. Small-Molecule Inhibitors of Haemophilus influenzae IgA1 Protease. *ACS Infect Dis* **5**, 1129–1138 (2019).

19. World Health Organization. *Prioritization of Pathogens to Guide Discovery, Research and Development of New Antibiotics for Drug-Resistant Bacterial Infections, Including Tuberculosis*.

20. Lamm, M. E., Emancipator, S. N., Robinson, J. K., Yamashita, M., Fujioka, H., Qiu, J. & Plaut, A. G. Microbial IgA protease removes IgA immune complexes from mouse glomeruli in vivo: Potential therapy for IgA nephropathy. *Am J Pathol* **172**, 31–36 (2008).

21. Xie, L.-S., Huang, J., Qin, W. & Fan, J.-M. Immunoglobulin A1 protease: A new therapeutic candidate for immunoglobulin A nephropathy. *Nephrology* **15**, 584–586 (2010).

22. Julian, B. A., Waldo, F. B., Rifai, A. & Mestecky, J. IgA Nephropathy, the Most Common Glomerulonephritis Worldwide - A Neglected Disease in the United States? *Am J Med* **84**, 129–132 (1988).

23. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

24. Fujiyama, Y., Kobayashi, K., Senda, S., Benno, Y., Bamba, T. & Hosoda, S. A Novel IgA Protease from Clostridium sp. Capable of Cleaving IgA1 and IgA2 A2m(1) but not IgA2 A2m(2) allotype paraproteins. *The Journal of Immunology* **134**, 573–576 (1985).

25. Baykov, I. K., Tikunov, A. Y., Babkin, I. V., Fedorets, V. A., Zhirakovskaia, E. V. & Tikunova, N. V. Tentaclins—A Novel Family of Phage Receptor-Binding Proteins That Can Be Hypermutated by DGR Systems. *Int J Mol Sci* **24**, (2023).

26. Cerdà-Costa, N. & Gomis-Rüth, F. X. Architecture and function of metallopeptidase catalytic domains. *Protein Science* **23**, 123–144 (2014).

27. Wang, Z., Rahkola, J., Redzic, J. S., Chi, Y. C., Tran, N., Holyoak, T., Zheng, H., Janoff, E. & Eisenmesser, E. Mechanism and inhibition of Streptococcus pneumoniae IgA1 protease. *Nat Commun* **11**, (2020).

28. Johnson, T. A., Qiu, J., Plaut, A. G. & Holyoak, T. Active site gating regulates substrate selectivity in a chymotrypsin-like serine protease. The structure of Haemophilus influenzae IgA1 protease. *J Mol Biol* **389**, 559–574 (2009).

29. Whelan, F., Lafita, A., Gilburt, J., Dégut, C., Griffiths, S. C., Jenkins, H. T., St John, A. N., Paci, E., Moir, J. W. B., Plevin, M. J. *et al.* Periscope Proteins are variable-length regulators of bacterial cell surface interactions. *Proc Natl Acad Sci U S A* **118**, (2021).

30. Northrop, D. B. On the Meaning of Km and V/K in Enzyme Kinetics. *J Chem Educ* **75**, 1153 (1998).

31. Reinholdt, J. & Kilian, M. A Sensitive Enzyme-Linked Immunosorbent Assay for IgA Protease Activity. *J Immunol Methods* **63**, 367–376 (1983).

32. Brezski, R. J. & Jordan, R. E. Cleavage of IgGs by proteases associated with invasive diseases. *MAbs* **2**, 212–220 (2010).

33. Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G. A., Bileschi, M. L., Bork, P., Bridge, A., Colwell, L. *et al.* InterPro in 2022. *Nucleic Acids Res* **51**, D418-427 (2023).

34. Mosbah, A., Beläch, A., Bornet, O., Beläch, J. P., Henrissat, B. & Darbon, H. Solution Structure of the Module X2-1 of Unknown Function of the Cellulosomal Scaffolding Protein CipC of Clostridium cellulolyticum. *J Mol Biol* **304**, 201–217 (2000).

35. Whelan, F. The long and the short of Periscope Proteins. *Biochem Soc Trans* **50**, 1293–1302 (2022).

Appendix A

The following are the full sequences of the expression vectors and their associated construct as described in *Chapter 2*. The primary sequences of the proteins after purification (*i.e.*, after affinity tag cleavage) are also detailed below.

1. NTD

GATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGAGGCCGTTGAGCACCGCCGCCGCGCAAGGAATGGTGCATGCAA
GGAGATGGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTC
GGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATA
GGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACCATGGGTCATCACCATCATCATCACGGGTCGGACTCAGAAGTC
AATCAAGAAGCTAAGCCAGAGGTCAAGCCAGAAGTCAAGCCTGAGACTCACATCAATTTAAAGGTGTCCGATGGATCTTCAGAGATCTTCTTCAAGATCAAAAAGACCACT
CCTTTAAGAAGGCTGATGGAAGCGTTCGCTAAAAGACAGGGTAAGGAAATGGACTCCTTAAGATTCTTGTACGACGGTATTAGAATTCAAGCTGATCAGGCCCCTGAAGATT
TGGACATGGAGGATAACGATATTATTGAGGCTCACCGCGAACAGATTGGAGGTGCTTCAAAACCCGACATAAAGGTAGGAGATTACGTTAAAATGGGTGTTTATAATAACG
CGAGCATTCTGTGGCGTTGCGTTTCTATCGATAACAACGGTCCGTTAATGCTGGCAGACAAGATTGTGGATACCCTCGCATACGACGCGAAAACCAACGACAACTCTAACAG
CAAGTCCCACAGCCGTAGCTACAAACGTGATGATTATGGCTCCAACTACTGGAAGGACTCAAACATGCGTAGCTGGCTCAATAGCACGGCTGCGGAAGGTAAAGTAGACTG
GCTGTGCGGTAACCCGCCCAAAGATGGCTATGTTAGCGGTGTGGGTGCGTACAACGAAAAAGCCGGCTTTCTGAACGCATTTAGCAAGTCTGAAATCGCAGCAATGAAAAC
CGTTACCCAACGTTCTCTGGTTTCGCATCCGGAGTACAACAAAGGTATCGTTGACGGCGACGCCAATTCTGATCTGTTGTACTACACAGACATCAGTGAGGCGGTTGCGAACT
ATGACTCTAGCTACTTCGAAACTACTACGGAGAAGGTGTTTCTGTTGGACGTTAAACAAGCAAATGCCGTGTGGAAAAACCTGAAGGGCTATTACGTGGCGTATAACAACGA
CGGTATGGCGTGGCCGTATTGGCTGCGTACCCCGGTGACCGACTGTAACCACGATATGCGTTATATCAGTTCCAGCGGCCAGGTCGGTCGTTATGCTCCGTGGTATAGCGACC
TGGGCGTGAGACCAGCATTTTATCTGGACTCTGAGTATTTTGTTACTACGAGCGGCTCTGGCTCCCAAAGCAGTCCGTATATCGGCTCCGCGCCGAACAAACAAGAAGATGA
CTACACCTGACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACC
CCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGG
TGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATC
GGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGC
CCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCG
GCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAACAAAATATTAACGCTTACAATTTAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCC
TATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTAGAAAAACTCATCGAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAA
TACCATATTTTTGAAAAAGCCGTTTCTGTAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCGACTCGTCCAACATC
AATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCATGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTCCAG
ACTTGTTCAACAGGCCAGCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGAGACGAAATACGCGATCGCTGTTAA
AAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACACTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTT
CCCGGGGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTCGGAAGAGGCATAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCT
GTAACATCATTGGCAACGCTACCTTTGCCATGTTTCAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCGACATTATCGCGAGC
CCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGAGCAAGACGTTTCCCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTATGTAAG
CAGACAGTTTTATTGTTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGC
GTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGA
TACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGT
GGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACG
ACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGA
GCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGC
CTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACC
GCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGT
GCGGTATTTCACACCGCATATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCC
CCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCA
CCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCA
GAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATA
CCGATGAAACGAGAGAGGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACTGGCGGTATGGATGCGGCGGGACCAG
AGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGATGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAA

ASKPDIKVGDYVKMGVYNNASILWRCVSIDNNGPLMLADKIVDTLAYDAKTNDNSNSKSHSRSYKRDDYGSNYWKDSNMRSWLNSTAAEGKVDWLCGNPPKDGYVSGVGAYN
EKAGFLNAFSKSEIAAMKTVTQRSLVSHPEYNKGIVDGDANSDLLYYTDISEAVANYDSSYFETTTEKVFLLDVKQANAVWKNLKGYYVAYNNDGMAWPYWLRTPVTDCNHDM
RYISSSGQVGRYAPWYSDLGVRPAFYLDSEYFVTTSGSGSQSSPYIGSAPNKQEDDYT

2. NTD+MD

GATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGAGGCCGTTGAGCACCGCCGCCGCGCAAGGAATGGTGCATGCAA
GGAGATGGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTC
GGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATA
GGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACCATGGGTCATCACCATCATCATCACGGGTCGGACTCAGAAGTC
AATCAAGAAGCTAAGCCAGAGGTCAAGCCAGAAGTCAAGCCTGAGACTCACATCAATTTAAAGGTGTCCGATGGATCTTCAGAGATCTTCTTCAAGATCAAAAAGACCACT
CCTTTAAGAAGGCTGATGGAAGCGTTCGCTAAAAGACAGGGTAAGGAAATGGACTCCTTAAGATTCTTGTACGACGGTATTAGAATTCAAGCTGATCAGGCCCCTGAAGATT
TGGACATGGAGGATAACGATATTATTGAGGCTCACCGCGAACAGATTGGAGGTGCTTCAAAACCCGACATAAAGGTAGGAGATTACGTTAAAATGGGTGTTTATAATAACG
CGAGCATTCTGTGGCGTTGCGTTTCTATCGATAACAACGGTCCGTTAATGCTGGCAGACAAGATTGTGGATACCCTCGCATACGACGCGAAAACCAACGACAACTCTAACAG
CAAGTCCCACAGCCGTAGCTACAAACGTGATGATTATGGCTCCAACTACTGGAAGGACTCAAACATGCGTAGCTGGCTCAATAGCACGGCTGCGGAAGGTAAAGTAGACTG
GCTGTGCGGTAACCCGCCCAAAGATGGCTATGTTAGCGGTGTGGGTGCGTACAACGAAAAAGCCGGCTTTCTGAACGCATTTAGCAAGTCTGAAATCGCAGCAATGAAAAC
CGTTACCCAACGTTCTCTGGTTTCGCATCCGGAGTACAACAAAGGTATCGTTGACGGCGACGCCAATTCTGATCTGTTGTACTACACAGACATCAGTGAGGCGGTTGCGAACT
ATGACTCTAGCTACTTCGAAACTACTACGGAGAAGGTGTTTCTGTTGGACGTTAAACAAGCAAATGCCGTGTGGAAAAACCTGAAGGGCTATTACGTGGCGTATAACAACGA
CGGTATGGCGTGGCCGTATTGGCTGCGTACCCCGGTGACCGACTGTAACCACGATATGCGTTATATCAGTTCCAGCGGCCAGGTCGGTCGTTATGCTCCGTGGTATAGCGACC
TGGGCGTGAGACCAGCATTTTATCTGGACTCTGAGTATTTTGTTACTACGAGCGGCTCTGGCTCCCAAAGCAGTCCGTATATCGGCTCCGCGCCGAACAAACAAGAAGATGA
CTACACCATTTCCGAGCCAGCAGAAGACGCCAACCCGGATTGGAACGTTAGTACCGAGCAGAGCATCCAACTGACCCTCGGCCCTTGGTATTCCAACGACGGTAAATACAGC
AACCCGACCATCCCGGTTTACACCATCCAAAAAACGCGTTCCGACACCGAGAATATGGTGGTGGTCGTGTGCGGTGAGGGCTACACTAAGAGCCAGCAGGGTAAATTCATC
AACGACGTAAAGCGCCTGTGGCAGGATGCTATGAAGTATGAGCCGTACCGCTCCTACGCGGATCGTTTCAATGTGTACGCTCTGTGTACCGCGTCAGAGAGCACCTTCGACA
ACGGTGGTTCTACGTTCTTTGATGTGATAGTTGACAAATACAACTCCCCGGTTATTTCCAACAACCTGCATGGTTCCCAATGGAAGAACCACATTTTTGAGCGTTGTATTGGT
CCGGAGTTTATCGAAAAGATTCATGATGCGCATATTAAAAAAAAGTGCGATCCGAATACCATTCCGTCGGGGAGCGAGTATGAACCGTACTATTACGTGCATGACTACATTG
CGCAGTTTGCTATGGTTGTGAATACCAAAAGTGACTTCGGCGGTGCGTACAATAACCGTGAATACGGCTTCCATTATTTCATCAGCCCGAGCGACAGCTATCGTGCAAGCAA
AACCTTCGCGCACGAATTCGGTCACGGCCTGCTTGGCTTGGGTGATGAATACTCCAACGGGTACCTGTTGGATGACAAGGAACTGAAGAGCCTGAACCTGAGCAGCGTCGAG
GACCCGGAAAAAATCAAGTGGCGTCAATTGCTGGGTTTTCGCAACACATATACGTGCCGTAATGCGTACGGCTCTAAAATGCTGGTTTCTAGCTACGAGTGCATCATGCGCG
ACACGAATTATCAGTTCTGCGAAGTTTGCCGCCTGCAAGGTTTCAAACGCATGAGCCAGTTGGTGAAGGATGTCGACTTGTATGTCGCCACCCCAGAGGTTAAAGAATACAC
CGGCGCATACAGCAAGCCGTCCGACTTTACCGATTTGGAAACCTCCAGCTACTACAACTACACCTATAACCGCAATGACCGCTTGCTTAGCGGTAATTCGAAGAGCCGTTTC
AATACCAATATGAACGGTAAAAAAATCGAGTTGCGTACCGTTATACAGAATATTAGCGACAAGAATGCGCGTCAACTGAAATTTAAGATGTGGATTAAACACAGCGATGGT
AGCGTGGCAACGGACTCGTCCGGCAACCCGCTGCAGACCGTCCAGACCTTCGATATTCCGGTTTGGAACGATAAGGCTAACTTTTGGCCTCTGGGTGCGTTGGACCACATCA
AGAGCGACTTCAACAGCGGTCTGAAGAGCTGTAGCCTGATTTATCAGATTCCGTCTGATGCCCAACTAAAGTCCGGCGACACCGTTGCGTTCCAGGTCCTGGACGAAAATGG
CAATGTGTTGGCTGACGATAACACCGAAACCCAGTGACTCGAGCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGC
CACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTA
GCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCG

CCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGG
CCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTT
GATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAACAAAATATTAACGCTTACAATTTAGGTGGCACT
TTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTAGAAAAACTCATCGAGCATCAAATGAAACT
GCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTGTAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCG
GTCTGCGATTCCGACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCATGAGTGACGACTGAATCCGGTGAGAAT
GGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCAGCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGC
GAGACGAAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACACTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGATA
TTCTTCTAATACCTGGAATGCTGTTTTCCCGGGGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTCGGAAGAGGCATAAATTCCGTC
AGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTTCAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGC
ACCTGATTGCCCGACATTATCGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGAGCAAGACGTTTCCCGTTGAATATGGCTCATAA
CACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTTATTGTTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGG
ATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAA
GGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAA
TCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTG
CACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGG
TAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTT
GTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCC
TGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCTGAT
GCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGC
TACGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGG
GAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCC
GCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGCCTCCGTGTAAGGGGG
ATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACTG
GCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGATGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAG
ATCCGGAA

ASKPDIKVGDYVKMGVYNNASILWRCVSIDNNGPLMLADKIVDTLAYDAKTNDNSNSKSHSRSYKRDDYGSNYWKDSNMRSWLNSTAAEGKVDWLCGNPPKDGYVSGVGAYN
EKAGFLNAFSKSEIAAMKTVTQRSLVSHPEYNKGIVDGDANSDLLYYTDISEAVANYDSSYFETTTEKVFLLDVKQANAVWKNLKGYYVAYNNDGMAWPYWLRTPVTDCNHDM
RYISSSGQVGRYAPWYSDLGVRPAFYLDSEYFVTTSGSGSQSSPYIGSAPNKQEDDYTISEPAEDANPDWNVSTEQSIQLTLGPWYSNDGKYSNPTIPVYTIQKTRSDTENMVVVVCG
EGYTKSQQGKFINDVKRLWQDAMKYEPYRSYADRFNVYYALCTASESTFDNGGSTFFDVIVDKYNSPVISNNLHGSQWKNHIFERCIGPEFIEKIHDAHIKKKKCDPNTIPSGSEYEPYY
YVHDYIAQFAMVVNTKSDFGGAYNNREYGFHYFISPSDSYRASKTFAHEFGHGLLGLGDEYSNGYLLDDKELKSLNLSSVEDPEKIKWRQLLGFRNTYTCRNAYGSKMLVSSYECI
MRDTNYQFCEVCRLQGFKRMSQLVKDVDLYVATPEVKEYTGAYSKPSDFTDLETSSYYNYTYNRNDRLLSGNSKSRFNTNMNGKKIELRTVIQNISDKNARQLKFKMWIKHSDGS
VATDSSGNPLQTVQTFDIPVWNDKANFWPLGALDHIKSDFNSGLKSCSLIYQIPSDAQLKSGDTVAFQVLDENGNVLADDNTETQ

3. MD

GATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGAGGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAA
GGAGATGGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTC
GGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATA
GGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACCATGGGTCATCACCATCATCATCACGGGTCGGACTCAGAAGTC
AATCAAGAAGCTAAGCCAGAGGTCAAGCCAGAAGTCAAGCCTGAGACTCACATCAATTTAAAGGTGTCCGATGGATCTTCAGAGATCTTCTTCAAGATCAAAAAGACCACT
CCTTTAAGAAGGCTGATGGAAGCGTTCGCTAAAAGACAGGGTAAGGAAATGGACTCCTTAAGATTCTTGTACGACGGTATTAGAATTCAAGCTGATCAGGCCCCTGAAGATT
TGGACATGGAGGATAACGATATTATTGAGGCTCACCGCGAACAGATTGGAGGTACCGAGCAGAGCATCCAACTGACCCTCGGCCCTTGGTATTCCAACGACGGTAAATACA
GCAACCCGACCATCCCGGTTTACACCATCCAAAAAACGCGTTCCGACACCGAGAATATGGTGGTGGTCGTGTGCGGTGAGGGCTACACTAAGAGCCAGCAGGGTAAATTCA
TCAACGACGTAAAGCGCCTGTGGCAGGATGCTATGAAGTATGAGCCGTACCGCTCCTACGCGGATCGTTTCAATGTGTACGCTCTGTGTACCGCGTCAGAGAGCACCTTCGA
CAACGGTGGTTCTACGTTCTTTGATGTGATAGTTGACAAATACAACTCCCCGGTTATTTCCAACAACCTGCATGGTTCCCAATGGAAGAACCACATTTTTGAGCGTTGTATTG
GTCCGGAGTTTATCGAAAAGATTCATGATGCGCATATTAAAAAAAAGTGCGATCCGAATACCATTCCGTCGGGGAGCGAGTATGAACCGTACTATTACGTGCATGACTACAT
TGCGCAGTTTGCTATGGTTGTGAATACCAAAAGTGACTTCGGCGGTGCGTACAATAACCGTGAATACGGCTTCCATTATTTCATCAGCCCGAGCGACAGCTATCGTGCAAGC
AAAACCTTCGCGCACGAATTCGGTCACGGCCTGCTTGGCTTGGGTGATGAATACTCCAACGGGTACCTGTTGGATGACAAGGAACTGAAGAGCCTGAACCTGAGCAGCGTCG
AGGACCCGGAAAAAATCAAGTGGCGTCAATTGCTGGGTTTTCGCAACACATATACGTGCCGTAATGCGTACGGCTCTAAAATGCTGGTTTCTAGCTACGAGTGCATCATGCG
CGACACGAATTATCAGTTCTGCGAAGTTTGCCGCCTGCAAGGTTCAAACGCATGAGCCAGTTGGTGAAGGATGTCGACTTGTATGTCGCCACCCCAGAGGTTAAAGAATAC
ACCGGCGCATACAGCAAGCCGTCCGACTTTACCGATTTGGAAACCTCCAGCTACTACAACTACACCTATAACCGCAATGACCGCTTGCTTAGCGGTAATTCGAAGAGCCGTT
TCAATACCAATATGAACGGTAAAAAAATCGAGTTGCGTACCGTTATACAGAATATTAGCGACAAGAATGCGCGTCAACTGAAATTTAAGATGTGGATTAAACACAGCGATG
GTAGCGTGGCAACGGACTCGTCCGGCAACCCGCTGCAGACCGTCCAGACCTTCGATATTCCGGTTTGGAACGATAAGGCTAACTTTTGGCCTCTGGGTGCGTTGGACCACAT
CAAGAGCGACTTCAACAGCGGTCTGAAGAGCTGTAGCCTGATTTATCAGATTCCGTCTGATGCCCAACTAAAGTCCGGCGACACCGTTGCGTTCCAGGTCCTGGACGAAAAT
GGCAATGTGTTGGCTGACGATAACACCGAAACCCAGTGACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCT
GCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTG
TAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTT
CGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTG
GGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTT
TTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAACAAAATATTAACGCTTACAATTTAGGTGGCA
CTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTAGAAAAACTCATCGAGCATCAAATGAAA
CTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTGTAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATC
GGTCTGCGATTCCGACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCATGAGTGACGACTGAATCCGGTGAGAA
TGGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCAGCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAG
CGAGACGAAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACACTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGAT
ATTCTTCTAATACCTGGAATGCTGTTTTCCCGGGGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTCGGAAGAGGCATAAATTCCGT
CAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTTCAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCG
CACCTGATTGCCCGACATTATCGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGAGCAAGACGTTTCCCGTTGAATATGGCTCATA
ACACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTTATTGTTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAG
GATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGA
AGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTA
ATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGT
GCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCG
GTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTT
TGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCC
CTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCTGA
TGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCG
CTACGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCG
GGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATC
CGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGCCTCCGTGTAAGGGG
GATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACT
GGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGATGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCA
GATCCGGAA

TEQSIQLTLGPWYSNDGKYSNPTIPVYTIQKTRSDTENMVVVVCGEGYTKSQQGKFINDVKRLWQDAMKYEPYRSYADRFNVYYALCTASESTFDNGGSTFFDVIVDKYNSPVISNN
LHGSQWKNHIFERCIGPEFIEKIHDAHIKKKKCDPNTIPSGSEYEPYYYVHDYIAQFAMVVNTKSDFGGAYNNREYGFHYFISPSDSYRASKTFAHEFGHGLLGLGDEYSNGYLLDDKE
LKSLNLSSVEDPEKIKWRQLLGFRNTYTCRNAYGSKMLVSSYECIMRDTNYQFCEVCRLQGFKRMSQLVKDVDLYVATPEVKEYTGAYSKPSDFTDLETSSYYNYTYNRNDRLLS
GNSKSRFNTNMNGKKIELRTVIQNISDKNARQLKFKMWIKHSDGSVATDSSGNPLQTVQTFDIPVWNDKANFWPLGALDHIKSDFNSGLKSCSLIYQIPSDAQLKSGDTVAFQVLDE
NGNVLADDNTETQ

4. MD+CTD1

CGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGC
CTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTG
CTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGA
ACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGG
CTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGA
GGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAG
AAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACT
CTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCGATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGT
AGGTTGAGGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCG
CTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGT
AGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATA
TACCATGGGTCATCACCATCATCATCACGGGTCGGACTCAGAAGTCAATCAAGAAGCTAAGCCAGAGGTCAAGCCAGAAGTCAAGCCTGAGACTCACATCAATTTAAAGGT
GTCCGATGGATCTTCAGAGATCTTCTTCAAGATCAAAAAGACCACTCCTTTAAGAAGGCTGATGGAAGCGTTCGCTAAAAGACAGGGTAAGGAAATGGACTCCTTAAGATTC
TTGTACGACGGTATTAGAATTCAAGCTGATCAGGCCCCTGAAGATTTGGACATGGAGGATAACGATATTATTGAGGCTCACCGCGAACAGATTGGAGGTACCGAGCAGAGC
ATCCAACTGACCCTCGGCCCTTGGTATTCCAACGACGGTAAATACAGCAACCCGACCATCCCGGTTTACACCATCCAAAAAACGCGTTCCGACACCGAGAATATGGTGGTGG
TCGTGTGCGGTGAGGGCTACACTAAGAGCCAGCAGGGTAAATTCATCAACGACGTAAAGCGCCTGTGGCAGGATGCTATGAAGTATGAGCCGTACCGCTCCTACGCGGATC
GTTTCAATGTGTACGCTCTGTGTACCGCGTCAGAGAGCACCTTCGACAACGGTGGTTCTACGTTCTTTGATGTGATAGTTGACAAATACAACTCCCCGGTTATTTCCAACAAC
CTGCATGGTTCCCAATGGAAGAACCACATTTTTGAGCGTTGTATTGGTCCGGAGTTTATCGAAAAGATTCATGATGCGCATATTAAAAAAAAGTGCGATCCGAATACCATTC
CGTCGGGGAGCGAGTATGAACCGTACTATTACGTGCATGACTACATTGCGCAGTTTGCTATGGTTGTGAATACCAAAAGTGACTTCGGCGGTGCGTACAATAACCGTGAATA
CGGCTTCCATTATTTCATCAGCCCGAGCGACAGCTATCGTGCAAGCAAAACCTTCGCGCACGAATTCGGTCACGGCCTGCTTGGCTTGGGTGATGAATACTCCAACGGGTAC
CTGTTGGATGACAAGGAACTGAAGAGCCTGAACCTGAGCAGCGTCGAGGACCCGGAAAAAATCAAGTGGCGTCAATTGCTGGGTTTTCGCAACACATATACGTGCCGTAAT
GCGTACGGCTCTAAAATGCTGGTTTCTAGCTACGAGTGCATCATGCGCGACACGAATTATCAGTTCTGCGAAGTTTGCCGCCTGCAAGGTTTCAAACGCATGAGCCAGTTGGT
GAAGGATGTCGACTTGTATGTCGCCACCCCAGAGGTTAAAGAATACACCGGCGCATACAGCAAGCCGTCCGACTTTACCGATTTGGAAACCTCCAGCTACTACAACTACACC
TATAACCGCAATGACCGCTTGCTTAGCGGTAATTCGAAGAGCCGTTCAATACCAATATGAACGGTAAAAAAATCGAGTTGCGTACCGTTATACAGAATATTAGCGACAAGA
ATGCGCGTCAACTGAAATTTAAGATGTGGATTAAACACAGCGATGGTAGCGTGGCAACGGACTCGTCCGGCAACCCGCTGCAGACCGTCCAGACCTTCGATATTCCGGTTTG
GAACGATAAGGCTAACTTTTGGCCTCTGGGTGCGTTGGACCACATCAAGAGCGACTTCAACAGCGGTCTGAAGAGCTGTAGCCTGATTTATCAGATTCCGTCTGATGCCCAA
CTAAAGTCCGGCGACACCGTTGCGTTCCAGGTCCTGGACGAAAATGGCAATGTGTTGGCTGACGATAACACCGAAACCCAGCGCTATACTACGGTTAGCATTCAATATAAAT
TCGAAGACGGTTCGGAGATTCCGAATACCGCGGGTGGCACCTTCACCGTTCCGTACGGTACGAAGCTGGATTTGACCCCAGCCAAAACGCTGTATGATTATGAATTTATCAA
AGTTGATGGCTTAAATAAGCCGATTGTCTCCGACGGCACTGTGGTGACGTACTACTACAAGAACTAACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAA
GCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATAT
CCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTC
GCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAA
CTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAAC
AACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAACAA
AATATTAACGCTTACAATTTAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTA
GAAAAACTCATCGAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTGTAATGAAGGAGAAAACTCACCGAGGCAG
TTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCGACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACC
ATGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCAGCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAA
CCGTTATTCATTCGTGATTGCGCCTGAGCGAGACGAAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACACTGCCAGCGCA
TCAACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGGGGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTT
GATGGTCGGAAGAGGCATAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTTCAGAAACAACTCTGGCGCATCGG
GCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCGACATTATCGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGAGCAA
GACGTTTCCCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTTATTGTTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGC
GTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGG
ATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTA
GCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGC
AGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCG
AAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTT
CGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTT
TGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAG
TGAGCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCAATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGT
TAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTT
ACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCG
ATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGG
TCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACT
GGAACGTTGTGAGGGTAAACAACTGGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGATGTAGGTGTTCCACAGGG
TAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGCGCTGACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCT
CAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAACGACA
GGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGGCGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGTGCA
AGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAAAATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGA
TAAAGAAGACAGTCATAAGTGCGGCGACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGAGATCCCGGTGCCTAAT
GAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTT
GCGTATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGC
CCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGG
ACTCGGTAATGGCG

TEQSIQLTLGPWYSNDGKYSNPTIPVYTIQKTRSDTENMVVVVCGEGYTKSQQGKFINDVKRLWQDAMKYEPYRSYADRFNVYALCTASESTFDNGGSTFFDVIVDKYNSPVISNN
LHGSQWKNHIFERCIGPEFIEKIHDAHIKKKCDPNTIPSGSEYEPYYYVHDYIAQFAMVVNTKSDFGGAYNNREYGFHYFISPSDSYRASKTFAHEFGHGLLGLGDEYSNGYLLDDKE
LKSLNLSSVEDPEKIKWRQLLGFRNTYTCRNAYGSKMLVSSYECIMRDTNYQFCEVCRLQGFKRMSQLVKDVDLYVATPEVKEYTGAYSKPSDFTDLETSSYYNYTYNRNDRLLS
GNSKSRFNTNMNGKKIELRTVIQNISDKNARQLKFKMWIKHSDGSVATDSSGNPLQTVQTFDIPVWNDKANFWPLGALDHIKSDFNSGLKSCSLIYQIPSDAQLKSGDTVAFQVLDE
NGNVLADDNTETQRYTTVSIQYKFEDGSEIPNTAGGTFTVPYGTKLDLTPAKTLYDYEFIKVDGLNKPIVSDGTVVTYYYKN

5. MD+CTD1 E540A

CGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGC
CTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTG
CTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGA
ACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGG
CTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGA
GGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAG
AAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACT
CTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCGATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGT
AGGTTGAGGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCG
CTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGT
AGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATA
TACCATGGGTCATCACCATCATCATCACGGGTCGGACTCAGAAGTCAATCAAGAAGCTAAGCCAGAGGTCAAGCCAGAAGTCAAGCCTGAGACTCACATCAATTTAAAGGT
GTCCGATGGATCTTCAGAGATCTTCTTCAAGATCAAAAAGACCACTCCTTTAAGAAGGCTGATGGAAGCGTTCGCTAAAAGACAGGGTAAGGAAATGGACTCCTTAAGATTC
TTGTACGACGGTATTAGAATTCAAGCTGATCAGGCCCCTGAAGATTTGGACATGGAGGATAACGATATTATTGAGGCTCACCGCGAACAGATTGGAGGTACCGAGCAGAGC
ATCCAACTGACCCTCGGCCCTTGGTATTCCAACGACGGTAAATACAGCAACCCGACCATCCCGGTTTACACCATCCAAAAAACGCGTTCCGACACCGAGAATATGGTGGTGG

TCGTGTGCGGTGAGGGCTACACTAAGAGCCAGCAGGGTAAATTCATCAACGACGTAAAGCGCCTGTGGCAGGATGCTATGAAGTATGAGCCGTACCGCTCCTACGCGGATC
GTTTCAATGTGTACGCTCTGTGTACCGCGTCAGAGAGCACCTTCGACAACGGTGGTTCTACGTTCTTTGATGTGATAGTTGACAAATACAACTCCCCGGTTATTTCCAACAAC
CTGCATGGTTCCCAATGGAAGAACCACATTTTTGAGCGTTGTATTGGTCCGGAGTTTATCGAAAAGATTCATGATGCGCATATTAAAAAAAAGTGCGATCCGAATACCATTC
CGTCGGGGAGCGAGTATGAACCGTACTATTACGTGCATGACTACATTGCGCAGTTTGCTATGGTTGTGAATACCAAAAGTGACTTCGGCGGTGCGTACAATAACCGTGAATA
CGGCTTCCATTATTTCATCAGCCCGAGCGACAGCTATCGTGCAAGCAAAACCTTCGCGCACGCATTCGGTCACGGCCTGCTTGGCTTGGGTGATGAATACTCCAACGGGTACC
TGTTGGATGACAAGGAACTGAAGAGCCTGAACCTGAGCAGCGTCGAGGACCCGGAAAAAATCAAGTGGCGTCAATTGCTGGGTTTTCGCAACACATATACGTGCCGTAATG
CGTACGGCTCTAAAATGCTGGTTTCTAGCTACGAGTGCATCATGCGCGACACGAATTATCAGTTCTGCGAAGTTTGCCGCCTGCAAGGTTTCAAACGCATGAGCCAGTTGGTG
AAGGATGTCGACTTGTATGTCGCCACCCCAGAGGTTAAAGAATACACCGGCGCATACAGCAAGCCGTCCGACTTTACCGATTTGGAAACCTCCAGCTACTACAACTACACCT
ATAACCGCAATGACCGCTTGCTTAGCGGTAATTCGAAGAGCCGTTTCAATACCAATATGAACGGTAAAAAAATCGAGTTGCGTACCGTTATACAGAATATTAGCGACAAGAA
TGCGCGTCAACTGAAATTTAAGATGTGGATTAAACACAGCGATGGTAGCGTGGCAACGGACTCGTCCGGCAACCCGCTGCAGACCGTCCAGACCTTCGATATTCCGGTTTGG
AACGATAAGGCTAACTTTTGGCCTCTGGGTGCGTTGGACCACATCAAGAGCGACTTCAACAGCGGTCTGAAGAGCTGTAGCCTGATTTATCAGATTCCGTCTGATGCCCAAC
TAAAGTCCGGCGACACCGTTGCGTTCCAGGTCCTGGACGAAAATGGCAATGTGTTGGCTGACGATAACACCGAAACCCAGCGCTATACTACGGTTAGCATTCAATATAAATT
CGAAGACGGTTCGGAGATTCCGAATACCGCGGGTGGCACCTTCACCGTTCCGTACGGTACGAAGCTGGATTTGACCCCAGCCAAAACGCTGTATGATTATGAATTTATCAAA
GTTGATGGCTTAAATAAGCCGATTGTCTCCGACGGCACTGTGGTGACGTACTACTACAAGAACTAACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAG
CCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATC
CGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCG
CTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAAC
TTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACA
ACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAACAAA
ATATTAACGCTTACAATTTAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTAG
AAAAACTCATCGAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTGTAATGAAGGAGAAAACTCACCGAGGCAGT
TCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCGACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCA
TGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCAGCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAAC
CGTTATTCATTCGTGATTGCGCCTGAGCGAGACGAAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACACTGCCAGCGCAT
CAACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGGGGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTG
ATGGTCGGAAGAGGCATAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTTCAGAAACAACTCTGGCGCATCGGG
CTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCGACATTATCGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGAGCAAG
ACGTTTCCCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTTATTGTTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCG
TCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGA
TCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAG
CACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCA
GCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGA
AGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTC
GCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTT
GCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGT
GAGCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCAATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTT
AAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTA
CAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGAT
TCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTC
ACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGG
AACGTTGTGAGGGTAAACAACTGGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGATGTAGGTGTTCCACAGGGTA
GCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGCGCTGACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCA
GGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAACGACAGG
AGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGGCGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGTGCAAG
ATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAAAATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATA
AAGAAGACAGTCATAAGTGCGGCGACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGAGATCCCGGTGCCTAATGA
GTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGC
GTATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCC
CAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGAC
TCGGTAATGGCG

TEQSIQLTLGPWYSNDGKYSNPTIPVYTIQKTRSDTENMVVVVCGEGYTKSQQGKFINDVKRLWQDAMKYEPYRSYADRFNVYALCTASESTFDNGGSTFFDVIVDKYNSPVISNN
LHGSQWKNHIFERCIGPEFIEKIHDAHIKKKCDPNTIPSGSYEPYYYVHDYIAQFAMVVNTKSDFGGAYNNREYGFHYFISPSDSYRASKTFAHAFGHGLLGLGDEYSNGYLLDDKE
LKSLNLSSVEDPEKIKWRQLLGFRNTYTCRNAYGSKMLVSSYECIMRDTNYQFCEVCRLQGFKRMSQLVKDVDLYVATPEVKEYTGAYSKPSDFTDLETSSYYNYTYNRNDRLLS
GNSKSRFNTNMNGKKIELRTVIQNISDKNARQLKFKMWIKHSDGSVATDSSGNPLQTVQTFDIPVWNDKANFWPLGALDHIKSDFNSGLKSCSLIYQIPSDAQLKSGDTVAFQVLDE
NGNVLADDNTETQRYTTVSIQYKFEDGSEIPNTAGGTFTVPYGTKLDLTPAKTLYDYEFIKVDGLNKPIVSDGTVVTYYYKN

6. CTD2

CGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGC
CTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTG
CTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGA
ACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGG
CTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGA
GGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAG
AAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACT
CTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCGATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGT
AGGTTGAGGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCG
CTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGT
AGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATA
TACCATGGGTCATCACCATCATCATCACGGGTCGGACTCAGAAGTCAATCAAGAAGCTAAGCCAGAGGTCAAGCCAGAAGTCAAGCCTGAGACTCACATCAATTTAAAGGT
GTCCGATGGATCTTCAGAGATCTTCTTCAAGATCAAAAAGACCACTCCTTTAAGAAGGCTGATGGAAGCGTTCGCTAAAAGACAGGGTAAGGAAATGGACTCCTTAAGATTC
TTGTACGACGGTATTAGAATTCAAGCTGATCAGGCCCCTGAAGATTTGGACATGGAGGATAACGATATTATTGAGGCTCACCGCGAACAGATTGGAGGTGAACATACCCATA
ATCTGACGCTGGTGGCGGCGAAGGCTGCTACCTGCACAACCGCGGGCAACTCAGCCTACTACACGTGCGATGGCTGTGATAAATGGTTTGCGGACGCGACAGGCTCCGTCGA
AATCACCGACAAAACCAGCGTGAAGATCCCGGCGCCGGGTCATACCGCGGGAACCGAGTGGAAAAGCGACGATACCAATCATTGGCACGAGTGCACTGTGGCGGGTTGCG
GTGTGATTATCGAGAGCACTAAGAGTGCGCACACCGCGGGTGAATGGATCGTGGATACCCCAGCTACCGCGACTACTGCTGGCACCAAGCACAAAGAGTGTACTGTGTGTC
ATCGTGTTCTTGAGACGCAACCGATTCCGAGCACCGGTACCTAACTCGAGCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGG
CTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCG
CCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCTCGCC
ACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACG
TAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTA
TTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAACAAAATATTAACGCTTACAATTTAGGT
GGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAATTCTTAGAAAAACTCATCGAGCATCAAAT
GAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTGTAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTG
GTATCGGTCTGCGATTCCGACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCATGAGTGACGACTGAATCCGGT
GAGAATGGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCAGCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGC
CTGAGCGAGACGAAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACACTGCCAGCGCATCAACAATATTTTCACCTGAATC
AGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGGGGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTCGGAAGAGGCATAAAT
TCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTTCAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGAT
TGTCGCACCTGATTGCCCGACATTATCGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGAGCAAGACGTTTCCCGTTGAATATGGC
TCATAACACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTTATTGTTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGAT
CAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTT

TTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCT
CTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGG
GTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGT
ATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCG
ATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTT
ATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCG
CCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCAATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTAT
CGCTACGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTC
CGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCA
TCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGCCTCCGTGTAAGGG
GGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAA
CTGGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGATGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATG
CAGATCCGGAACATAATGGTGCAGGGCGCTGACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGACGTTTTGCAGC
AGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCACGATCATGCGCACCCG
TGGGGCCGCCATGCCGGCGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGTGCAAGATTCCGAATACCGCAAGCGA
CAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAAAATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGC
GGCGACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGAGATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATT
GCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTT
TTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTT
GATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCG

EHTHNLTLVAAKAATCTTAGNSAYYTCDGCDKWFADATGSVEITDKTSVKIPAPGHTAGTEWKSDDTNHWHECTVAGCGVIIESTKSAHTAGEWIVDTPATATTAGTKHKECTVC
HRVLETQPIPSTGT

7. CTD1-4

CGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGC
CTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTG
CTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGA
ACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGG
CTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGA
GGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAG
AAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACT
CTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCGATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGT
AGGTTGAGGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCG
CTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGT
AGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATA
TACCATGGGTCATCACCATCATCATCACGGGTCGGACTCAGAAGTCAATCAAGAAGCTAAGCCAGAGGTCAAGCCAGAAGTCAAGCCTGAGACTCACATCAATTTAAAGGT
GTCCGATGGATCTTCAGAGATCTTCTTCAAGATCAAAAAGACCACTCCTTTAAGAAGGCTGATGGAAGCGTTCGCTAAAAGACAGGGTAAGGAAATGGACTCCTTAAGATTC
TTGTACGACGGTATTAGAATTCAAGCTGATCAGGCCCCTGAAGATTTGGACATGGAGGATAACGATATTATTGAGGCTCACCGCGAACAGATTGGAGGTACCCAGCGCTATA
CTACGGTTAGCATTCAATATAAATTCGAAGACGGTTCGGAGATTCCGAATACCGCGGGTGGCACCTTCACCGTTCCGTACGGTACGAAGCTGGATTTGACCCCAGCCAAAAC
GCTGTATGATTATGAATTTATCAAAGTTGATGGCTTAAATAAGCCGATTGTCTCCGACGGCACTGTGGTGACGTACTACTACAAGAACAAGAACGAGGAACATACCCATAAT
CTGACGCTGGTGGCGGCGAAGGCTGCTACCTGCACAACCGCGGGCAACTCAGCCTACTACACGTGCGATGGCTGTGATAAATGGTTTGCGGACGCGACAGGCTCCGTCGAA
ATCACCGACAAAACCAGCGTGAAGATCCCGGCGCCGGGTCATACCGCGGAACCGAGTGGAAAAGCGACGATACCAATCATTGGCACGAGTGCACTGTGGCGGGTTGCGGT
GTGATTATCGAGAGCACTAAGAGTGCGCACACCGCGGGTGAATGGATCGTGGATACCCCAGCTACCGCGACTACTGCTGGCACCAAGCACAAAGAGTGTACTGTGTGTCAT
CGTGTTCTTGAGACGCAACCGATTCCGAGCACCGGTACCGAGCTGAAGATCATAGCCGGCGACAACCAGATTTATAACAAAGCCAGCGGTAGCGATGTAACGATCACCTGT
AATGGTGATTTTGCAAAATTCACCGGTATCAAAGTTGACGGCAGTGTCGTGGACAGCAGCAATTACACCGCGGTGTCCGGTTCTACGGTTCTGACACTGAAAGCGAGTTATC
TGGGCACTCTGACTGACGGTTCTCACACCATCACCTTCGTTTATACCGACGGCGAGGCGAACGCTAATTTAACGGTACGCACCGCTGGCAGCGGGCACATCCACGATTATGG
CACTGAGTGGAAATCCAATGCAGATAACCATTGGCACGAATGCAACTGCGGCGATAAGAAGGATGAGGCGGCTCACTCGTTCAAGTGGGTAGTCGATAAAGAAGCCACCGC
TACCAAAAAAGGGTTCTAAGCACGAAGAGTGCAAAATCTGCGGTTATAAGCGCTCTGCGGTGGAAATCCCGGCGACCGGTTAACTCGAGCACCACCACCACCACCACTGAGA
TCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTG
AAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCC
TAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCA
CCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCT
TGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTA
ACGCGAATTTTAACAAAATATTAACGCTTACAATTTAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCT
CATGAATTAATTCTTAGAAAAACTCATCGAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTGTAATGAAGGAGAA
AACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCGACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATC
AAGTGAGAAATCACCATGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCAGCCATTACGCTCGTCATCAAAATCA
CTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGAGACGAAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGG
AACACTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGGGGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGT
ACGGATAAAATGCTTGATGGTCGGAAGAGGCATAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTTCAGAAACA
ACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCGACATTATCGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAAT
CGCGGCCTAGAGCAAGACGTTTCCCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTTATTGTTCATGACCAAAATCCCTTAACGTGAGT
TTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCG
GTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCA
CTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGT
TACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAA
GCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTAT
AGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGC
CTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGA
GCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCAATGGTGCACTCTCAGTACAATCTGCT
CTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTG
CTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAG
CGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCG
GTTTTTTCCTGTTTGGTCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCTCACGATACGGGTTACTGATGATGA
ACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACTGGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGATGT
AGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGCGCTGACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGAC
CATTCATGTTGTTGCTCAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCC
GGGTCCTCAACGACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGGCGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTT
GAGCGAGGGCGTGCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAAAATGACCCAGAGCGCTGCCGGCACCTGTC
CTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCGACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGAG
ATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCG
GGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGT
CCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACC
AACGCGCAGCCCGGACTCGGTAATGGCG

TQRYTTVSIQYKFEDGSEIPNTAGGTFTVPYGTKLDLTPAKTLYDYEFIKVDGLNKPIVSDGTVVTYYYKNKNEEHTHNLTLVAAKAATCTTAGNSAYYTCDGCDKWFADATGSVE
ITDKTSVKIPAPGHTAGTEWKSDDTNHWHECTVAGCGVIIESTKSAHTAGEWIVDTPATATTAGTKHKECTVCHRVLETQPIPSTGTELKIIAGDNQIYNKASGSDVTITCNGDFAKF
TGIKVDGSVVDSSNYTAVSGSTVLTLKASYLGTLTDGSHTITFVYTDGEANANLTVRTAGSGHIHDYGTEWKSNADNHWHECNCGDKKDEAAHSFKWVVDKEATATKKGSKHEE
CKICGYKRSAVEIPATG

8. MD+CTD1-4

ATGCTGAATGAGGGCATCGTTCCCACTGCGATGCTGGTTGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCCATTACCGAGTCCGGGCTGCGCGTTGGTGCGGATATCT
CGGTAGTGGGATACGACGATACCGAAGATAGCTCATGTTATATCCCGCCGTTAACCACCATCAAACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGACCGCTTGCTGCA
ACTCTCTCAGGGCCAGGCGGTGAAGGGCAATCAGCTGTTGCCAGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCC
GATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGACTCATGACCAAAATCCCTTAACGTGAGTTACGCGCGCGTCGTTCCACTGAGCGTCAGA
CCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAG
AGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTTCTTCTAGTGTAGCCGTAGTTAGCCCACCACTTCAAGAACTCTGTAGCACCG
CCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTC
GGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAG
AAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCT
CTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACA
TGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAG
GAAGCGGAAGGCGAGAGTAGGGAACTGCCAGGCATCAAACTAAGCAGAAGGCCCCTGACGGATGGCCTTTTTGCGTTTCTACAAACTCTTTCTGTGTTGTAAAACGACGGCC
AGTCTTAAGCTCGGGCCCCCTGGGCGGTTCTGATAACGAGTAATCGTTAATCCGCAAATAACGTAAAAACCCGCTTCGGCGGGTTTTTTTATGGGGGGAGTTTAGGGAAAGA
GCATTTGTCAGAATATTTAAGGGCGCCTGTCACTTTGCTTGATATATGAGAATTATTTAACCTTATAAATGAGAAAAAAGCAACGCACTTTAAATAAGATACGTTGCTTTTTC
GATTGATGAACACCTATAATTAAACTATTCATCTATTATTTATGATTTTTTGTATATACAATATTTCTAGTTTGTTAAAGAGAATTAAGAAAATAAATCTCGAAAATAATAAA
GGGAAAATCAGTTTTTGATATCAAAATTATACATGTCAACGATAATACAAAATATAATACAAACTATAAGATGTTATCAGTATTTATTATGCATTTAGAATAAATTTTGTGTC
GCCCTTAATTGTGAGCGGATAACAATTACGAGCTTCATGCACAGTGAAATCATGAAAAATTTATTTGCTTTGTGAGCGGATAACAATTATAATATGTGGAATTGTGAGCGCTC
ACAATTCCACAACGGTTTCCCTCTAGAAATAATTTTGTTTAACTTTTAGGAGGTAAAACATATGGCGAAGTATTACGGCAGTATGAAAATCGAAGAAGGTAAACTGGTAATC
TGGATTAACGGCGATAAAGGCTATAACGGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGAATTAAAGTCACCGTTGAGCATCCGGATAAACTGGAAGAGAAA
TTCCCACAGGTTGCGGCAACTGGCGATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTGGCTGAAATCACCCCGGACAAAGC
GTTCCAGGACAAGCTGTATCCGTTTACCTGGGATGCCGTACGTTACAACGGCAAGCTGATTGCTTACCCGATCGCTGTTGAAGCGTTATCGCTGATTTATAACAAAGATCTGC
TGCCGAACCCGCCAAAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAAGGTAAGAGCGCGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGC
CGCTGATTGCTGCTGACGGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTACGACATTAAAGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTGACCTTCCTGG
TTGACCTGATTAAAAACAAACACATGAATGCAGACACCGATTACTCCATCGCAGAAGCTGCCTTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGT
CCAACATCGACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGTCAACCATCCAAACCGTTCGTTGGCGTGCTGAGCGCAGGTATTAACGCCGCCAG
TCCGAACAAAGAGCTGGCAAAAGAGTTCCTCGAAAACTATCTGCTGACTGATGAAGGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCTTA
CGAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACTATGGAAAACGCCCAGAAAGGTGAAATCATGCCGAACATCCCGCAGATGTCCGCTTCTGGTATGCCGTGCG
TACTGCGGTGATCAACGCCGCCAGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGACTCGTATCACTAAAAGCGGAAGCGGAGAGAACCTGTACTTCCAATC
CCTGGTGCCGCGCGGCAGCGGATCCACCGAGCAGAGCATCCAACTGACCCTCGGCCCTTGGTATTCCAACGACGGTAAATACAGCAACCCGACCATCCCGGTTTACACCATC
CAAAAAACGCGTTCCGACACCGAGAATATGGTGGTCGTCGTGCGGTGAGGGCTACACTAAGAGCCAGCAGGGTAAATTCATCAACGACGTAAAGCGCCTGTGGCAGGAT
GCTATGAAGTATGAGCCGTACCGCTCCTACGCGGATCGTTTCAATGTGTACGCTCTGTGTACCGCGTCAGAGAGCACCTTCGACAACGGTGGTTCTACGTTCTTTGATGTGAT
AGTTGACAAATACAACTCCCCGGTTATTTCCAACAACCTGCATGGTTCCCAATGGAAGAACCACATTTTTGAGCGTTGTATTGGTCCGGAGTTTATCGAAAAGATTCATGATG
CGCATATTAAAAAAAAAGTGCGATCCGAATACCATTCCGTCGGGGAGCGAGTATGAACCGTACTATTACGTGCATGACTACATTGCGCAGTTTGCTATGGTTGTGAATACCAA
AAGTGACTTCGGCGGTGCGTACAATAACCGTGAATACGGCTTCCATTATTTCATCAGCCCGAGCGACAGCTATCGTGCAAGCAAAACCTTCGCGCACGAATTCGGTCACGGC
CTGCTTGGCTTGGGTGATGAATACTCCAACGGGTACCTGTTGGATGACAAGGAACTGAAGAGCCTGAACCTGAGCAGCGTCGAGGACCCGGAAAAAATCAAGTGGCGTCAA
TTGCTGGGTTTTCGCAACACATATACGTGCCGTAATGCGTACGGCTCTAAAATGCTGGTTTCTAGCTACGAGTGCATCATGCGCGACACGAATTATCAGTTCTGCGAAGTTTG
CCGCCTGCAAGGTTTCAAACGCATGAGCCAGTTGGTGAAGGATGTCGACTTGTATGTCGCCACCCCAGAGGTTAAAGAATACACCGGCGCATACAGCAAGCCGTCCGACTTT
ACCGATTTGGAAACCTCCAGCTACTACAACTACACCTATAACCGCAATGACCGCTTGCTTAGCGGTAATTCGAAGAGCCGTTTCAATACCAATATGAACGGTAAAAAAATCG
AGTTGCGTACCGTTATACAGAATATTAGCGACAAGAATGCGCGTCAACTGAAATTTAAGATGTGGATTAAACACAGCGATGGTAGCGTGGCAACGGACTCGTCCGGCAACC
CGCTGCAGACCGTCCAGACCTTCGATATTCCGGTTTGGAACGATAAGGCTAACTTTTGGCCTCTGGGTGCGTTGGACCACATCAAGAGCGACTTCAACAGCGGTCTGAAGAG
CTGTAGCCTGATTTATCAGATTCCGTCTGATGCCCAACTAAAGTCCGGCGACACCGTTGCGTTCCAGGTCCTGGACGAAAATGGCAATGTGTTGGCTGACGATAACACCGAA
ACCCAGCGCTATACTACGGTTAGCATTCAATATAAATTCGAAGACGGTTCGGAGATTCCGAATACCGCGGGTGGCACCTTCACCGTTCCGTACGGTACGAAGCTGGATTTGA
CCCCAGCCAAAACGCTGTATGATTATGAATTTATCAAAGTTGATGGCTTAAATAAGCCGATTGTCTCCGACGGCACTGTGGTGACGTACTACTACAAGAACAAGAACGAGGA
ACATACCCATAATCTGACGCTGGTGGCGGCGAAGGCTGCTACCTGCACAACCGCGGGCAACTCAGCCTACTACACGTGCGATGGCTGTGATAAATGGTTTGCGGACGCGACA
GGCTCCGTCGAAATCACCGACAAAACCAGCGTGAAGATCCCGGCGCCGGGTCATACCGCGGGAACCGAGTGGAAAAGCGACGATACCAATCATTGGCACGAGTGCACTGTG
GCGGGTTGCGGTGTGATTATCGAGAGCACTAAGAGTGCGCACACCGCGGGTGAATGGATCGTGGATACCCCAGCTACCGCGACTACTGCTGGCACCAAGCACAAAGAGTGT
ACTGTGTGTCATCGTGTTCTTGAGACGCAACCGATTCCGAGCACCGGTACCGAGCTGAAGATCATAGCCGGCGACAACCAGATTTATAACAAAGCCAGCGGTAGCGATGTAA
CGATCACCTGTAATGGTGATTTTGCAAAATTCACCGGTATCAAAGTTGACGGCAGTGTCGTGGACAGCAGCAATTACACCGCGGTGTCCGGTTCTACGGTTCTGACACTGAA
AGCGAGTTATCTGGGCACTCTGACTGACGGTTCTCACACCATCACCTTCGTTTATACCGACGGCGAGGCGAACGCTAATTTAACGGTACGCACCGCTGGCAGCGGGCACATC
CACGATTATGGCACTGAGTGGAAATCCAATGCAGATAACCATTGGCACGAATGCAACTGCGGCGATAAGAAGGATGAGGCGGCTCACTCGTTCAAGTGGGTAGTCGATAAA
GAAGCCACCGCTACCAAAAAGGGTTCTAAGCACGAAGAGTGCAAAATCTGCGGTTATAAGCGCTCTGCGGTGAAATCCCGGCGACCGGATCCCACCATCACCATCACCAC
TAAGACTACAAAGACGATGACGACAAGTAACTCGAGCCCCAAGGGCGACACCCCCTAATTAGCCCGGGCGAAAGGCCCAGTCTTTCGACTGAGCCTTTCGTTTTATTTGATG
CCTGGCAGTTCCCTACTCTCGCATGGGGAGTCCCCACACTACCATCGGCGCTACGGCGTTTCACTTCTGAGTTCGGCATGGGGTCAGGTGGGACCACCGCGCTACTGCCGCCA
GGCAAACAAGGGGTGTTATGAGCCATATTCAGGTATAAATGGGCTCGCGATAATGTTCAGAATTGGTTAATTGGTTGTAACACTGACCCCTATTTGTTTATTTTTCTAAATAC
ATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAATATTGAAAAAGGAAGAATATGAGCCATATTCAACGGGAAACGTCGAGGCCGCGATTAA
ATTCCAACATGGATGCTGATTTATATGGGTATAAATGGGCTCGCGATAATGTCGGGCAATCAGGTGCGACAATCTATCGCTTGTATGGGAAGCCCGATGCGCCAGAGTTGTT
TCTGAAACATGGCAAAGGTAGCGTTGCCAATGATGTTACAGATGAGATGGTCAGACTAAACTGGCTGACGGAATTTATGCCACTTCCGACCATCAAGCATTTTATCCGTACT
CCTGATGATGCATGGTTACTCACCACTGCGATCCCCGGAAAAACAGCGTTCCAGGTATTAGAAGAATATCCTGATTCAGGTGAAAATATTGTTGATGCGCTGGCAGTGTTCCT
GCGCCGGTTGCACTCGATTCCTGTTTGTAATTGTCCTTTTAACAGCGATCGCGTATTTCGCCTCGCTCAGGCGCAATCACGAATGAATAACGGTTTGGTTGATGCGAGTGATTT
TGATGACGAGCGTAATGGCTGGCCTGTTGAACAAGTCTGGAAAGAAATGCATAAACTTTTGCCATTCTCACCGGATTCAGTCGTCACTCATGGTGATTTCTCACTTGATAACC
TTATTTTTGACGAGGGGAAATTAATAGGTTGTATTGATGTTGGACGAGTCGGAATCGCAGACCGATACCAGGATCTTGCCATCCTATGGAACTGCCTCGGTGAGTTTTCTCCT
TCATTACAGAAACGGCTTTTTCAAAAATATGGTATTGATAATCCTGATATGAATAAATTGCAATTTCATTTGATGCTCGATGAGTTTTTCTAAGCGGCGCGCCATCGAATGGC
GCAAAACCTTTCGCGGTATGGCATGATAGCGCCCGGAAGAGAGTCAATTCAGGGTGGTGAATATGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCTTA
TCAGACCGTTTCCCGCGTGGTGAACCAGGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGCGGCGATGGCGGAGCTGAATTACATTCCCAACCGCGTGGCACA
ACAACTGGCGGGCAAACAGTCGTTGCTGATTGGCGTTGCCACCTCCAGTCTGGCCCTGCACGCGCCGTCGCAAATTGTCGCGGCGATTAAATCTCGCGCCGATCAACTGGGT
GCCAGCGTGGTGGTGTCGATGGTAGAACGAAGCGGCGTCGAAGCCTGTAAAGCGGCGGTGCACAATCTTCTCGCGCAACGCGTCAGTGGGCTGATCATTAACTATCCGCTGG
ATGACCAGGATGCCATTGCTGTGGAAGCTGCCTGCACTAATGTTCCGGCGTTATTTCTTGATGTCTCTGACCAGACACCCATCAACAGTATTATTTTCTCCCATGAGGACGGT
ACGCGACTGGGCGTGGAGCATCTGGTCGCATTGGGTCACCAGCAAATCGCGCTGTTAGCGGGCCCATTAAGTTCTGTCTCGGCGCGTCTGCGTCTGGCTGGCTGGCATAAAT
ATCTCACTCGCAATCAAATTCAGCCGATAGCGGAACGGGAAGGCGACTGGAGTGCCATGTCCGGTTTTCAACAAACCATGCAA

GSGSTEQSIQLTLGPWYSNDGKYSNPTIPVYTIQKTRSDTENMVVVVCGEGYTKSQQGKFINDVKRLWQDAMKYEPYRSYADRFNVYALCTASESTFDNGGSTFFDVIVDKYNSPV
ISNNLHGSQWKNHIFERCIGPEFIEKIHDAHIKKKCDPNTIPSGSEYEPYYYVHDYIAQFAMVVNTKSDFGGAYNNREYGFHYFISPSDSYRASKTFAHEFGHGLLGLGDEYSNGYLL
DDKELKSLNLSSVEDPEKIKWRQLLGFRNTYTCRNAYGSKMLVSSYECIMRDTNYQFCEVCRLQGFKRMSQLVKDVDLYVATPEVKEYTGAYSKPSDFTDLETSSYYNYTYNRND
RLLSGNSKSRFNTNMNGKKIELRTVIQNISDKNARQLKFKMWIKHSDGSVATDSSGNPLQTVQTFDIPVWNDKANFWPLGALDHIKSDFNSGLKSCSLIYQIPSDAQLKSGDTVAFQ
VLDENGNVLADDNTETQRYTTVSIQYKFEDGSEIPNTAGGTFTVPYGTKLDLTPAKTLYDYEFIKVDGLNKPIVSDGTVVTYYYKNKNEEHTHNLTLVAAKAATCTTAGNSAYYTC
DGCDKWFADATGSVEITDKTSVKIPAPGHTAGTEWKSDDTNHWHECTVAGCGVIIESTKSAHTAGEWIVDTPATATTAGTKHKECTVCHRVLETQPIPSTGTELKIIAGDNQIYNKA
SGSDVTITCNGDFAKFTGIKVDGSVVDSSNYTAVSGSTVLTLKASYLGTLTDGSHTITFVYTDGEANANLTVRTAGSGHIHDYGTEWKSNADNHWHECNCGDKKDEAAHSFKWVV
DKEATATKKGSKHEECKICGYKRSAVEIPATGSHHHHHH

9. FL

ATGCTGAATGAGGGCATCGTTCCCACTGCGATGCTGGTTGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCCATTACCGAGTCCGGGCTGCGCGTTGGTGCGGATATCT
CGGTAGTGGGATACGACGATACCGAAGATAGCTCATGTTATATCCCGCCGTTAACCACCATCAAACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGACCGCTTGCTGCA
ACTCTCTCAGGGCCAGGCGGTGAAGGGCAATCAGCTGTTGCCAGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCC
GATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGACTCATGACCAAAATCCCTTAACGTGAGTTACGCGCGCGTCGTTCCACTGAGCGTCAGA
CCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAG
AGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTTCTTCTAGTGTAGCCGTAGTTAGCCCACCACTTCAAGAACTCTGTAGCACCG
CCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTC
GGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAG

AAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCT
CTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACA
TGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAG
GAAGCGGAAGGCGAGAGTAGGGAACTGCCAGGCATCAAACTAAGCAGAAGGCCCCTGACGGATGGCCTTTTTGCGTTTCTACAAACTCTTTCTGTGTTGTAAAACGACGGCC
AGTCTTAAGCTCGGGCCCCCTGGGCGGTTCTGATAACGAGTAATCGTTAATCCGCAAATAACGTAAAAACCCGCTTCGGCGGGTTTTTTTATGGGGGGAGTTTAGGGAAAGA
GCATTTGTCAGAATATTTAAGGGCGCCTGTCACTTTGCTTGATATATGAGAATTATTTAACCTTATAAATGAGAAAAAAGCAACGCACTTTAAATAAGATACGTTGCTTTTTC
GATTGATGAACACCTATAATTAAACTATTCATCTATTATTTATGATTTTTTGTATATACAATATTTCTAGTTTGTTAAAGAGAATTAAGAAAATAAATCTCGAAAATAATAAA
GGGAAAATCAGTTTTTGATATCAAAATTATACATGTCAACGATAATACAAAATATAATACAAACTATAAGATGTTATCAGTATTTATTATGCATTTAGAATAAATTTTGTGTC
GCCCTTAATTGTGAGCGGATAACAATTACGAGCTTCATGCACAGTGAAATCATGAAAAATTTATTTGCTTTGTGAGCGGATAACAATTATAATAGATTCGTTGAGCGCTC
ACAATTCCACACAACGGTTTCCCTCTAGAAATAATTTTGTTTAACTTTTAGGAGGTAAAACATATGGCGAAGTATTACGGCAGTATGAAAATCGAAGAAGGTAAACTGGTAATC
TGGATTAACGGCGATAAAGGCTATAACGGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGAATTAAAGTCACCGTTGAGCATCCGGATAAACTGGAAGAGAAA
TTCCCACAGGTTGCGGCAACTGGCGATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTGGCTGAAATCACCCCGGACAAAGC
GTTCCAGGACAAGCTGTATCCGTTTACCTGGGATGCCGTACGTTACAACGGCAAGCTGATTGCTTACCCGATCGCTGTTGAAGCGTTATCGCTGATTTATAACAAAGATCTGC
TGCCGAACCCGCCAAAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAAGGTAAGAGCGCGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGC
CGCTGATTGCTGCTGACGGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTACGACATTAAAGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTGACCTTCCTGG
TTGACCTGATTAAAAACAAACACATGAATGCAGACACCGATTACTCCATCGCAGAAGCTGCCTTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGT
CCAACATCGACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGTCAACCATCCAAACCGTTCGTTGGCGTGCTGAGCGCAGGTATTAACGCCGCCAG
TCCGAACAAAGAGCTGGCAAAAGAGTTCCTCGAAAACTATCTGCTGACTGATGAAGGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCTTA
CGAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACTATGGAAAACGCCCAGAAAGGTGAAATCATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCG
TACTGCGGTGATCAACGCCGCCAGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGACTCGTATCACTAAAAGCGGAAGCGGAGAGAACCTGTACTTCCAATC
CCTGGTGCCGCGCGGCAGCGGATCCGCTTCAAAACCCGACATAAAGGTAGGAGATTACGTTAAAATGGGTGTTTATAATAACGCGAGCATTCTGTGGCGTTGCGTTTCTATC
GATAACAACGGTCCGTTAATGCTGGCAGACAAGATTGTGGATACCCTCGCATACGACGCGAAAACCAACGACAACTCTAACAGCAAGTCCCACAGCCGTAGCTACAAACGT
GATGATTATGGCTCCAACTACTGGAAGGACTCAAACATGCGTAGCTGGCTCAATAGCACGGCTGCGGAAGGTAAAGTAGACTGGCTGTGCGGTAACCCGCCCAAAGATGGC
TATGTTAGCGGTGTGGGTGCGTACAACGAAAAAGCCGGCTTTCTGAACGCATTTAGCAAGTCTGAAATCGCAGCAATGAAAACCGTTACCCAACGTTCTCTGGTTTCGCATC
CGGAGTACAACAAAGGTATCGTTGACGGCGACGCCAATTCTGATCTGTTGTACTACACAGACATCAGTGAGGCGGTTGCGAACTATGACTCTAGCTACTTCGAAACTACTAC
GGAGAAGGTGTTTCTGTTGGACGTTAAACAAGCAAATGCCGTGTGGAAAAACCTGAAGGGCTATTACGTGGCGTATAACAACGACGGTATGGCGTGGCCGTATTGGCTGCG
TACCCCGGTGACCGACTGTAACCACGATATGCGTTATATCAGTTCCAGCGGCCAGGTCGGTCGTTATGCTCCGTGGTATAGCGACCTGGGCGTGAGACCAGCATTTTATCTGG
ACTCTGAGTATTTTGTTACTACGAGCGGCTCTGGCTCCCAAAGCAGTCCGTATATCGGCTCCGCGCCGAACAAACAAGAAGATGACTACACCATTTCCGAGCCAGCAGAAGA
CGCCAACCCGGATTGGAACGTTAGTACCGAGCAGAGCATCCAACTGACCCTCGGCCCTTGGTATTCCAACGACGGTAAATACAGCAACCCGACCATCCCGGTTTACACCATC
CAAAAAACGCGTTCCGACACCGAGAATATGGTGGTGGTCGTGTCGGTGAGGGCTACACTAAGAGCCAGCAGGGTAAATTCATCAACGACGTAAAGCGCCTGTGGCAGGAT
GCTATGAAGTATGAGCCGTACCGCTCCTACGCGGATCGTTTCAATGTGTACGCTCTGTGTACCGCGTCAGAGAGCACCTTCGACAACGGTGGTTCTACGTTCTTTGATGTGAT
AGTTGACAAATACAACTCCCCGGTTATTTCCAACAACCTGCATGGTTCCCAATGGAAGAACCACATTTTTGAGCGTTGTATTGGTCCGGAGTTTATCGAAAAGATTCATGATG
CGCATATTAAAAAAAAGTGCGATCCGAATACCATTCCGTCGGGGAGCGAGTATGAACCGTACTATTACGTGCATGACTACATTGCGCAGTTTGCTATGGTTGTGAATACCAA
AAGTGACTTCGGCGGTGCGTACAATAACCGTGAATACGGCTTCCATTATTTCATCAGCCCGAGCGACAGCTATCGTGCAAGCAAAACCTTCGCGCACGAATTCGGTCACGGC
CTGCTTGGCTTGGGTGATGAATACTCCAACGGGTACCTGTTGGATGACAAGGAACTGAAGAGCCTGAACCTGAGCAGCGTCGAGGACCCGGAAAAAATCAAGTGGCGTCAA
TTGCTGGGTTTTCGCAACACATATACGTGCCGTAATGCGTACGGCTCTAAAATGCTGGTTTCTAGCTACGAGTGCATCATGCGCGACACGAATTATCAGTTCTGCGAAGTTTG
CCGCCTGCAAGGTTTCAAACGCATGAGCCAGTTGGTGAAGGATGTCGACTTGTATGTCGCCACCCCAGAGGTTAAAGAATACACCGGCGCATACAGCAAGCCGTCCGACTTT
ACCGATTTGGAAACCTCCAGCTACTACAACTACACCTATAACCGCAATGACCGCTTGCTTAGCGGTAATTCGAAGAGCCGTTTCAATACCAATATGAACGGTAAAAAAATCG
AGTTGCGTACCGTTATACAGAATATTAGCGACAAGAATGCGCGTCAACTGAAATTTAAGATGTGGATTAAACACAGCGATGGTAGCGTGGCAACGGACTCGTCCGGCAACC
CGCTGCAGACCGTCCAGACCTTCGATATTCCGGTTTGGAACGATAAGGCTAACTTTTGGCCTCTGGGTGCGTTGGACCACATCAAGAGCGACTTCAACAGCGGTCTGAAGAG
CTGTAGCCTGATTTATCAGATTCCGTCTGATGCCCAACTAAAGTCCGGCGACACCGTTGCGTTCCAGGTCCTGGACGAAAATGGCAATGTGTTGGCTGACGATAACACCGAA
ACCCAGCGCTATACTACGGTTAGCATTCAATATAAATTCGAAGACGGTTCGGAGATTCCGAATACCGCGGTGGCACCTTCACCGTTCCGTACGGTACGAAGCTGGATTTGA
CCCCAGCCAAAACGCTGTATGATTATGAATTTATCAAAGTTGATGGCTTAAATAAGCCGATTGTCTCCGACGGCACTGTGGTGACGTACTACTACAAGAACAAGAACGAGGA
ACATACCCATAATCTGACGCTGGTGGCGGCGAAGGCTGCTACCTGCACAACCGCGGGCAACTCAGCCTACTACACGTGCGATGGCTGTGATAAATGGTTTGCGGACGCGACA
GGCTCCGTCGAAATCACCGACAAAACCAGCGTGAAGATCCCGGCGCCGGGTCATACCGCGGGAACCGAGTGGAAAAGCGACGATACCAATCATTGGCACGAGTGCACTGTG
GCGGGTTGCGGTGTGATTATCGAGAGCACTAAGAGTGCGCACACCGCGGGTGAATGGATCGTGGATACCCCAGCTACCGCGACTACTGCTGGCACCAAGCACAAAGAGTGT
ACTGTGTGTCATCGTGTTCTTGAGACGCAACCGATTCCGAGCACCGGTACCGAGCTGAAGATCATAGCCGGCGACAACCAGATTTATAACAAAGCCAGCGGTAGCGATGTAA
CGATCACCTGTAATGGTGATTTTGCAAAATTCACCGGTATCAAAGTTGACGGCAGTGTCGTGGACAGCAGCAATTACACCGCGGTGTCCGGTTCTACGGTTCTGACACTGAA
AGCGAGTTATCTGGGCACTCTGACTGACGGTTCTCACACCATCACCTTCGTTTATACCGACGGCGAGGCGAACGCTAATTTAACGGTACGCACCGCTGGCAGCGGGCACATC
CACGATTATGGCACTGAGTGGAAATCCAATGCAGATAACCATTGGCACGAATGCAACTGCGGCGATAAGAAGGATGAGGCGGCTCACTCGTTCAAGTGGGTAGTCGATAAA
GAAGCCACCGCTACCAAAAAGGGTTCTAAGCACGAAGAGTGCAAAATCTGCGGTTATAAGCGCTCTGCGGTGGAAATCCCGGCGACCGGTACTTCAACCGCGCCAACCGAT
ACCACGAAACCGAATGACACCACCAAGCCGGGTAACACGAACGGCTCTGAAAAGGGATCCCACCATCACCATCACCACTAAGACTACAAAGACGATGACGACAAGTAACT
CGAGCCCCAAGGGCGACACCCCCTAATTAGCCCGGGCGAAAGGCCCAGTCTTTCGACTGAGCCTTTCGTTTTATTTGATGCCTGGCAGTTCCCTACTCTCGCATGGGGAGTCC
CCACACTACCATCGGCGCTACGGCGTTTCACTTCTGAGTTCGGCATGGGGTCAGGTGGGACCACCGCGCTACTGCCGCCAGGCAAACAAGGGGTGTTATGAGCCATATTCAG
GTATAAATGGGCTCGCGATAATGTTCAGAATTGGTTAATTGGTTGTAACACTGACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAAC
CCTGATAAATGCTTCAATAATATTGAAAAAGGAAGAATATGAGCCATATTCAACGGGAAACGTCGAGGCCGCGATTAAATTCCAACATGGATGCTGATTTATATGGGTATAA
ATGGGCTCGCGATAATGTCGGGCAATCAGGTGCGACAATCTATCGCTTGTATGGGAAGCCCGATGCGCCAGAGTTGTTTCTGAAACATGGCAAAGGTAGCGTTGCCAATGAT
GTTACAGATGAGATGGTCAGACTAAACTGGCTGACGGAATTTATGCCACTTCCGACCATCAAGCATTTTATCCGTACTCCTGATGATGCATGGTTACTCACCACTGCGATCCC
CGGAAAAACAGCGTTCCAGGTATTAGAAGAATATCCTGATTCAGGTGAAAATATTGTTGATGCGCTGGCAGTGTTCCTGCGCCGGTTGCACTCGATTCCTGTTTGTAATTGTC
CTTTTAACAGCGATCGCGTATTTCGCCTCGCTCAGGCGCAATCACGAATGAATAACGGTTTGGTTGATGCGAGTGATTTTGATGACGAGCGTAATGGCTGGCCTGTTGAACAA
GTCTGGAAAGAAATGCATAAACTTTTGCCATTCTCACCGGATTCAGTCGTCACTCATGGTGATTTCTCACTTGATAACCTTATTTTTGACGAGGGGAAATTAATAGGTTGTATT
GATGTTGGACGAGTCGGAATCGCAGACCGATACCAGGATCTTGCCATCCTATGGAACTGCCTCGGTGAGTTTTCTCCTTCATTACAGAAACGGCTTTTTCAAAAATATGGTAT
TGATAATCCTGATATGAATAAATTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAAGCGGCGCGCCATCGAATGGCGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCG
GAAGAGAGTCAATTCAGGGTGGTGAATATGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCGTTTCCCGCGTGGTGAACCAGGCCAGCC
ACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGCGGCGATGGCGGAGCTGAATTACATTCCCAACCGCGTGGCACAACAACTGGCGGGCAAACAGTCGTTGCTGATTGGCG
TTGCCACCTCCAGTCTGGCCCTGCACGCGCCGTCGCAAATTGTCGCGGCGATTAAATCTCGCGCCGATCAACTGGGTGCCAGCGTGGTGGTGTCGATGGTAGAACGAAGCGG
CGTCGAAGCCTGTAAAGCGGCGGTGCACAATCTTCTCGCGCAACGCGTCAGTGGGCTGATCATTAACTATCCGCTGGATGACCAGGATGCCATTGCTGTGGAAGCTGCCTGC
ACTAATGTTCCGGCGTTATTTCTTGATGTCTCTGACCAGACACCCATCAACAGTATTATTTTCTCCCATGAGGACGGTACGCGACTGGGCGTGGAGCATCTGGTCGCATTGGG
TCACCAGCAAATCGCGCTGTTAGCGGGCCCATTAAGTTCTGTCTCGGCGCGTCTGCGTCTGGCTGGCTGGCATAAATATCTCACTCGCAATCAAATTCAGCCGATAGCGGAAC
GGGAAGGCGACTGGAGTGCCATGTCCGGTTTTCAACAAACCATGCAA

GSGSASKPDIKVGDYVKMGVYNNASILWRCVSIDNNGPLMLADKIVDTLAYDAKTNDNSNSKSHSRSYKRDDYGSNYWKDSNMRSWLNSTAAEGKVDWLCGNPPKDGYVSGV
GAYNEKAGFLNAFSKSEIAAMKTVTQRSLVSHPEYNKGIVDGDANSDLLYYTDISEAVANYDSSYFETTTEKVFLLDVKQANAVWKNLKGYYVAYNNDGMAWPYWLRTPVTDC
NHDMRYISSSGQVGRYAPWYSDLGVRPAFYLDSEYFVTTSGSGSQSSPYIGSAPNKQEDVTISEPAEDANPDWNVSTEQSIQLTLGPWYSNDGKYSNPTIPVYTIQKTRSDTENMV
VVVCGEGYTKSQQGKFINDVKRLWQDAMKYEPYRSYADRFNVYYALCTASESTFDNGGSTFFDVIVDKYNSPVISNNLHGSQWKNHIFERCIGPEFIEKIHDAHIKKKKCDPNTIPSGSE
YEPYYYVHDYIAQFAMVVNTKSDFGGAYNNREYGFHYFISPSDSYRASKTFAHEFGHGLLGLGDEYSNGYLLDDKELKSLNLSSVEDPEKIKWRQLLGFRNTYTCRNAYGSKMLV
SSYECIMRDTNYQFCEVCRLQGFKRMSQLVKDVDLYVATPEVKEYTGAYSKPSDFTDLETSSYYNYTYNRNDRLLSGNSKSRFNTNMNGKKIELRTVIQNISDKNARQLKFKMWIK
HSDGSVATDSSGNPLQTVQTFDIPVWNDKANFWPLGALDHIKSDFNSGLKSCSLIYQIPSDAQLKSGDTVAFQVLDENGNVLADDNTETQRYTTVSIQYKFEDGSEIPNTAGGTFTVP
YGTKLDLTPAKTLYDYEFIKVDGLNKPIVSDGTVVTYYYKNKNEEHTHNLTLVAAKAATCTTAGNSAYYTCDGCDKWFADATGSVEITDKTSVKIPAPGHTAGTEWKSDDTNHW
HECTVAGCGVIIESTKSAHTAGEWIVDTPATATTAGTKHKECTVCHRVLETQPIPSTGTELKIIAGDNQIYNKASGSDVTITCNGDFAKFTGIKVDGSVVDSSNYTAVSGSTVLTLKAS
YLGTLTDGSHTITFVYTDGEANANLTVRTAGSGHIHDYGTEWKSNADNHWHECNCGDKKDEAAHSFKWVVDKEATATKKGSKHEECKICGYKRSAVEIPATGTSTAPTDTTKPN
DTTKPGNTNGSEKGSHHHHHH