UNIVERSITY OF CALIFORNIA, SAN DIEGO

Understanding receptor specificity through the massively variable major tropism determinant of *Bordetella* bacteriophage

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

in

Chemistry

by

Jason L. Miller

Committee in charge

Professor Partho Ghosh, Chair
Professor Raffi Aroian
Professor Gourisankar Ghosh
Professor Simpson Joseph
Professor Charles Perrin

2006

This dissertation of Jason L. Miller is approved, and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2006

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my advisor and scientific mentor, Dr. Partho Ghosh, for providing the direction, knowledge, and support toward my graduate education.

I am grateful to my colleagues from the Ghosh lab for their participation in five of the most intellectual and entertaining years of my life. My special thanks to Stephen McMahon, Jeff Lawton, and Don Kerkow for their involvement with the Mtd project and establishing the foundation for my graduate work. Thanks also to Jenn Whiles-Lillig, Mike Marino, Lori Beutow, Sara Birtalan, Mani Banerjee, and Jonathan Mikolosko, who always made themselves available to share their scientific knowledge. I thank the graduate students who follow my footsteps, Loren and Alicia, for all your support during the completion of my graduate work. Special thanks go to Jyothi Kumaran and Brent Hamaoka for sharing their experiences and making strong contributions to my scientific training. Last and certainly not least, my most heartfelt gratitude goes to Rebecca Phillips and Case McNamara. Rebecca has been an incredible friend and scientific colleague during my tenure at UCSD; our surfing experiences will never be forgotten. My sports talks and friendship with Case has made a challenging graduate career extremely enjoyable and I thank him for postponing his leave from the Ghosh lab until my defense.

I am very grateful to my collaborators, Jeff Miller, Asher Hodes and Roman Barbalat who without their help, I would not have been able to carry out so many

successful experiments. In addition, I would like to thank Simon Bergqvist and Betsy Komives for their training and advice with my binding kinetics assays. My appreciation goes to both the Gouri Ghosh and Alex Hoffman lab members for their shared resources and experimental assistance.

And to my family and friends, thank you for all the love and support. To my mom and dad, I thank you for providing me with so many opportunities in life and for encouraging me to make most of my talents. I want to express my love and gratitude to my wife, Monica, who has provided her strength and support during these difficult years of my life and made every success worth while.

The text of chapter 2, in full, is a reprint of the material as it appears in the journal Nature Structural and Molecular Biology. The dissertation author was the primary researcher and/or author and the co-authors (S. McMahon, J. Lawton, D. Kerkow, A. Hodes, M. Marti-Renom, S. Doulatov, E. Narayanan, A. Sali, J.F. Miller, and P. Ghosh) listed in this publication contributed to or supervised the research which forms the basis for this chapter.

The text of chapter 3, in full, is a reprint of the material in current preparation for publication. The dissertation author was the primary researcher and/or author and the co-authors (A. Hodes, R. Barbalat, J.F. Miller P. Ghosh) listed in this publication contributed to or supervised the research which forms the basis for this chapter.

VITA

2000 B.A., Chemistry and Biology, California State University, Sacramento

2003 M.S.., Chemistry, University of California, San Diego

2006 Ph.D., Chemistry, University of California, San Diego

PUBLICATIONS

McMahon, S.A., Miller, J.L., Lawton, J.A., Kerkow, D.E., Hodes, A., Marti-Renom, M.A., Doulatov, S., Narayanan, E., Sali, A., Miller, J.F., Ghosh, P. The C-type lectin fold as an evolutionary solution for massive sequence variation. *Nat Struct Mol Biol*. 2005 Oct;12(10):886-92.

Miller, J.L., Hodes, A., Barbalat, R., Miller, J.F., Ghosh, P. Receptor recognition by Mtd, the massively sequence variable protein of a diversity-generating retroelement. *In preparation.*

ABSTRACT OF THE DISSERTATION

Understanding receptor specificity through the massively variable major tropism determinant of *Bordetella* bacteriophage

by

Jason L. Miller

Doctor of Philosophy in Chemistry

University of California, San Diego, 2006

Professor Partho Ghosh, Chair

Only a few protein folds are known to tolerate massive sequence variation for the sake of binding diversity. The immunoglobulin (Ig)-fold found in the structural design of antibodies and T-cell receptors as well as the leucine-rich repeat-fold (LRR's) discovered in the variable lymphocyte receptors of the agnathous hagfish are capable of withstanding recombinant variability on the order of 10^{14}-10^{16} unique protein sequences. Our studies have elucidated the structure of *Bordetella* bacteriophage major tropism determinant (Mtd) which utilizes the C-type lectin (CLec) fold to accommodate variation for nearly 10^{13} sequences. Additionally, we

have identified other diversity generating retroelement (DGR) encoded proteins in prokaryotes and bacteriophage that also use the CLec-fold for sequence variation.

The structure of Mtd is stabilized as an intertwined, pyramidal shaped trimer consisting of three distinct domains with the variable residues organized as discrete receptor binding sites located near the C-termini. The variable residues are presented on surface exposed loops that are stabilized by a network of hydrogen bonds along the C-Lec-fold resulting in a highly static scaffold for mutagenic variation. Association between Mtd-P1 variant with its native receptor, pertactin ectodomain (Prn-E) was characterized and used as a model for understanding receptor binding specificity.

Mtd-P1 was co-crystallized with Prn-E and the structure solved to 3.16 Å resolution. We show that despite a difference in structure fold and interaction modalities, Mtd uses its variable residues for target recognition in an antibody-like manner. In addition, we have identified and characterized an antibody-like affinity maturation process where a non-productive interaction is converted to a productive one. More significantly, we have discovered that phage avidity acts as a differential amplifier for binding affinity and specificity. These results indicate that the C-Lec fold acts as a stable platform for sequence variation with enough flexibility for multiple receptor recognition. Finally, our evidence suggests that multivalency is conserved in conferring efficacy and specificity to sequence variable repertoires.

I.

Introduction

Introduction

Complex behavior in biological systems is encoded at many levels, but one of the most fundamental involves specific recognition between macromolecules. Understanding the basis for specificity of macromolecular interactions remains an important and unsolved problem in biomedicine. Although antibody-antigen interactions have been extensively characterized and served as a model for understanding specificity, whether or not this generally applies to ligand-receptor interactions is unknown. Current technologies for designing and manipulating ligands include *in vivo* selection by fusing peptides to bacteriophage (phage display) or expressing recombinant proteins on bacterial cell surfaces (Benhar, 2001). These affinity selection assays are based on the premise that small changes in peptide sequence design result in unique structural changes that will allow a favorable ligand-protein interaction. A rich area for exploring this type of recognition is in host-pathogen interactions, since millions of years of evolutionary pressure have led to the selection of highly specific macromolecular interactions between pathogen and host.

A remarkable host-pathogen system for addressing specificity was recently characterized between a temperate bacteriophage that infects *Bordetella bronchiseptica* (Liu, 2002). Remarkably, the phage changes host tropism by genetically varying the sequence of a host receptor-binding protein, called major tropism determinant (Mtd, 40-kD). Depending on the virulence stage of its host, various antigen receptors are expressed to the surface of *Bordetella*. Upon infection, the BvgAS signal transduction pathway of *Bordetella* subspecies is activated (Bvg^+

phase), resulting in the expression of toxins, adhesion molecules, and a type III secretion system (Akerley et al., 1995; Uhl and Miller, 1996). During this phase, *Bordetella* secretes the cell surface antigen, pertactin, which allows the pathogen to adhere to pulmonary epithelial cells in the mucosal layers of the mammalian respiratory tract (Emsley et al., 1996). Inactivation of the BvgAS signal transduction system (Bvg$^-$ phase) results in the repression of virulence and colonization factors, followed by induction of motility genes essential for *ex vivo* survival by *Bordetella (Akerley et al., 1995; Uhl and Miller, 1996).* At intermediate activation levels of BvgAS, other genes such as bipA are preferentially expressed at the cell surface (Cotter and Miller, 1997).

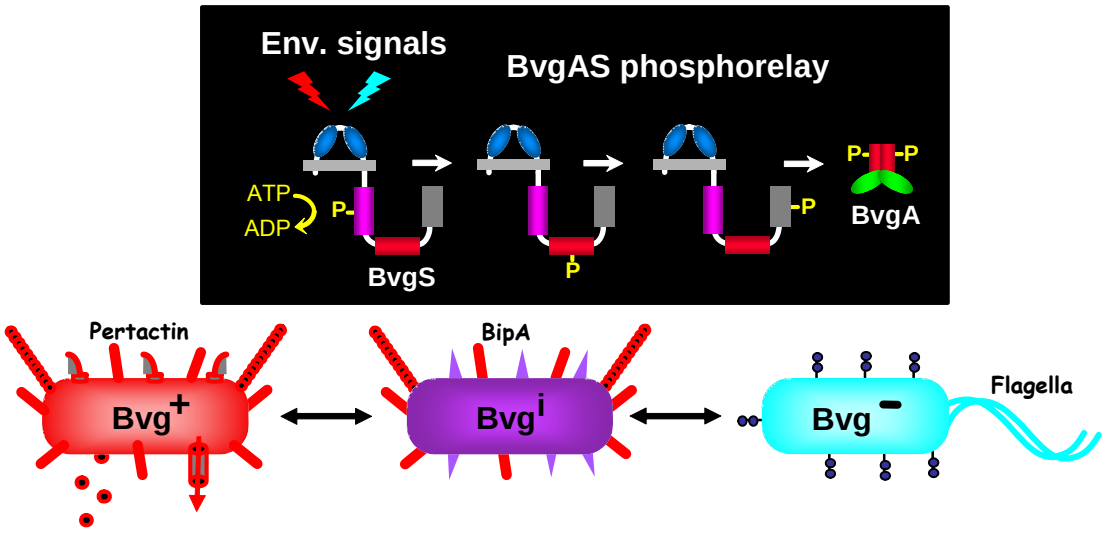


Figure 1.1. Schematic of *Bordetella* BvgAS signal transduction pathway. (Above) A two component phosphorelay system involves activation of BvsS receptor complex by environmental signals, resulting in autophosphorylation and phosphate transfer to transcriptional response regulator, BvgA. (Below) Activation of BvgAS during infection (Bvg$^+$) leads to expression of pertactin to *Bordetella* surface. During intermediate levels of activation (Bvg^i) BipA is preferentially expressed. Inactivation of BvgAS (Bvg$^-$) leads to motility gene expression such as flagella production. Figure adapted from Jeff Miller, UCLA.

Studies have shown that the phage require Mtd to attach to receptors present during these phases (Doulatov, 2004; Liu, 2002). Deletion of Mtd results in non-infectious phage particles with tail fibers absent (Liu, 2002). The diversity of this gene is created through a diversity-generating retroelement (DGR) contained with the *Bordetella* phage genome. The phage DGR includes Mtd with a C-terminal variable region (VR), a template region (TR) with >90% identity to the VR, and a reverse transcriptase (RT) with sequence similarity to group II introns and retroviral reverse transcriptases (Liu, 2002; Liu, 2004). Mtd variation occurs as a result of RT introducing errors at specific adenines located within TR, then replacement of the C-terminal variable region (VR) of Mtd with the mutant copy of TR. The only adenines left unperturbed are located within codons for the terminal eight amino acids of Mtd, that serve as an initiation site for mutagenic homing (IMH). The IMH sequence is responsible for the unidirectional transfer of sequence information from TR to the Mtd VR (Doulatov, 2004).

Consequently, Mtd variability is specific to 12 adenine-encoded amino acids scattered throughout the C-terminal 45 residues of the VR. Of these 12 residues, 10 result from the template codon AAC, capable of 14 potential unique amino acids from adenine-substitution. The other 2 residues contain adenine as the first base of their template codons, resulting in 3 possible amino acid variations. As a result, the theoretical diversity of unique Mtd variants is calculated as:

$$15^{10} \times 4^{2} = 9.2 \times 10^{12} \text{ Possible Mtd Variants}$$

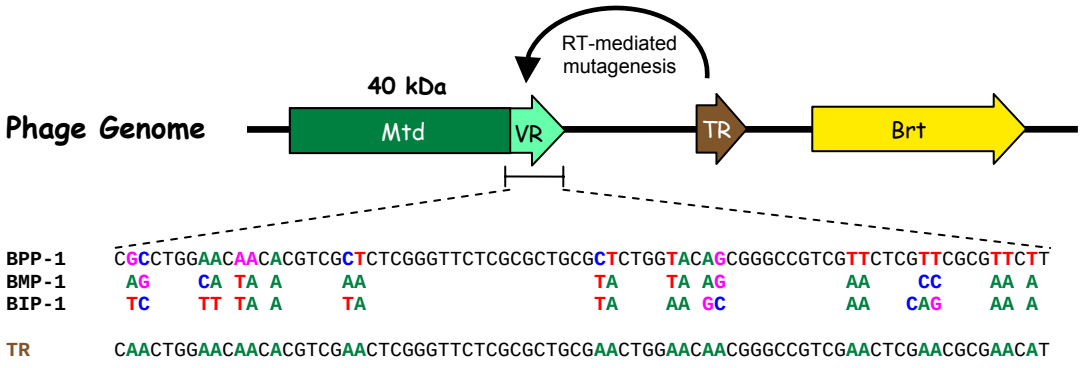


Figure 1.2. *Bordetella* phage diversity generating retroelement (DGR). The genomic organization of the phage DGR includes the Mtd gene with a variable region (VR) located at 3' end. Downstream of Mtd locus resides the template region (TR) with 90% identity to VR. Located immediately following TR, the *Bordetella* phage reverse transcriptase (Brt) gene is responsible for mutagenesis of TR for directed transfer and replacement of the Mtd VR. Alignment of TR with VR's from 3 different phage display 12 codons (colored) with mutable adenines provided by TR for Mtd amino acid variability.

This variability has allowed the phage to recognize different surface proteins by modifying the C-terminal sequence of the Mtd gene product. Five unique Mtd variants whose amino acid sequence have been characterized include: Mtd-P1 isolated from Bvg^+ tropic phage dependent on pertactin (BPP-1), Mtd-P3c from Bvg^+ tropic phage with no dependence on pertactin for infection (BPP-3c), Mtd-M1 isolated from Bvg^- tropic phage (BMP-1), Mtd-I1 found in indiscriminant phage capable of infecting both Bvg^+ or Bvg^- phage (BIP-1), Mtd-U1 potentially from non-infective phage (BUP-1) but with a VR resembling the TR. The three distinct phage tropisms switch at

characteristic frequencies. BPP-1 forms plaques on Bvg^+ phase *Bordetella* at a 10^6-fold higher efficiency than on Bvg^- phase cells. Similar preferential infectivity was seen in BMP-1 phage for Bvg^- *Bordetella*. Additionally, switching between BIP and BMP phage occurs at frequencies similar to return switching to the BPP tropism.

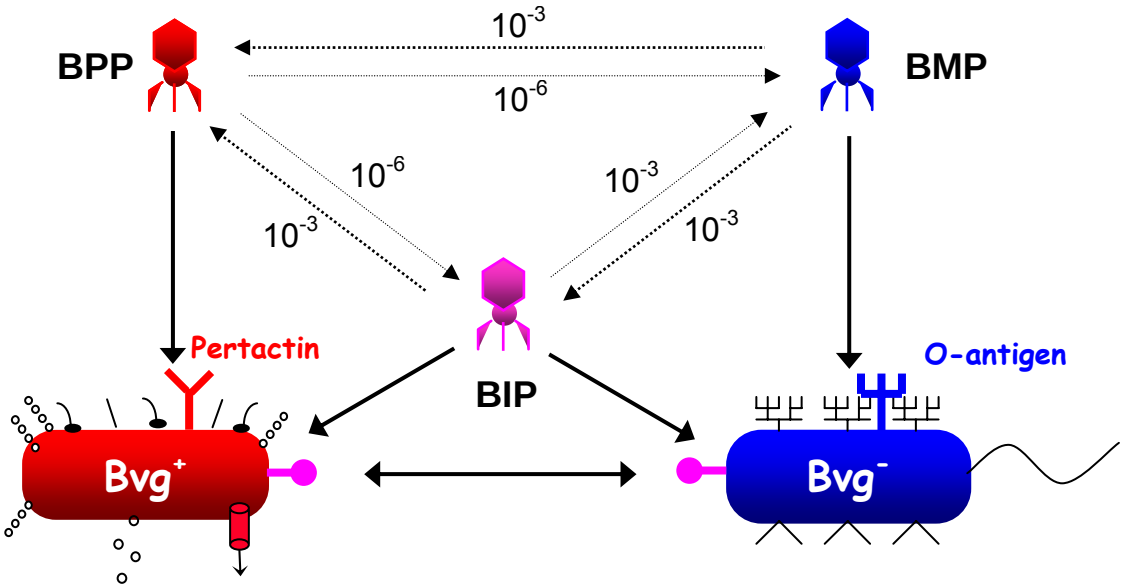


Figure 1.3. *Bordetella* phage switching frequencies. Bvg-plus phage (BPP) use pertactin as receptor for infection. Indirect evidence points to O-antigen as a potential receptor for Bvg-minus phage (BMP). Bvg-indiscriminant phage infects *Bordetella* in both plus or minus phase, indicating it recognizes a receptor expressed in both phases. Image borrowed from Doulatov, S. et al., Nature, 2004.

Previous studies have shown BPP-1 plaquing efficiency is decreased in *Bordetella* when the surface antigen, pertactin, was deleted from the genome. Ectopic expression of pertatin restored full infectivity by BPP-1, implicating pertactin as the primary determinant for binding (Liu, 2002). Indirect evidence suggests BMP-1 recognizes the glycosyl groups of the O-antigen (unpublished data). Currently,

infectious phage requiring intermediate levels of BvgAS activation have not yet been identified.

The ability of *Bordetella* phage to switch receptor specificities for its own survival mirrors the adaptive immune response tailoring antibodies for antigen recognition. Only the Ig-fold of antibodies and T-cell receptors, capable of producing ~10^{14}-10^{16} sequence variants (Davis and Bjorkman, 1988), and the leucine-rich repeat (LRR)-fold of the agnathous hagfish lymphocyte receptors with an estimated 10^{14} sequences (Pancer, 2004), are comparable to the ~10^{13} variable sequences of Mtd. Until now, Ig-fold immune proteins and their interactions with antigenic targets have served as the sole paradigm for understanding recognition by sequence variable repertoires.

The goal of the research presented in this dissertation is to understand how sequence alterations in the variable region are tolerated structurally by Mtd, and likewise how these variations lead to formation of a wide array of receptor-binding sites. In addition, we wanted to characterize the interaction between Mtd-P1 variant and one of its host receptors, pertactin, as model for ligand-receptor specificity. These studies show that the Mtd variable binding domain is stabilized by a trimeric C-type lectin fold, thus presenting a novel structural solution for stabilization of highly variable proteins. We also provide evidence that the C-type lectin fold is used for accommodating massive sequence variation for numerous prokaryotic organisms and

bacteriophage with related DGR's. Structural elucidation and direct binding studies of Mtd with pertactin demonstrate similarities to antibody recognition of antigen targets. Furthermore, we show that phage avidity acts as differential amplifier to increase affinity and specificity to *Bordetella* surface receptors.

Bibliography

Akerley, B. J., Cotter, P. A., and Miller, J. F. (1995). Ectopic expression of the flagellar regulon alters development of the Bordetella-host interaction. Cell *80*, 611-620.

Benhar, I. (2001). Biotechnological applications of phage and cell display. Biotechnology Advances *19*, 1-33.

Cotter, P. A., and Miller, J. F. (1997). A mutation in the Bordetella bronchiseptica bvgS gene results in reduced virulence and increased resistance to starvation, and identifies a new class of Bvg-regulated antigens. Mol Microbiol *24*, 671-685.

Davis, M. M., and Bjorkman, P. J. (1988). T-cell antigen receptor genes and T-cell recognition. Nature *334*, 395-402.

Doulatov, S. (2004). Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. Nature *431*, 476-481.

Emsley, P., Charles, I. G., Fairweather, N. F., and Isaacs, N. W. (1996). Structure of Bordetella pertussis virulence factor P.69 pertactin. Nature *381*, 90-92.

Liu, M. (2002). Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage. Science *295*, 2091-2094.

Liu, M. (2004). Genomic and genetic analysis of Bordetella bacteriophages encoding reverse transcriptase-mediated tropism-switching cassettes. J Bacteriol *186*, 1503-1517.

Pancer, Z. (2004). Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. Nature *430*, 174-180.

Uhl, M. A., and Miller, J. F. (1996). Integration of multiple domains in a two-component sensor protein: the Bordetella pertussis BvgAS phosphorelay. EMBO J *15*, 1028-1036.

II.

The C-type lectin fold as an evolutionary solution for massive sequence variation

Abstract

Only few instances are known of protein folds that tolerate massive sequence variation for the sake of binding diversity. The most extensively characterized is the immunoglobulin (Ig)-fold. We now add to this the C-type lectin (CLec)-fold, as found unexpectedly in major tropism determinant (Mtd), a retroelement-encoded receptor-binding protein of *Bordetella* bacteriophage. Variation in Mtd, with its ~10^{13} possible sequences, enables phage adaptation to *Bordetella*. Mtd is an intertwined, pyramid-shaped trimer, with variable residues organized by its CLec-fold into discrete receptor-binding sites. The CLec-fold is seen to provide a strikingly static scaffold for combinatorial display of variable residues, likely reflecting a different evolutionary solution for balancing diversity against stability than in the Ig-fold. A bias in Mtd variants for the receptor pertactin is found, and evidence that the CLec-fold is used broadly for sequence variation by related retroelements is presented.

Introduction

Major tropism determinant (Mtd, 40 kDa), the receptor-binding protein of *Bordetella* bacteriophage, varies massively in sequence (Liu, 2002; Liu, 2004). Variation in Mtd depends on a phage-encoded retroelement that belongs to a novel family of retroelements implicated in generating sequence diversity in various phage and bacterial genomes (Doulatov, 2004). Remarkably, the *Bordetella* bacteriophage retroelement can produce ~10^{13} different sequence variants of Mtd, rivaling the ~10^{14}–10^{16} possible sequences of antibodies and T-cell receptors, whose immunoglobulin (Ig)-fold has been the sole paradigm for how proteins tolerate massive sequence variation (Chothia and Lesk, 1987; Davis and Bjorkman, 1988). Lack of similarity in Mtd to other proteins suggests that it represents a novel evolutionary solution to this problem. We therefore set out to understand how Mtd accommodates massive sequence variation and creates diverse receptor-binding sites.

Diversity in Mtd is required for phage adaptation to *Bordetella* as this bacterial pathogen varies its gene expression pattern according to infectious cycle. This expression pattern is regulated by the BvgAS two-component system, such that virulence factors, for example adhesins and the type III secretion system, are expressed only in the pathogenic or Bvg^+ phase of *Bordetella* (Akerley et al., 1995; Uhl and Miller, 1996). In contrast, other genes, including those required for motility, are expressed only in the environmental or Bvg^- phase. Certain other genes are expressed preferentially at intermediate levels of Bvg activation (Cotter and Miller, 1997). For the phage, this means that potential receptors appear and disappear as

Bordetella responds to its environment. Changes in *Bordetella* are met by emergence of Mtd variants that maintain phage infectivity by utilizing newly expressed *Bordetella* surface molecules as receptors.

Mtd variants are produced by a unique adenine-specific mutagenesis process involving a retroelement-encoded reverse transcriptase (*Bordetella* reverse transcriptase, Brt) (Fig. 2.1). The process relies on two nearly identical direct repeats, called variable (VR) and template regions (TR) (Liu, 2002; Liu, 2004). In brief, sequence information in the variable region, which encodes the C-terminal 45 amino acids of Mtd, is replaced by sequence information from the template region, which remains unchanged by this process. The mutagenic aspect comes from the fact that adenines in the template region are transmitted to the variable region with poor fidelity, being replaced at random by any base (Liu, 2002). As a consequence, variability in Mtd is focused to 12 adenine-encoded amino acids that are scattered across its C-terminal variable region (Fig. 2.1).

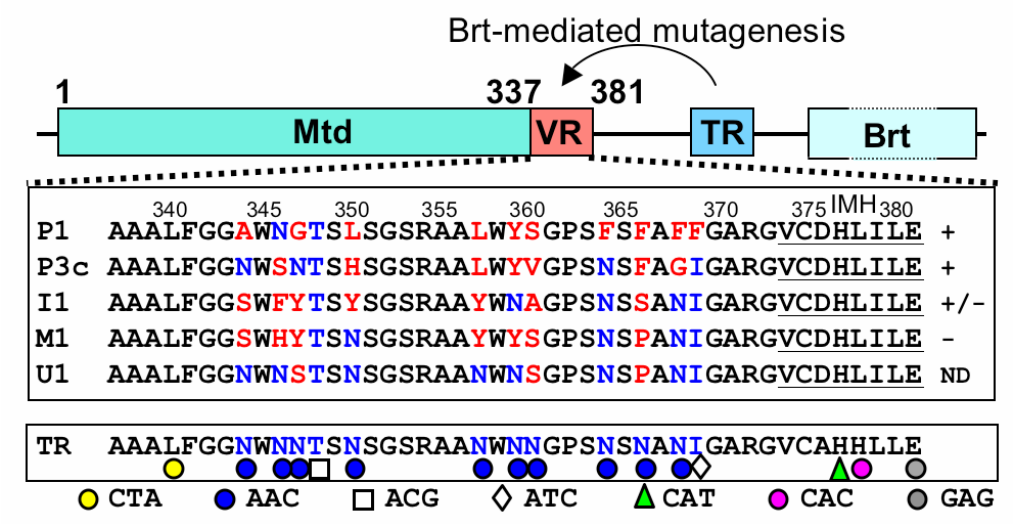


Figure 2.1. Mtd variation. Genome organization of the *Bordetella* bacteriophage retroelement (to scale, except for Brt). Variable region (VR) sequences for the five Mtd variants studied and the predicted sequence of the template region (TR) are shown (variable positions differing between VR and TR in red, and identical to TR in blue). Region corresponding to the initiation of mutagenic homing (IMH) sequence is underlined. Tropism is indicated to the right of sequences by +, Bvg^+; –, Bvg^-; +/–, either Bvg^+ or Bvg^-; ND, not determined. Adenine-containing codons in TR are denoted by colored symbols. Adenine-specific mutagenesis of codons denoted by blue or white symbols results in nonsynonymous substitutions; the three adenine-containing codons at the end of the sequence are part of the IMH element and invariant.

Experimental Procedures

Expression and Purification

Ancestries of Mtd-P1, Mtd-M1, Mtd-I1, and Mtd-P3c have previously been described (Liu, 2002). Coding sequences of Mtd (1–381) variants and Prn-E (38–640) were amplified by PCR from B. bronchiseptica RB50 or phage lysates. Mtd-U1 was obtained from a phage lysate induced from a BMP-1 lysogen.Mtd-M1 was expressed in E. coli BL21 (DE3) as an N-terminally his-tagged protein using pET28b (Novagen), and all other Mtd variants were expressed as glutathione S-transferase (GST) fusion proteins using pGEX-2T (Pharmacia). After being grown to A_{600} 0.6 at 37 °C, bacterial cultures were induced with 1 mM IPTG and grown overnight at 25 °C. Bacteria expressing his-tagged Mtd-M1 were lysed by sonication in 500 mM NaCl, 50 mM phosphate buffer, pH 8.0, 5-25 mM imidazole, and 1 mM PMSF; the lysate supernatant was applied to a metal chelation column (Poros MC/M), and Mtd-M1 was eluted from the column using an imidazole gradient (25-500 mM). Mtd-M1 was dialyzed in thrombin cleavage buffer (150 mM NaCl, 50 mM Tris or phosphate buffer, pH 8.0, and including or not 2 mM $CaCl_2$) and the his-tag was removed by overnight thrombin cleavage (2-3 units/mg Mtd-M1). Mtd-M1 was dialyzed in 150 mM NaCl, 10-50 mM Tris, pH 8.0 and further purified by metal chelation (Poros MC/M) and size exclusion chromatography (Superdex 200).

Bacteria expressing other GST-Mtd fusion proteins were lysed by sonication in 150 mM NaCl, 50 mM phosphate buffer, pH 7.4, 2 mM DTT, 2 mM EDTA, 1 mM PMSF, and 1% Triton X-100. Lysate supernatants were applied to glutathione-agarose (Sigma) columns, and GST-Mtd fusion proteins were eluted using 10 mM reduced glutathione, 50 mM Tris, pH 8.0. GST-Mtd fusion proteins were cleaved with thrombin as above to separate GST from Mtd, dialyzed in 50 mM Tris pH 8.0, and applied to an anion exchange column (Poros 20HQ), which retains GST but not Mtd. Mtd was then dialyzed in 150 mM NaCl, 50 mM Tris, pH 8.0 and further purified by size exclusion chromatography (Superdex 200). Purified Mtd was dialyzed in 150 mM NaCl, 10 mM Tris, concentrated to ~20 mg ml^{-}1 (e280 65361 M^{-}1cm^{-}1 for Mtd-P1), and snap-frozen in liquid N_2 for storage.

For selenomethionine-substituted Mtd protein, similar procedures were followed except that the E. coli strain B834 (DE3) was used for expression and was grown in synthetic minimal media supplemented with 200 mg/L seleno-D/L-methionine (Sigma) as described previously (Budisa et al., 1995). Selenomethionine-labeled Mtd was purified as described for unlabeled protein, except that buffers were generally supplemented with 2-10 mM β-mercaptoethanol throughout purification, and the purified protein was crystallized in the presence of 2 mM DTT. Prn-E was expressed as an N-terminally his-tagged protein in pET28b (having the vector derived sequence MGSSHHHHHHSSGLVPRGSHM at its N-terminus) as inclusion bodies in E. coli BL21 (DE3). Published procedures were followed for its refolding and purification (Emsley et al., 1994). Prn-E was dialyzed in 100 mM NaCl, 10 mM Tris,

pH 8.0, concentrated to ~25 mg ml^{-1} (e280 45857 $M^{-}1cm^{-}1$), and snap-frozen in liquid N_2 for storage.

Crystallization

Crystals of Mtd variants were grown by vapor diffusion at 25°C by mixing equal volumes of protein with various precipitants. Crystals of native Mtd-P1 grown in 30% (w/v) PEG 550 MME, 100 mM HEPES, pH 7.5, and 50mM $CaCl_2$ were used for data collection. These crystals belong to space group P63, and the asymmetric unit contains an Mtd monomer and ~60% solvent. Crystals were cryoprotected in the crystallization solution, and flash cooled to ~110K for data collection at 1.1055 Å wavelength (ID-19, APS, Argonne, IL).

Crystals of Mtd-P1 L52M/V53M containing seleno-methionine were grown in 24% (w/v) PEG 550 MME, 100mM HEPES, pH 7.5, and 50mM $MgCl_2$, belong to space group P213, and contain an Mtd monomer and ~50% solvent in the asymmetric unit. Crystals were cryoprotected in crystallization solution supplemented up to 30% (w/v) PEG 550 MME, and flash cooled to ~110K for data collection (ID-19, APS, Argonne, IL).

Crystals of Mtd-M1 and Mtd-I1 were grown in 34% (v/v) isopropanol, 300 mM sodium citrate, and 100 mM HEPES, pH 7.5, belong to space group P63, and contain an Mtd monomer and ~60% solvent in the asymmetric unit. Crystals were cryoprotected in crystallization solution supplemented with 15% (Mtd-M1) or 30%

(Mtd-I1) glycerol, and flash cooled to ~110K for data collection at 1.54179 Å wavelength (UCSD X-ray Facility, rotating anode source).

Crystals of Mtd-P3c were grown in 0.5 M Li_2SO_4, 0.75 M NaCl, and 100 mM Tris, pH 8.5, belong to space group P6322, and contain an Mtd monomer and ~50% solvent in the asymmetric unit. Crystals were cryoprotected in crystallization solution supplemented with 30% glycerol, and flash cooled to ~110K for data collection at 1.2309 Å wavelength (ID-19, APS, Argonne, IL).

Crystals of Mtd-U1 were grown in 15% PEG 4K, 100 mM MOPS, pH 6.5, and 0.2 M $MgCl_2$, belong to space group I222, and contain an Mtd trimer and ~65% solvent in the asymmetric unit. Crystals were cryoprotected in crystallization solution supplemented up to 30% PEG 4K, and flash cooled to ~110K for data collection at 0.9792 Å wavelength (ID-19, APS, Argonne, IL).

Structure Determination

For phase determination by MAD, methionine substitutions were introduced in Mtd-P1 at Leu52 and Val53 via Quikchange (Stratagene), and seleno-methionine was biosynthetically incorporated into the mutated protein, which was then purified and crystallized. Inverse-beam, three-wavelength diffraction data to 1.5 Å resolution were collected at ~110K from cryoprotected crystals (19-ID, Advanced Photon Source), and processed and scaled using HKL2000 (Otwinowski and Minor, 1997). Positions of three seleno-methionines (Met100, Met182, Met245) and initial phases were determined using SOLVE (Terwilliger, 1997), and solvent flattening was performed

using DM (Winn, 2003). Automated model building with ARP/wARP (Perrakis, 1997) produced a trace of residues 5–380 (Table 2.1). This model was used for molecular replacement with Amore (Winn, 2003) of native Mtd-P1, -P3c, -I1, -M1, and -U1 (Table 2.2). O (Jones et al., 1991) was used for manual model building, and REFMAC5 (Winn, 2003) for refinement of 95% of data (remaining 5% used for validation refinement protocol). Waters having $\geq 3\sigma$ F_o–F_c density and within 2.6–3.4 Å of a hydrogen bond donor or acceptor were modeled. At least one divalent cation was modeled and refined in each structure. All variable residues are within electron density in $2F_o$–F_c maps calculated with model phases, except for His346 in Mtd-M1 and Phe346 in Mtd-I1; density is seen only to the Cβ atom for these residues. Coordinates and structure factors have been deposited in the Protein Data Bank (PDB accession codes 1YU0, 1YU1, 1YU2, 1YU3, 1YU4).

Coprecipitation Assays

Prn-E (80 μM, his-tagged) was incubated (5 min, 25 °C) with Mtd variants (42 μM trimer) in binding buffer (150 mM NaCl, 50 mM Tris, pH 8.0). Ni^{2+}-NTA agarose beads were added and incubated for 5 min with the protein mixture, after which beads were washed three times with binding buffer supplemented with 25 mM imidazole. Bound proteins were eluted with binding buffer supplemented with 500 mM imidazole and visualized by SDS-PAGE and Coomassie staining. The final wash was found to be devoid of protein.

Table 2.1. Data collection, phasing, and refinement statistics for MAD (SeMet) Mtd-P1 structures.

	Native		Se-Met	
Data collection	Mtd-P1*		Mtd-P1 (L52M/V53M)*	
Space group	$P6_3$		$P2_13$	
Cell dimensions				
a, *b*, *c* (Å)	92.26, 92.26, 102.23		112.31, 112.31, 112.31	
α, β, γ (°)	90.0, 90.0, 120.0		90.0, 90.0, 90.0	
		Peak	*Inflection*	*Remote*
Wavelength (Å)	1.1055	.9791	.97932	.95372
Resolution (Å)**	50.0-1.56 (1.62-1.56)	50.0-1.50 (1.55-1.50)	50.0-1.49 (1.54-1.49)	50.0-1.55 (1.61-1.55)
R_{merge}	6.1 (12.4)	9.8 (54.7)	9.8 (55.6)	10.9 (82.1)
$I / \sigma I$	23.5 (14.2)	12.9 (2.3)	13.7 (2.2)	19.1 (2.4)
Completeness (%)	98.6 (100)	99.5 (100)	97.6 (80.7)	99.6 (100)
Redundancy	5.9 (5.3)	4.7 (4.0)	4.7 (4.0)	9.5 (8.9)
Refinement				
Resolution (Å)	20.33-1.56 (1.59-1.56)			
No. reflections	69,212			
R_{work} / R_{free}	12.8/14.5			
No. atoms				
Protein	2,758			
Ligand/ion	2			
Water	488			
B-factors				
Protein	9.2			
Ligand/ion	7.9			
Water	23.1			
R.m.s deviations				
Bond lengths (Å)	0.009			
Bond angles (°)	1.23			

*Single crystal used for data set.
**Highest resolution shell is shown in parenthesis.

Table 2.2. Data collection and refinement statistics for Mtd-P3, Mtd-I1, Mtd-M1, and Mtd-U1 molecular replacement structures.

Data collection	Mtd-P3c*	Mtd-I1*	Mtd-M1*	Mtd-U1*
Space group	$P6_322$	$P6_3$	$P6_3$	I222
Cell dimensions				
a, *b*, *c* (Å)	66.55, 66.55, 334.95	92.31, 92.31, 102.29	93.53, 93.53, 103.15	86.68, 129.44, 312.65
α, β, γ (°)	90.0, 90.0, 120.0	90.0, 90.0, 120.0	90.0, 90.0, 120.0	90.0, 90.0, 90.0
Resolution (Å)**	50.0-2.07 (2.14-2.07)	50.0-2.52 (2.61-2.52)	50.0-1.86 (1.92-1.85)	50.0-1.87 (1.94-1.87)
R_{merge}	13.9 (73.3)	16.0 (53.5)	4.8 (32.1)	13.7 (81.3)
$I / \sigma I$	14.1 (3.1)	10.2 (2.7)	37.6 (5.1)	12.5 (2.0)
Completeness (%)	95.14 (94.2)	99.9 (99.5)	98.8 (85.3)	99.7 (99.0)
Redundancy	10.0 (9.5)	5.5 (5.0)	5.7 (5.2)	5.7 (5.2)
Refinement				
Resolution (Å)	28.99-2.07 (2.12-2.07)	26.01-2.52 (2.59-2.52)	42.6-1.86 (1.91-1.86)	22.4-1.87 (1.92-1.87)
No. reflections	26,802	16,724	42,618	144,743
R_{work} / R_{free}	18.2/22.1	17.0/21.3	16.3/17.8	21.0/30.7
No. atoms				
Protein	2,753	2,759	2,770	8,238
Ligand/ion	5	1	1	5
Water	369	164	352	1,167
B-factors				
Protein	22.7	23.3	22.8	24.2
Ligand/ion	62.1	38.7	23.4	18.2
Water	30.5	22.6	30.8	31.0
R.m.s. deviations				
Bond lengths (Å)	0.007	0.008	0.005	0.006
Bond angles (°)	1.03	1.01	0.986	0.989

*Single crystal used for data set.
**Highest resolution shell is shown in parenthesis.

Isothermal Titration Calorimetry (ITC)

ITC measurements were performed at 15 and 23 °C (MicroCal MCS, Amherst, MA) and data were analyzed with Origin software. Mtd-P1 (450 μM trimer) was injected from a stirring syringe into a calorimeter cell containing Prn-E (45 μM); both proteins were in 150 mM NaCl, 50 mM Tris, pH 8.0. Estimation of the lower bound for K_D of binding between Mtd-M1 and Prn-E derives from the following argument. Mtd-M1 (400 μM) was used in ITC experiments with Prn-E (40 μM), following the same protocol as for Mtd-P1, but no heat of binding was observable. This means that value of the parameter *c* ($c = M_{tot}K_A$) (Wiseman et al., 1989) for Mtd-M1 is likely to be < 1, since values of $1 < c < 500$ yield interpretable binding isotherms. M_{tot} is the concentration of the component in the cell (Prn-E), and K_A is the association constant. Therefore, with the parameter *c* likely to be < 1 for Mtd-M1, the binding affinity can be estimated to be $1/K_A \geq K_D \geq 40$ μM. For comparison, the value of *c* for Mtd-P1 (450 μM) binding to Prn-E (45 μM) is ~14. In addition, interpretable and consistent binding data for Mtd-P1 were also obtained in experiments carried out with *c* ~3.

Static Light Scattering (SLS)

SLS was performed on protein samples (in 150 mM NaCl, 50 mM Tris, pH 8.0) applied at 0.5 mL/min to a size exclusion column (Superdex 200) using a BioCad Sprint HPLC. Data were acquired at scattering angles between 52 and 163° at 25°C with a Wyatt Technologies DAWN DSP laser photometer. Molecular masses were determined using Astra software.

Putative variable proteins and fold assignment

The BUILD_PROFILE and PP_SCAN modules (N.E. and A.S., unpublished results) as implemented in MODELLER (Sali and Blundell, 1993) were used to assign possible folds for the twelve potentially variable protein sequences listed below:

Table 2.3. Identified species and domains used for protein fold assignment

			ORF	Genbank Accession Number	Residues used for fold assignment
1	VHML	*Vibrio harveyi* bacteriophage VHML	ORF35	AAN12335	1-410
2	*B.l.*	*Bifidobacterium longum*	Blon03000687	ZP_00120812	123-526
3	*B.t.*	*Bacteroides thetaiotaomicron* VPI-5482	BT2316	NP_811229	494-888
4	*T. d.*	*Treponema denticola* ATCC 35405	TDE2269 (DUF-323)*	NP_972869	1-329
5	*T.e.* 1A	*Trichodesmium erythraeum* IMS101	(DUF-323)*	ZP_00326406	628-877
6	*T.e.* 1B	*Trichodesmium erythraeum* IMS101	(DUF-323)*	ZP_00326410	41-305
7	*T.e.* 2	*Trichodesmium erythraeum* IMS101	(DUF-323)*	ZP_00324769	1-187
8	*N.* PCC 1	*Nostoc sp.* PCC 7120	(DUF-323)*	NP_489056	32-273
9	*N.* PCC 2A	*Nostoc sp.* PCC 7120	(DUF-323)*	NP_487534	1-250
10	*N.* PCC 2B	*Nostoc sp.* PCC 7120	(DUF-323)*	NP_487535	29-279
11	*N.p.* 1	*Nostoc punctiforme* PCC 73102	(DUF-323)*	ZP_00111384	397-648
12	*N.p.* 2	*Nostoc punctiforme* PCC 73102	(DUF-323)*	ZP_00111383	28-279

*DUF-323 (domain of unknown function) indicates protein identified by BLAST as belonging to this family.

The fold assignment procedure consisted of the following two steps. First, for each sequence, a profile (or multiple alignment) containing all detectable homologous sequences from a non-redundant sequence database was constructed using BUILD_PROFILE. Second, each profile was compared against a library of profiles

corresponding to sequences of structural domains from the SCOP database (version 1.65) (Murzin et al., 1995; Sellers, 1974). Statistically significant hits were collected and their folds assigned to the query sequence.

The BUILD_PROFILE module of MODELLER relies on local dynamic programming (Needleman and Wunsch, 1970) to align a sequence against a profile and uses a robust procedure to determine the statistical significance of the alignments (Pearson, 1998). Profiles were constructed by iteratively scanning the sequence or the profile from a previous iteration against a database of non-redundant sequences (see below) until convergence was reached or a maximum of five iterations were carried out. A sequence was selected and included in the profile if the e-value for the alignment was < 1.

The non-redundant sequence database used to construct sequence profiles was derived from the UniProt database (Apweiler et al., 2004). The UniProt database was parsed by MODELLER to remove redundancy and to filter low complexity sequences. The database was parsed such that sequences were removed if they were shorter than 30 residues or longer than 3,000 residues, had more than 40 residues or 40% of their length consisting of low complexity residues, and were redundant (90% sequence identity with other members). The filtering is essential to maintain the integrity and speed of the BUILD_PROFILE and PP_SCAN searches. Low complexity sequences tend to yield large amount of false positives.

Each profile was then compared against a library of profiles from SCOP domain structures using the PP_SCAN method in MODELLER. The PP_SCAN method compares profiles and aligns sequences using the correlation coefficient between two positions in each of the profiles (Marti-Renom et al., 2004). This method allowed us to determine if any of the sequences is similar to any of the 9,501 nonredundant SCOP domain sequences available in our library of profiles.

BUILD_PROFILE results for Mtd (three domains):

Mtd N-terminal domain (residues 5-52):

The profile contains only one sequence. The profile was added to the SCOP library of profiles to be compared against sequences of potentially variable proteins.

Mtd Intermediate domain (residues 53-179):

BUILD_PROFILE did not find any sequences similar to the intermediate domain of Mtd with an e-value < 1. This profile (or single sequence) was added to the SCOP library of profiles to be compared against sequences of potentially variable proteins.

Mtd C-terminal domain (residues 180-379):

BUILD_PROFILE found 246 sequences similar to the C-terminal domain of Mtd with an e-value < 1. This profile was added to the SCOP library of profiles to be compared against sequences of potentially variable proteins.

BUILD_PROFILE and PP_SCAN results for putative variable proteins:

1. *VHML (1-410).* BUILD_PROFILE found 260 sequences similar to this sequence with an e-value < 1.

Table 2.4. Significant hits for VHML (1-410) with e-value < .1

PDB code	SCOP code	SCOP family	Sequence identity	E-value	Length alignment
1flg:A	b.70.1.1	Quinoprotein alcohol dehydrogenase-like	13.7%	0.00	293
1g72:A	b.70.1.1	Quinoprotein alcohol dehydrogenase-like	13.9%	0.00	306
1h4i:A	b.70.1.1	Quinoprotein alcohol dehydrogenase-like	12.0%	0.00	307
1kb0:A_2	b.70.1.1	Quinoprotein alcohol dehydrogenase-like	12.1%	0.00	351
1kv9:A_2	b.70.1.1	Quinoprotein alcohol dehydrogenase-like	11.2%	0.00	306
Mtd C-terminal domain	**d.169.1.1**	**C-type lectin**	**17.8%**	**0.18E-07**	**146**

2. *B.l. (123-526).* BUILD_PROFILE did not find any sequences similar to his sequence with an e-value < 1. The PP_SCAN algorithm against SCOP sequences yielded no significant matches.

Upon re-examination of the *B.l.* (123-526) sequence, a low complexity region was found (residues 318-333). Low complexity regions have amino acid distributions significantly different from those observed in globular proteins. The search was carried out again with the low complexity region omitted and two fragments (123-317 and 334-526).

Results for residues 123-317: BUILD_PROFILE found 908 sequences similar to this sequence with an e-value < 1.

Table 2.5. Significant hits for *B.l.* (123-317) with e-value < .1

PDB code	SCOP code	SCOP family	Sequence identity	E-value	Length alignment
1k0b:A	d.144.1.7	Protein kinase	7.8%	0.9e-1	104

PP_SCAN could only find hits with low significance, which indicates that this shorter sequence aligns with very low sequence identity to any of the SCOP folds in our library.

Results for residues 334-526: BUILD_PROFILE did not find any sequence similar to this sequence with an e-value < 1.

PP_SCAN found significant hits (e-value < 0.1):

Table 2.6. Significant hits for *B.l.* (334-526) with e-value < .1

PDB code	SCOP code	SCOP family	Sequence identity	E-value	Length alignment
1n11:A	d.211.1.1	Ankyrin repeat	20.6%	0.2e-2	63

PP_SCAN could only find hits with low significance, which indicates that this shorter sequence aligns with very low sequence identity to any of the SCOP folds in our library.

3. *B.t. (494-888).* BUILD_PROFILE found 12,012 sequences similar to this sequence with an e-value < 1. The large number of hits is consistent with low complexity regions that generate large numbers of false positives.

Table 2.7. Significant hits for *B.t.* (494-888) with e-value < .1

PDB code	SCOP code	SCOP family	Sequence identity	E-value	Length alignment
1g9m:G	d.171.1.1	gp120 core	12.5%	0.00	271
1g9n:G	d.171.1.1	gp120 core	11.5%	0.00	277
1gc0:A	c.67.1.3	Cystathionine synthase-like	10.6%	0.72E-01	143

Upon re-examination of this sequence, two low complexity regions were discovered at the ends of the sequence (residues 785-795 and 865-876).

Results for residues 494-784: BUILD_PROFILE found 439 sequences similar to this shorter sequence with an e-value < 1.

Table 2.8. Significant hits for *B.t.* (494-784) with e-value < .1

PDB code	SCOP code	SCOP family	Sequence identity	E-value	Length alignment
1acc:_	f.11.1.1	Anthrax protective protein	10.1%	0.3e-5	168
1epw:A3	d.92.1.7	Clostridium neurotoxin	8.9%	0.1e-4	114
1bhe:_	b.80.1.3	Galacturonase	7.8%	0.3e-3	131

PP_SCAN could only find hits with low significance, which indicates that this shorter sequence aligns with very low sequence identity to any of the SCOP folds in our library.

Results for 796-864: BUILD_PROFILE found 66 sequences similar to this shorter sequence with an e-value < 1.

Table 2.9. Significant hits for *B.t.* (796-864) with e-value < .1

PDB code	SCOP code	SCOP family	Sequence identity	E-value	Length alignment
1nvr:A	d.144.1.7	Protein Kinase like	11.8%	0.01	51
1ln6:A	f.14.1.2	G coupled receptor	10.6%	0.06	24

PP_SCAN could only find hits with low significance, which indicates that this sequence aligns with very low sequence identity to any of the SCOP folds in our library.

4. *T.d. (1-329).* BUILD_PROFILE found 169 sequences similar to this sequence with an e-value < 1.

Table 2.10. Significant hits for *T.d.* (1-329) with e-value < .1

PDB code	SCOP code	SCOP family	Sequence identity	E-value	Length alignment
Mtd C-terminal domain	**d.169.1.1**	**C-type lectin**	**17.0%**	**0.00**	**144**
1o6v:A_1	b.1.18.15	Internalin Ig-like domain	19.7%	0.73E-03	63
1h6u:A_1	b.1.18.15	Internalin Ig-like domain	16.7%	0.14E-02	54
1koa:__2	d.144.1.7	Protein kinases, catalytic subunit	14.5%	0.28E-01	116

5. *T.e.* 1A (628-877). BUILD_PROFILE found 238 sequences similar to this sequence with an e-value < 1.

Table 2.11. Significant hits for *T.e.* (628-877) with e-value < .1

PDB code	SCOP code	SCOP family	Sequence identity	E-value	Length alignment
Mtd C-terminal domain	**d.169.1.1**	**C-type lectin**	**14.7%**	**0.00**	**184**
1h4g:A*	b.29.1.11	Xylanase/endoglucanase	15.6%	0.4e-07	141

*Ten other xylanases/endoglucanases had hits with e-value < 0.1

6. *T.e.* 1B (41-305). BUILD_PROFILE found 188 sequences similar to this sequence with an e-value < 1.

Table 2.12. Significant hits for *T.e.* (41-305) with e-value < .1

PDB code	SCOP code	SCOP family	Sequence identity	E-value	Length alignment
Mtd C-terminal domain	**d.169.1.1**	**C-type lectin**	**14.4%**	**0.00**	**193**

7. *T.e.* 2 (1-187). BUILD_PROFILE found 222 sequences similar to this sequence with an e-value < 1.

Table 2.13. Significant hits for *T.e.* 2 (1-187) with e-value < .1

PDB code	SCOP code	SCOP family	Sequence identity	E-value	Length alignment
Mtd C-terminal domain	**d.169.1.1**	**C-type lectin**	**14.4%**	**0.00**	**186**
1yna:_*	b.29.1.11	Xylanase/endoglucanase	15.6%	0.1e-8	133

8. *N.* PCC 1 (32-273). BUILD_PROFILE found 135 sequences similar to this sequence with an e-value smaller < 1.

Table 2.14. Significant hits for N.PCC 1 (32-273) with e-value < .1

PDB code	SCOP code	SCOP family	Sequence identity	E-value	Length alignment
Mtd C-terminal domain	d.169.1.1	C-type lectin	14.4%	0.00	186

9. *N.* PCC 2A (1-250). BUILD_PROFILE found 167 sequences similar to this sequence with an e-value < 1.

Table 2.15. Significant hits for N.PCC 2A (1-250) with e-value < .1

PDB code	SCOP code	SCOP family	Sequence identity	E-value	Length alignment
Mtd C-terminal domain	d.169.1.1	C-type lectin	15.9%	0.00	176

10. *N.* PCC 2B (29-279). BUILD_PROFILE found 166 sequences similar to this sequence with an e-value < 1.

Table 2.16. Significant hits for N.PCC 2B (29-279) with e-value < .1

PDB code	SCOP code	SCOP family	Sequence identity	E-value	Length alignment
Mtd C-terminal domain	d.169.1.1	C-type lectin	16.6%	0.00	186

11. *N.p.* 1 (397-648). BUILD_PROFILE found 146 sequences similar to this sequence with an e-value <1.

Table 2.17. Significant hits for *N.p.* 1 (397-648) with e-value < .1

PDB code	SCOP code	SCOP family	Sequence identity	E-value	Length alignment
Mtd C-terminal domain	d.169.1.1	C-type lectin	17.1%	0.00	179

12. *N. p.* 2 (28-279). BUILD_PROFILE found 145 sequences similar to this sequence with an e-value < 1.

Table 2.18. Significant hits for *N.p.* 2 (28-279) with e-value < .1

PDB code	SCOP code	SCOP family	Sequence identity	E-value	Length alignment
Mtd C-terminal domain	d.169.1.1	C-type lectin	18.6%	0.00	180

Results

Overall Structure of Mtd

Crystal structures of four Mtd variants, P1, M1, P3c, and I1 (Fig. 2.1, Table 2.1), that differ in receptor specificity and host tropism were determined (resolution limits 1.56–2.52 Å). Mtd-P1 confers phage infectivity against pathogenic (Bvg^+) but not environmental (Bvg^-) phase *Bordetella*, because it uses the Bvg^+-specific surface protein and virulence factor pertactin as a receptor (Doulatov, 2004; Emsley et al., 1996). In contrast, Mtd-M1 confers infectivity against Bvg^- phase *Bordetella* only, indicating that it binds an unknown minus-specific receptor, possibly glycosyl groups of the O-antigen as suggested by indirect evidence. Mtd-P3c confers infectivity against the plus phase only but uses an unknown receptor other than pertactin, and Mtd-I1 confers infectivity against both plus and minus phases (i.e. it is indiscriminate), indicating it uses an unknown receptor expressed by both phases. The structure of a fifth variant, Mtd-U1, was also determined. Mtd-U1 is a direct descendant of Mtd-M1 and is not known to be infective, but is of interest because its variable region resembles the template region in sequence.

Tropic variants of Mtd are nearly identical in overall structure, indicating no large conformational changes result from sequence variation. Mtd is an intertwined, pyramid-shaped trimer (Fig. 2.2), corresponding in size and shape to knobs seen at the ends of phage tail fibers (Liu, 2004). A largely hydrophobic interface buries >4,500 $Å^2$ of surface area in each protomer, consistent with obligatory trimerization observed

by static light scattering (Fig. 2.3). Hydrogen bonds and a shared cation are also involved in trimerization. The majority (69%) of the interface area is composed of non-polar residues.

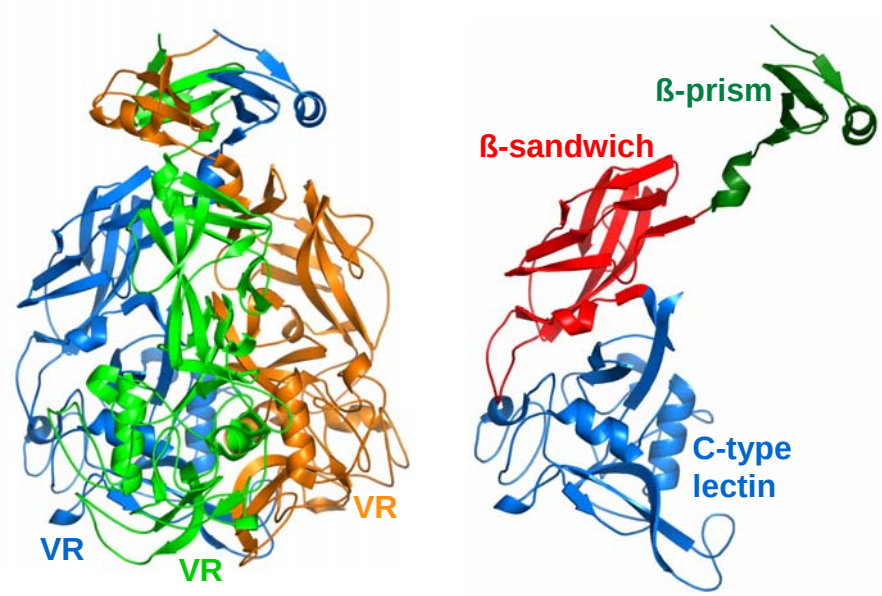


Figure 2.2. Mtd structure. Left, Structure of Mtd trimer (~90 Å height, ~50 Å base) in ribbon representation with protomers colored blue, green, and gold, and locations of VR indicated. Mtd-P1 is depicted, but other Mtd variants are identical (average rmsd 0.33 Å, 376 Cα atoms). Right, Mtd protomer in ribbon representation with β-prism domain in green, β-sandwich in red, and C-type lectin in blue. Molecular graphics were made with PyMol (DeLano, 2002).

Each protomer is also joined to its neighbor via 20 hydrogen bonds, one electrostatic interaction (between Glu-234 and Arg-354), and at least one shared divalent cation. In each of the Mtd structures this cation (Mg^{2+} or Ca^{2+}) is coordinated in tetragonal bipyramidal geometry with side chain oxygen atoms of Glu-312 and the main chain carbonyl oxygen atoms of Phe-313. In the Mtd-M1 trimer, a second Mg^{2+} ion occupies an adjacent binding site, forming similar coordination geometry with the

other side chain oxygen atoms of Glu-312 and several ordered water molecules. The cation-binding sites in Mtd are not similar to the calcium-binding sites observed in carbohydrate-binding C-type lectins.

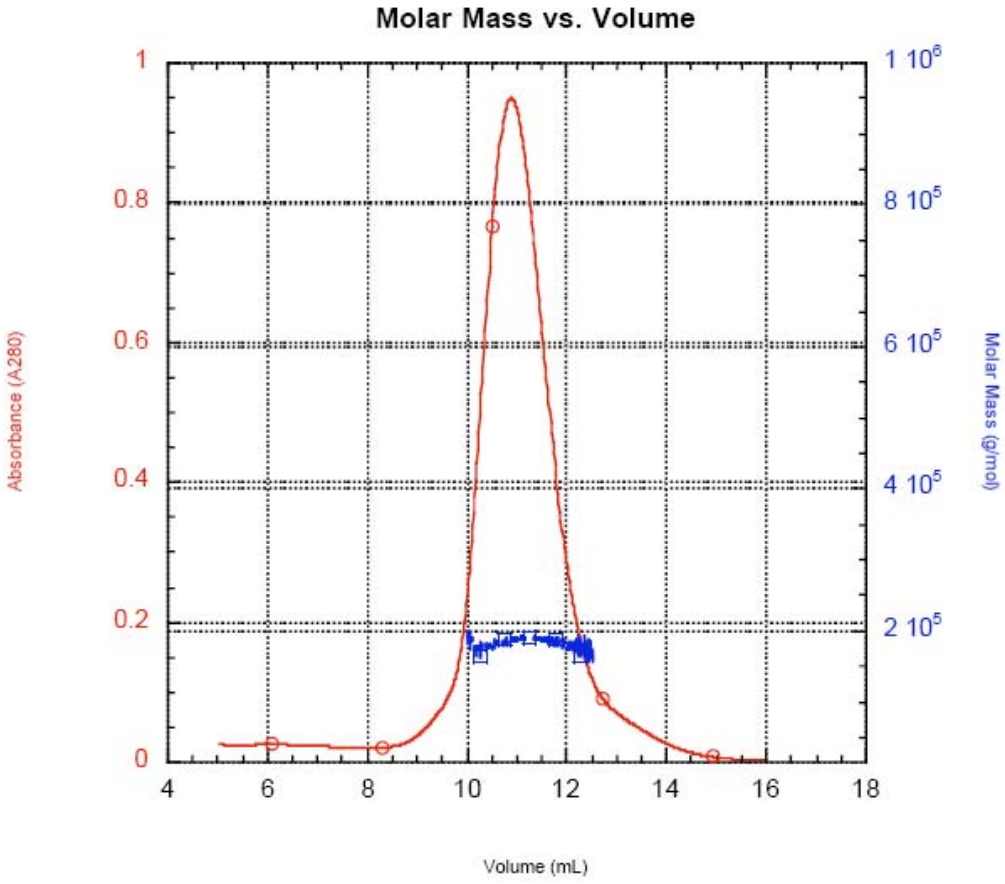


Figure 2.3. Mtd-P1 and Prn-E associate in a 3:1 complex. Static light scattering of the Mtd-P1/Prn-E complex applied to a size exclusion column. Protein absorbance at 280 nm (red) and calculated molar mass (blue) as a function of retention time are shown. The calculated molecular mass of the peak is 181.8 ± 3.272 kDa, and the predicted molecular mass for a 3:1 Mtd-P1/Prn-E complex is 179.8 kDa. No dissociation was observed for the Mtd-P1/Prn-E complex eluting from the column, and therefore an assumption of 100% mass recovery was made in calculating molecular mass.

Mtd protomers are composed of N-terminal, intermediate, and C-terminal domains (Fig. 2.4). At the apex of the pyramid, trimerization of the N-terminal domain (residues 1-48) of Mtd forms a threefold symmetric β-prism, resembling the pseudo-threefold symmetric β-prisms of monocot lectins (rmsd 2.4 Å, 60 Cα atoms) but lacking residues identified in these proteins to bind carbohydrates (Hester et al., 1995; Sauerborn et al., 1999). The intermediate domain forms a β-sandwich containing three- and four-stranded antiparallel sheets with a near right-angle turn in the middle, and has a novel fold. Functional roles for the β-prism and β-sandwich domains are not known, but tethering Mtd to the phage surface seems a likely possibility. The overall intertwining nature of the protein, which is important for function of the C-terminal variable domain, is best appreciated by noticing that the β-prism domain occupies a different face of the pyramid than the other domains (Fig. 2.2).

C-type lectin fold in the variable domain

The C-terminal domain is unexpectedly found to have a C-type lectin (CLec)-fold (Weis et al., 1991), as identified by structural homology searches (Holm and Sander, 1993) (Fig. 2.5). The CLec-fold of Mtd is related to those of divergently and convergently evolved CLec proteins, such as macrophage mannose receptor (2.7 Å rmsd, 101 Cα, Z=6.5) (Feinberg, 2000) and intimin (3.0 Å rmsd, 89 Cα, Z=5.8) (Batchelor, 2000; Luo, 2000), respectively. Significantly, the CLec-fold is not functionally restricted to calcium-dependent carbohydrate binding but is instead

recognized to constitute a general ligand-binding (including protein) motif (Drickamer, 1999).

The typical topological features of the ~110–130 residue CLec-fold, as also seen in Mtd, are a two-stranded antiparallel β-sheet formed by the domain's N- and C-termini (β1β5) which are connected by two α-helices and a three-stranded, antiparallel β-sheet (β2β3β4) (Fig. 2.5). The β2β3β4 sheet in Mtd contains an additional three-residue strand, β4', and a short 3_{10}-helix. The CLec-fold in Mtd has convergently evolved, having different core residues than divergently evolved CLec proteins and lacking residues required for calcium- or carbohydrate-binding (Weis et al., 1991). Likewise, none of the four disulfide-bond forming cysteines found in many CLec domains is found in Mtd, confirming that disulfides are not required for stability of the CLec fold. Unique to Mtd are two ~40-residue inserts that interrupt secondary structure elements and stabilize the variable region (see below).

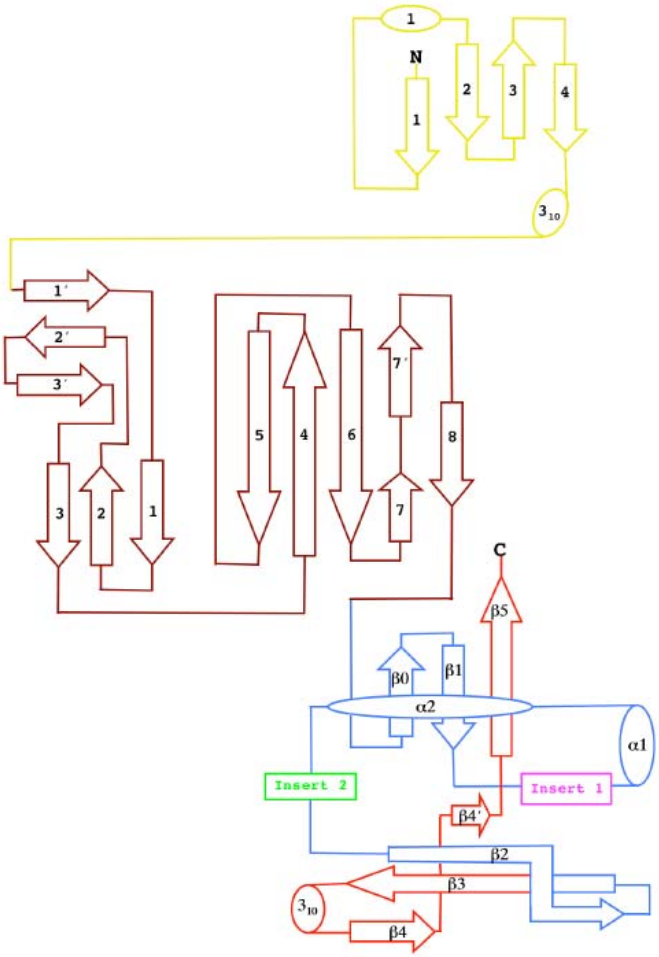


Figure 2.4. . **Topology diagram of Mtd.** N-terminal (β-prism) domain in yellow and intermediate (β-sandwich) domains in brown, and coloring for C-type lectin domain in blue and red. Ovals represent α-helices and 3_{10}-helices, and arrows β-strands; boxes denote the two unique insert regions in the C-type lectin fold of Mtd.

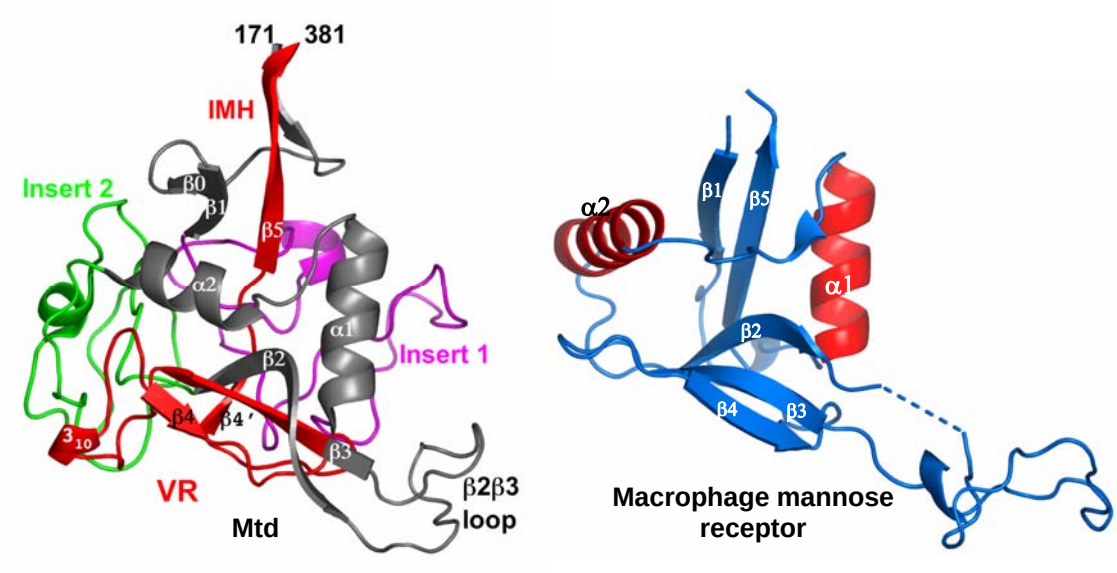


Figure 2.5. C-type Lectin Homology Left, C-type lectin domain of Mtd in ribbon representation with VR in red, insert 1 (residues 200–236, between β1 and α1) in pink, insert 2 (residues 264–305, between α2 and β2) in green, and remaining portions in gray. Secondary structure elements are labeled, including short 3_{10}-helix in VR. The last strand (β5) in the C-type lectin fold corresponds to the IMH and is positioned in the core of the trimer. Right, the C-type lectin domain of mannose macrophage receptor (residues 639–763) in ribbon representation with α-helices in red, and β-strands and loops in blue.

Variable Region

All 12 adenine-encoded, variable residues of Mtd are organized into a solvent-exposed, receptor-binding site on the external face of the β2β3β4β4' sheet (Fig. 2.6A), indicating an elegant coevolution between the genetic mechanism of variation and the physical target of this variation. Notably, this same face has been shown to be responsible for protein-protein interactions in the CLec proteins Ly49A (Tormo et al., 1999) and intimin (Batchelor, 2000; Luo, 2000).

All but two of the variable residues in the template region are encoded by AAC (Fig. 2.1). Adenine-specific mutagenesis of AAC-encoded Asn permits substitution

by 14 other amino acids covering the gamut of chemical character. For example, Trp cannot be encoded, but Phe and Tyr can, and likewise Glu and Lys cannot be encoded, but Asp and Arg (also His) can. Significantly, the use of the AAC codon rules out a nonsense codon being introduced. Residues 348 and 369 are encoded by ACG (Thr) and ATC (Ile), respectively, in the template region, and adenine-specific mutagenesis of these permits substitution by three other amino acids (Ser, Pro, Ala for 348; Val, Leu, Phe for 369). There appears to be no structural necessity for residue 348 to be small, but 369 may need to be hydrophobic in order to pack between the invariant residues Trp307 and Trp309.

The binding site also contains two invariant, solvent-exposed aromatic residues, Trp307 and Trp345, which are likely to partake in receptor interactions (Fig. 2.6A). Two additional aromatic residues, Tyr322 and Tyr333, are contributed by the β2β3 loop from a neighboring protomer that intertwines into the binding site (Fig. 2.6A). The β2β3 loop varies greatly in structure among CLec proteins, and is responsible for calcium-dependent carbohydrate binding among a subset (Weis et al., 1992). Variable residues along with the above invariant aromatic residues together constitute ~900 $Å^2$ of surface area per protomer of Mtd-P1.

Concordance between the genetic mechanism and physical target of variation is also seen in the β5 strand, the very C-terminus of Mtd. The β5 strand is encoded by a 21-bp genetic element (initiation of mutagenic homing, IMH) that sets the directionality of information transfer and is not replaced by the template region (Doulatov, 2004). It is therefore invariant despite having adenine-encoded amino acids, and is also noteworthy in having codons that differ between variable and

template region at bases other than adenine (i.e., encoding residues 376–379). Invariance of the IMH at the nucleotide level is echoed at the protein level, with β5 positioned in the central core of trimer, making close intra- and inter-molecular contacts that could be disrupted by variation.

Mtd Variants

Structural comparison of tropic variants of Mtd reveals that the main chain conformation of the CLec-domain is remarkably invariant to large differences in sequence (Fig. 2.6B, rmsd 0.27 Å, 45 Cα). This is made more striking by the fact that more than half the variable residues are located on loops (344, 346, 347, 359, 360, 364, and 366) (Fig. 2.6A). Providing stabilization to these loops are the two inserts in the CLec-domain along with trimeric assembly. The inserts form hydrogen bonds to the main and invariant side chains (e.g., Ser 351, 353, and 365) of the variable region (Fig. 2.6C). Trimeric assembly results in a similar pattern of hydrogen bonds, for example between the β2β3 loop of one protomer and the invariant side chain of Arg354 of another protomer (Fig. 2.6D). The β2β3 loop has the same intertwining conformation in all Mtd variants examined, being positioned exclusively over invariant residues (i.e., 351–356) (Figs. 2.6A and 2.6D). The β-prism and β-sandwich domains reinforce overall trimeric assembly, and therefore may have indirect roles in stabilizing the backbone of the variable region.

The binding sites of the five Mtd variants studied differ greatly in their pattern of hydrophobicity. Figure 2.7 shows that Mtd-P1 and Mtd-I1 have highly hydrophobic

binding sites, and that the continuity of the hydrophobic surface decreases successively for Mtd-P3c, -M1, and -U1, with this last one having nine template region-encoded, mostly hydrophilic residues (Fig. 2.1). The binding sites of Mtd-P1 and -I1 accommodate four to five large, exposed hydrophobic residues, and although a preponderance of exposed hydrophobic surface is typically correlated with protein instability (Takano et al., 1998), both Mtd-P1 and -I1 are found to be highly stable proteins. Protein stability is likely aided by the hydrophilic surface formed by invariant residues surrounding the binding site.

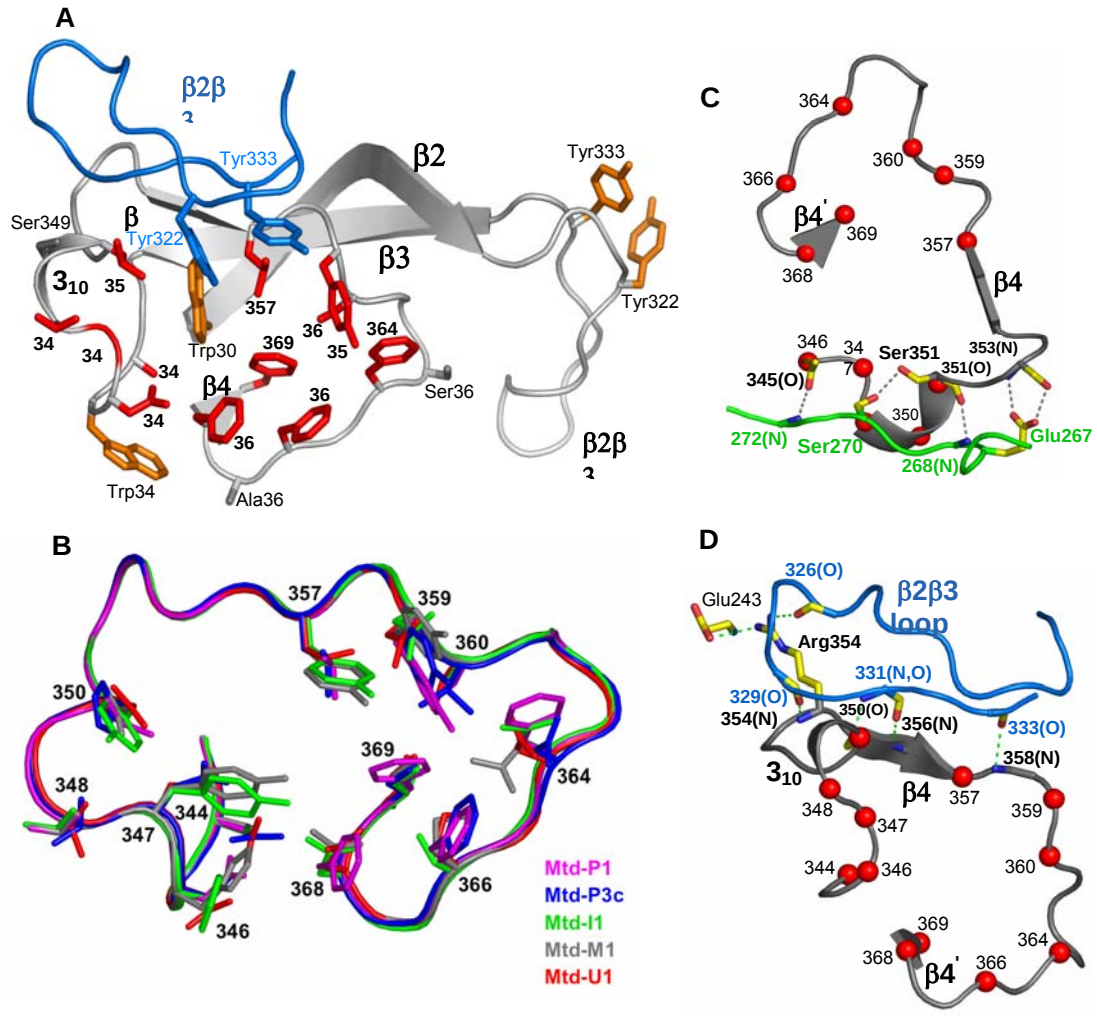


Figure 2.6A-D. Mtd Variable Region **(A)** Receptor-binding site of Mtd (gray, residues 307–369) with variable residues (red) and invariant residues (aromatic in orange; others in gray) in bonds representation. β2β3 loop from a neighboring protomer and its invariant tyrosines 322 and 333 colored blue. **(B)** Superposition of variable region of Mtd variants. Variable residues are in bonds representation and numbered, and backbones are in trace representation. **(C)** Stabilization of VR by insert 2. Main chain and side chain (Glu267 and Ser270) of insert 2 (green) form hydrogen bonds to main chain and side chain (Ser351 and Ser353) of VR (gray). Positions of variable residues marked by a red sphere (at Cα), and hydrogen bonds represented as dashed lines. Main chain atoms indicated in parenthesis. **(D)** Stabilization of VR by trimeric assembly. Main chain of β2β3 loop (blue) and side chain of Glu243 from a neighboring protomer form hydrogen bonds with main chain and side chain (Arg354) atoms of VR. Representation is same as in preceding panel.

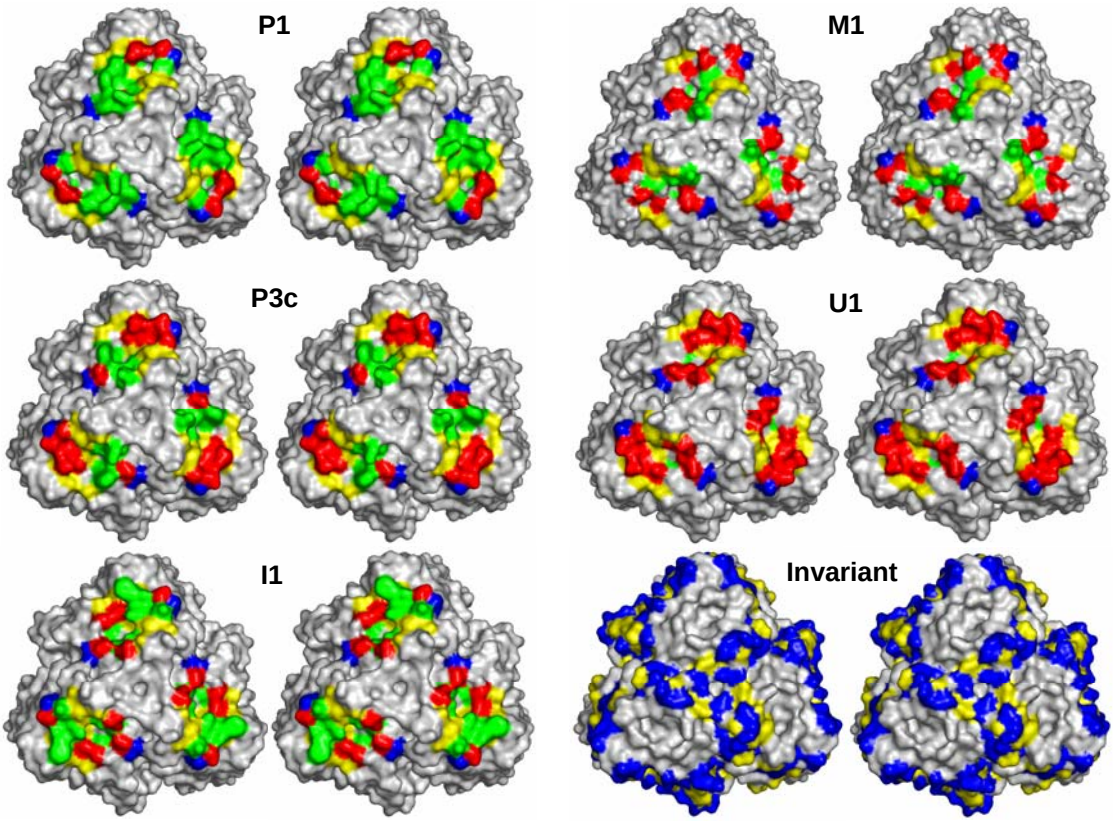


Figure 2.7. Receptor-binding site. Molecular surface in stereo of receptor-binding sites of Mtd variants. Variable (green) and invariant (yellow) hydrophobic residues (Ala, Val, Leu, Ile, Phe, Tyr, Trp, and Met), and variable (red) and invariant (blue) hydrophilic residues (Ser, Thr, Asn, Gln, Asp, Glu, His, Lys, Arg, and Cys) are colored. 'Invariant' shows surrounding surface, with receptor-binding site residues uncolored and all others colored.

Pertactin Interactions

Direct association between Mtd-P1 and its *Bordetella* receptor pertactin (Doulatov, 2004) was examined using purified components. The ectodomain of pertactin (Prn-E) was incubated with Mtd variants and found by a coprecipitation assay to associate with Mtd-P1 but not with Mtd-I1 or -U1 (Fig. 2.8). Intriguingly, pertactin was also

found to associate with Mtd-P3c and weakly with Mtd-M1, both direct descendants of Mtd-P1 (Liu, 2002). Tyr359 is the only variable residue common to these three pertactin-binding Mtd variants (Fig. 2.1), and is also highly conserved among plus tropic phage (Liu, 2002). While an explanation for the importance of Tyr359 requires further investigation, our results suggest that this residue is a dominant pertactin-binding determinant. Pertactin cross-reactivity, as conferred by the single residue Tyr359, along with the observation that the variable region is usually replaced in short patches rather than *en bloc* (Doulatov, 2004) provide an explanation for why phage tropism is vastly skewed towards the plus phase. This skew is evident in the fact that plus-tropic phage convert to minus-tropic or indiscriminate (infecting both plus and minus *Bordetella)* phage at a frequency of 10^{-6}, in contrast to the thousand-fold greater frequency of 10^{-3} at which minus-tropic and indiscriminate phage convert to plus-tropic phage (Liu, 2002). It is possible that this skew affects *Bordetella* variation between its pathogenic (plus) and environmental (minus) phases.

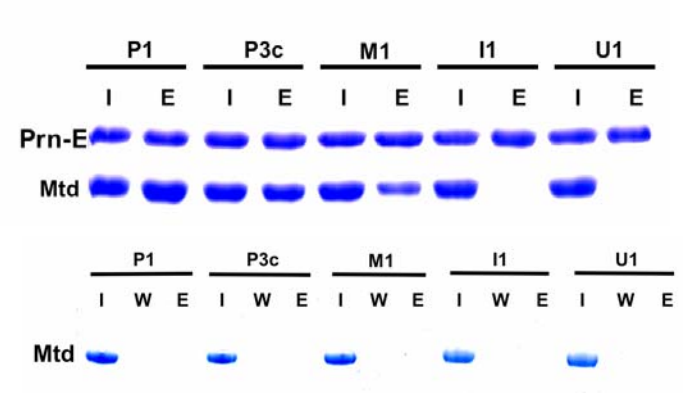
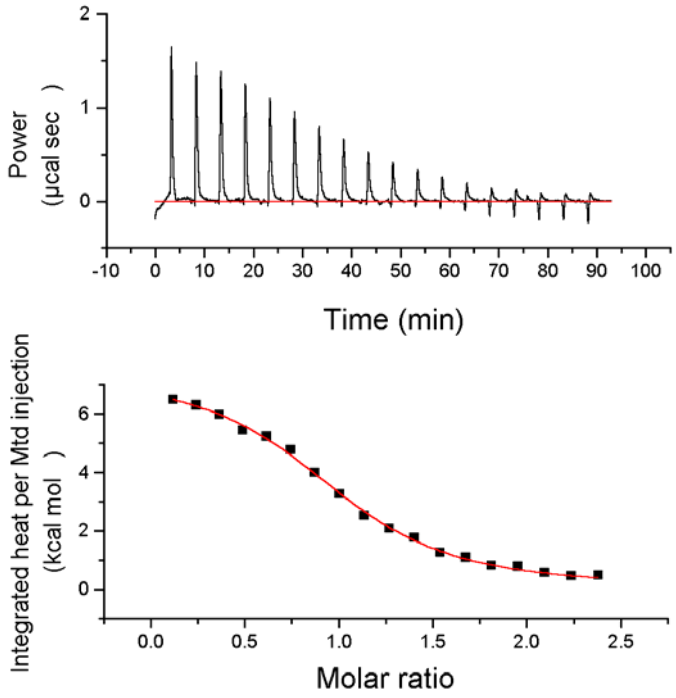


Figure 2.8. Pertactin Associations. Top panel shows association of Mtd variants with the his-tagged ectodomain of pertactin (Prn-E) assessed by coprecipitation using Ni^{2+}-NTA agarose beads and visualized using SDS-PAGE and Coomassie staining. The first lane for each variant shows relative amounts of protein incubated ('I'), and the second lane shows relative amounts of protein eluted ('E') from beads after washing. Lower panel shows coprecipitation of Mtd variants with Ni^{2+}-NTA agarose beads in the absence of his-tagged ectodomain of pertactin (Prn-E), visualized using SDS-PAGE and Coomassie staining. The first lane for each variant shows protein incubated with beads ('I'), the second lane shows the last wash ('W') out of 3, and the last lane shows protein eluted ('E') from beads after washing.

Potentially trivalent Mtd-P1 is found by static light scattering to associate with only a single pertactin molecule (Fig. 2.3). This suggests that pertactin sterically occludes other binding sites, associates with Mtd pseudosymmetrically, or both. The binding sites in Mtd are fixed in relation to one another by trimeric intertwining and not flexibly disposed to adapt to target surfaces as in antibodies. The entropically-driven association between Mtd-P1 and pertactin has a modest K_D of ~3 μM, as assessed by isothermal titration calorimetry (ITC), which also provides further evidence for 3:1 Mtd:pertactin stoichiometry (Fig. 2.9). Since *Bordetella* bacteriophage appears to have 12 Mtd trimers in total (Liu, 2002; Liu, 2004), this modest affinity is likely to translate to high avidity in binding outer membrane-

confined pertactin. The affinity of Mtd-M1 for pertactin is too low to be quantified precisely, but a K_D of $\geq$ 40 μM can be estimated (see Methods), distinguishing between infectious (3 μM) and noninfectious ($\geq$ 40 μM) interactions.



Temperature (°C)	K_d (μM)	ΔH (kcal mol^{-1})	N	ΔG (kcal mol^{-1})	TΔS (kcal mol^{-1})
23	3.04 ± .48	7.248 ± .082	1.00	–7.47	14.2
15.5	3.42 ± .48	7.785 ± .122	.981	–7.21	14.7

Figure 2.9. Mtd-P1 binding kinetics. Isothermal titration calorimetry of binding between Mtd-P1 and Prn-E. Top panel shows the endothermic reaction due to successive 15 μL injections of Mtd-P1 into a solution of Prn-E. Middle panel shows integrated heats of injection plotted against molar ratio of Mtd-P1 trimer to Prn-E. Bottom panel summarizes data from experiments carried out at 15.5 and 23 °C, with free energy and entropy changes calculated using the relationships $\Delta G = -RT \ln K_A$ and $\Delta G = \Delta H - T\Delta S$. N is the number of Mtd trimers calculated to bind Prn-E.

Variable proteins of other retroelements

Nine retroelements having similar reverse transcriptases to the one in the *Bordetella* bacteriophage retroelement and similar adenine-specific differences between template and variable regions have been identified in phage and bacterial genomes (Doulatov, 2004). Potentially variable proteins of these related retroelements have low sequence identity to Mtd (~17%), but structural and mechanistic considerations enable sequence alignment consisting of the β2β3β4β4' sheet of the CLec-fold (Fig. 2.10). Most particularly, the invariant Mtd binding site residue Trp345 is present in a highly conserved 'GGXW' motif. Invariant residues (Ser351, Ser353, Arg354) involved in loop stabilization, trimeric contacts, or both (Figs. 2.6C and 2.6D) are also generally conserved. As in Mtd, residues differing between variable and template regions or ones that could potentially vary through an adenine-specific mechanism are located chiefly between the β3 and β5 strands.

These conclusions are supported by profile-profile based sequence alignments, which indicate statistically significant matches (e-values $0–10^{-8}$) between the C-terminus of Mtd and C-termini of potentially variable proteins from *Treponema denticola*, *Vibrio harveyi* ML phage, *Trichodesmium*, and *Nostoc* (Appendix). An additional piece of evidence comes from the observation that variable regions and IMH elements of related retroelements are consistently located at the very C-terminus of potentially variable proteins (Doulatov, 2004). This property is not necessitated by genetic mechanisms of variation, but is consistent with features of the CLec-fold of

Mtd (Fig. 2.5). Taken together, these results provide evidence that the CLec-fold is used broadly as a scaffold for sequence variation by related retroelements.

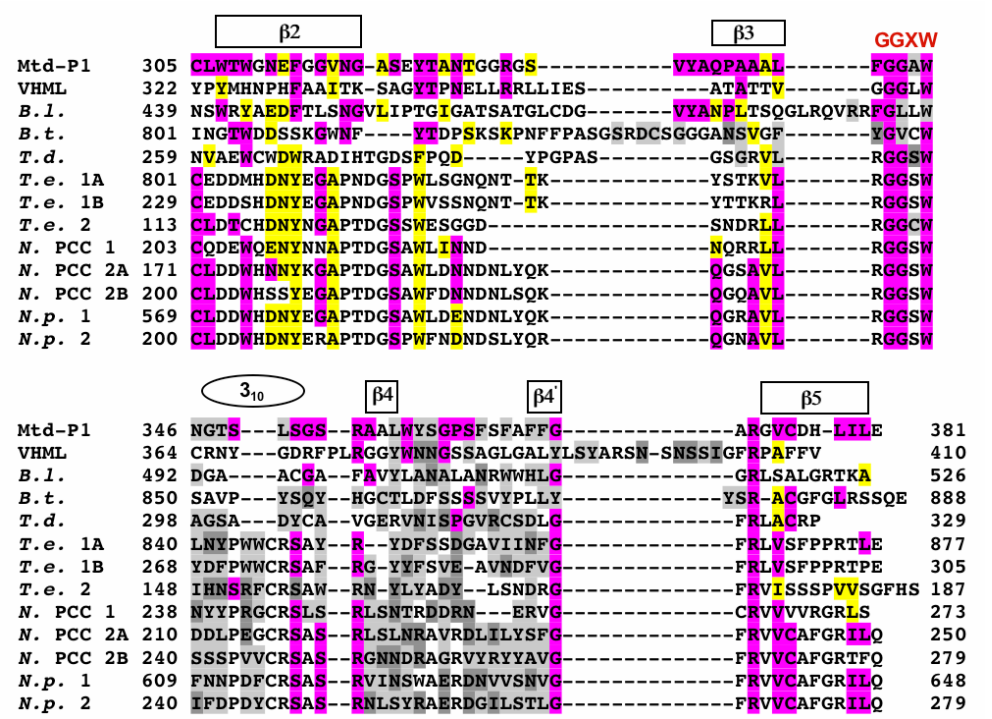


Figure 2.10. Structure-based sequence alignment of Mtd with potentially variable proteins of related retroelements. Residues colored pink in Mtd are identical and yellow are chemically conserved (grouped as: Trp, Phe, Tyr; Ala, Val, Leu, Ile; Ser, Thr; Asn, Gln, Asp, Glu; Arg, Lys) to residues in potentially variable proteins. Light gray corresponds to variable residues in Mtd, and residues that differ between VR and TR in genomic sequences of potentially variable proteins. Dark gray corresponds to residues that could vary by an adenine-specific mechanism in potentially variable proteins. In assigning color, grays take precedence over pink and yellow, such that certain putatively variable residues are also identical or conserved. Secondary structure elements (box for β-strand, and oval for 3_{10} helix) for Mtd are denoted above the alignment, and the 'GGXW' motif is also denoted. VHML, Vibrio harveyi bacteriophage VHML; B.l., Bifidobacterium longum; B.t., Bacteroides thetaiotaomicron; T. d., Treponema denticola; T.e., Trichodesmium erythraeum IMS101; N. PCC, Nostoc sp. PCC 7120; N.p., Nostoc punctiforme PCC 73102. A number of these sequences are identified as belonging to the DUF-323 family, suggesting that DUF-323 proteins are likely to have C-type lectin folds.

In contrast to the CLec domain, neither the β-prism nor β-sandwich domain of Mtd is conserved in proteins of related retroelements. Indeed some of these putatively variable proteins are quite short (e.g., *T.e.* 2) and appear to be composed of only the CLec domain, whereas others are quite large (e.g., *B.t.* and *T.e.* 1A*)* and have additional, uncharacterized domains. The biological functions of these proteins are currently unknown, but those from *B. thetaiotaomicron* and *T. denticola* are predicted to be lipoproteins that localize to the outer bacterial surface, and the one from *B. longum* to have a signal sequence that is consistent with secretion. These three bacterial species are part of the normal human microbiota, and variability in their CLec-fold may play a role in modulating interactions with host tissues.

Discussion

Highly variable proteins may be conceptually divided into predators and prey. Predator proteins are those, such as antibodies, that function by binding unanticipated ligands. Massive diversity in predator proteins enhances the likelihood, for example, of an antibody binding to a novel antigen with sufficient affinity to stimulate an immune response, or Mtd to a novel bacterial cell surface receptor to initiate an infection. For prey proteins, variability is responsible for evasion from predatory binding proteins, resulting in the phenomenon known as antigenic variation. Interestingly, pertactin itself is antigenically variant (Mooi, 1998), requiring Mtd to keep pace with pertactin variation as driven by antibody selection.

While predators are extremely rare, a fairly sizeable number of antigenically variable prey proteins are known, with well characterized examples being trypanosomal variable surface glycoproteins (VSG) (Blum, 1993) and gonococcal pilins (Parge, 1995). The selection pressure for diversity in antigenically variable proteins is not nearly as great as in predator proteins. For example, diversity in VSG (Van der Ploeg, 1982) is estimated to be ~10^3, considerably lower than the ~10^{14}–10^{16} of immune system proteins (Davis and Bjorkman, 1988) and ~10^{13} of Mtd (Liu, 2002). In addition, structural demands on prey proteins are much less severe than those on predator proteins. Antigenic variation need only lessen the affinity with which an antibody binds, which may be achieved through small changes, while the binding site

of a predatory protein needs to accommodate almost any sequence in order to have sufficient diversity to bind almost any ligand. This latter demand has been met successfully by the Ig-fold in the vertebrate immune system. It may have also been met by the leucine-rich repeat fold in the immune system of jawless fish, although in this case the extent of diversity is not known and antigen binding has not yet been demonstrated (Pancer, 2004). Our work provides the first evidence that evolution has made use not only of the Ig-fold but also the CLec-fold as a paradigm for massive sequence variation in predatory binding proteins.

Are there consequences for using one fold over the other? In contrast to the CLec-fold of Mtd, which displays a fixed number of variable residues on an invariant backbone, the number of variable residues displayed on loops by the immunoglobulin (Ig)-fold of antibodies and T-cell receptors is not fixed and neither is the backbone conformation. This is made possible by variable residues in the Ig-fold being continuous in primary sequence rather than dispersed as in the CLec-fold. It is reasonable to expect that the Ig-fold provides greater binding diversity because of its combined sequence and conformational variation. However, there appears to be a cost for greater diversity, as certain conformations of the Ig-fold are deleterious to protein stability and promote fibril formation, as in the case of light chain amyloidosis (Bellotti et al., 2000; Helms and Wetzel, 1995; Pokkuluri, 2002; Stevens, 2000). It seems likely that the CLec-fold, although more limited in diversity, may be less susceptible to unstable conformations and protein misfolding errors because of its

static scaffold, and that evolution has arrived at different balances between diversity and stability in the CLec- and Ig-folds.

The text of this chapter, in full, is a reprint of the material as it appears in the journal Nature Structural and Molecular Biology. The dissertation author was the primary researcher and/or author and the co-authors (S. McMahon, J. Lawton, D. Kerkow, A. Hodes, M. Marti-Renom, S. Doulatov, E. Narayanan, A. Sali, J.F. Miller, and P. Ghosh) listed in this publication contributed to or supervised the research which forms the basis for this chapter.

Bibliography

Akerley, B. J., Cotter, P. A., and Miller, J. F. (1995). Ectopic expression of the flagellar regulon alters development of the Bordetella-host interaction. Cell *80*, 611-620.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M.*, et al.* (2004). UniProt: the Universal Protein knowledgebase. Nucleic Acids Res *32*, D115-119.

Batchelor, M. (2000). Structural basis for recognition of the translocated intimin receptor (Tir) by intimin from enteropathogenic Escherichia coli. EMBO J *19*, 2452-2464.

Bellotti, V., Mangione, P., and Merlini, G. (2000). Review: immunoglobulin light chain amyloidosis[ndash]the archetype of structural and pathogenic variability. J Struct Biol *130*, 280-289.

Blum, M. L. (1993). A structural motif in the variant surface glycoproteins of Trypanosoma brucei. Nature *362*, 603-609.

Budisa, N., Steipe, B., Demange, P., Eckerskorn, C., Kellermann, J., and Huber, R. (1995). High-level biosynthetic substitution of methionine in proteins by its analogs 2-aminohexanoic acid, selenomethionine, telluromethionine and ethionine in Escherichia coli. Eur J Biochem *230*, 788-796.

Chothia, C., and Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. J Mol Biol *196*, 901-917.

Cotter, P. A., and Miller, J. F. (1997). A mutation in the Bordetella bronchiseptica bvgS gene results in reduced virulence and increased resistance to starvation, and identifies a new class of Bvg-regulated antigens. Mol Microbiol *24*, 671-685.

Davis, M. M., and Bjorkman, P. J. (1988). T-cell antigen receptor genes and T-cell recognition. Nature *334*, 395-402.

DeLano, W. L. (2002). The PyMOL molecular graphics system.

Doulatov, S. (2004). Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. Nature *431*, 476-481.

Drickamer, K. (1999). C-type lectin-like domains. Curr Opin Struct Biol *9*, 585-590.

Emsley, P., Charles, I. G., Fairweather, N. F., and Isaacs, N. W. (1996). Structure of Bordetella pertussis virulence factor P.69 pertactin. Nature *381*, 90-92.

Emsley, P., McDermott, G., Charles, I. G., Fairweather, N. F., and Isaacs, N. W. (1994). Crystallographic characterization of pertactin, a membrane-associated protein from Bordetella pertussis. J Mol Biol *235*, 772-773.

Feinberg, H. (2000). Structure of a C-type carbohydrate recognition domain from the macrophage mannose receptor. J Biol Chem *275*, 21539-21548.

Helms, L. R., and Wetzel, R. (1995). Destabilizing loop swaps in the CDRs of an immunoglobulin VL domain. Protein Sci *4*, 2073-2081.

Hester, G., Kaku, H., Goldstein, I. J., and Wright, C. S. (1995). Structure of mannose-specific snowdrop (Galanthus nivalis) lectin is representative of a new plant lectin family. Nat Struct Biol *2*, 472-479.

Holm, L., and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. J Mol Biol *233*, 123-138.

Jones, T. A., Zou, J.-Y., Cowan, S. W., and Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr A *47*, 110-119.

Liu, M. (2002). Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage. Science *295*, 2091-2094.

Liu, M. (2004). Genomic and genetic analysis of Bordetella bacteriophages encoding reverse transcriptase-mediated tropism-switching cassettes. J Bacteriol *186*, 1503-1517.

Luo, Y. (2000). Crystal structure of enteropathogenic Escherichia coli intimin-receptor complex. Nature *405*, 1073-1077.

Marti-Renom, M. A., Madhusudhan, M. S., and Sali, A. (2004). Alignment of protein sequences by their profiles. Protein Sci *13*, 1071-1087.

Mooi, F. R. (1998). Polymorphism in the Bordetella pertussis virulence factors P.69/pertactin and pertussis toxin in The Netherlands: temporal trends and evidence for vaccine-driven evolution. Infect Immun *66*, 670-675.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol *247*, 536-540.

Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol *48*, 443-453.

Otwinowski, Z., and Minor, W. (1997). Processing of x-ray diffraction data collected in oscillation mode. Methods Enzymol *276*, 307-326.

Pancer, Z. (2004). Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. Nature *430*, 174-180.

Parge, H. E. (1995). Structure of the fibre-forming protein pilin at 2.6 A resolution. Nature *378*, 32-38.

Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. J Mol Biol *276*, 71-84.

Perrakis, A. (1997). wARP: improvement and extension of crystallographic phases by weighted averaging of multiple-refined dummy atomic models. Acta Crystallogr D Biol Crystallogr *53*, 448-455.

Pokkuluri, P. R. (2002). Increasing protein stability by polar surface residues: domain-wide consequences of interactions within a loop. Biophys J *82*, 391-398.

Sali, A., and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol *234*, 779-815.

Sauerborn, M. K., Wright, L. M., Reynolds, C. D., Grossmann, J. G., and Rizkallah, P. J. (1999). Insights into carbohydrate recognition by Narcissus pseudonarcissus lectin: the crystal structure at 2 A resolution in complex with alpha1[ndash]3 mannobiose. J Mol Biol *290*, 185-199.

Sellers, P. (1974). Theory and Computation of Evolutionary Distances. Siam Journal on Applied Mathematics *26*, 787-793.

Stevens, F. J. (2000). Four structural risk factors identify most fibril-forming kappa light chains. Amyloid *7*, 200-211.

Takano, K., Yamagata, Y., and Yutani, K. (1998). A general rule for the relationship between hydrophobic effect and conformational stability of a protein: stability and structure of a series of hydrophobic mutants of human lysozyme. J Mol Biol *280*, 749-761.

Terwilliger, T. C. (1997). Multiwavelength anomalous diffraction phasing of macromolecular structures: analysis of MAD data as single isomorphous replacement with anomalous scattering data using the MADMRG Program. Methods Enzymol *276*, 530-537.

Tormo, J., Natarajan, K., Margulies, D. H., and Mariuzza, R. A. (1999). Crystal structure of a lectin-like natural killer cell receptor bound to its MHC class I ligand. Nature *402*, 623-631.

Uhl, M. A., and Miller, J. F. (1996). Integration of multiple domains in a two-component sensor protein: the Bordetella pertussis BvgAS phosphorelay. EMBO J *15*, 1028-1036.

Van der Ploeg, L. H. (1982). An analysis of cosmid clones of nuclear DNA from Trypanosoma brucei shows that the genes for variant surface glycoproteins are clustered in the genome. Nucleic Acids Res *10*, 5905-5923.

Weis, W. I., Drickamer, K., and Hendrickson, W. A. (1992). Structure of a C-type mannose-binding protein complexed with an oligosaccharide. Nature *360*, 127-134.

Weis, W. I., Kahn, R., Fourme, R., Drickamer, K., and Hendrickson, W. A. (1991). Structure of the calcium-dependent lectin domain from a rat mannose-binding protein determined by MAD phasing. Science *254*, 1608-1615.

Winn, M. D. (2003). An overview of the CCP4 project in protein crystallography: an example of a collaborative project. J Synchrotron Radiat *10*, 23-25.

Wiseman, T., Williston, S., Brandts, J. F., and Lin, L. N. (1989). Rapid measurement of binding constants and heats of binding using a new titration calorimeter. Anal Biochem *179*, 131-137.

III.

Receptor recognition by Mtd, the massively sequence variable protein of a diversity-generating retroelement

Abstract

Recognition of diverse targets through massive protein sequence variation was thought to occur exclusively in the adaptive immune system until discovery of diversity-generating retroelements (DGR). Here, we present the first structure of a DGR encoded variable protein bound to a receptor. Mtd from the *Bordetella* bacteriophage DGR binds the receptor pertactin through antibody-like surface complementarity but in a pseudosymmetric fashion that sets it apart from antibodies and demonstrates remarkable adaptability in its static binding sites. We tracked evolution of receptor specificity in Mtd to understand how modest differences in receptor affinities are translated into all-or-none phage infectivity. We find a rationale in the avidity effect, which derives from multiple phage-bound Mtd molecules interacting cooperatively with surface-localized pertactin. Significantly, avidity not only strengthens Mtd-pertactin interactions for infectivity but also amplifies affinity differences, thereby providing specificity. We suggest these principles are generally applicable to the immune system and other sequence variable repertoires.

Introduction

Diversity-generating retroelements (DGR) have been identified in the genomes of ~25 prokaryotic organisms and bacteriophages to date (Doulatov, 2004) (M. Gingery and J.F.M, unpublished results), and are likely to be responsible for massive protein sequence variation in many of these. The most extensively characterized DGR is encoded by *Bordetella* bacteriophage (Doulatov, 2004; Liu, 2002; Liu, 2004). DGR-based variation of *Bordetella* bacteriophage Mtd (40 kDa), which functions as the phage's receptor-binding protein, enables the use of alternative receptors during infection (Liu, 2002). This is necessary as the expression pattern of *Bordetella* surface molecules (i.e., potential phage receptors) changes according to environmental cues (Liu, 2002; Uhl and Miller, 1996). The expression pattern is regulated by the BvgAS two-component system, with virulence factors being expressed only in the pathogenic or Bvg^+ phase of *Bordetella* (Akerley et al., 1995; Liu, 2002; Uhl and Miller, 1996). In contrast, genes required for motility are expressed only in the environmental or Bvg^- phase, while others are expressed preferentially at intermediate levels of Bvg activation (Cotter and Miller, 1997). Pertactin, a Bvg^+-specific protein that localizes to the outer membrane and belongs to the autotransporter family (Emsley et al., 1996; Mattoo and Cherry, 2005), is the only Mtd receptor whose identity is currently known, but the existence of others has been demonstrated genetically (Liu, 2002; McMahon et al., 2005).

Due to the unique template-based and adenine-specific mode of DGR mutagenesis (Liu, 2002), variation in Mtd is focused to 12 adenine-encoded residues scattered across a 45-residue C-terminal region (Fig. 3.1). Mtd has previously been shown to accommodate massive sequence variation (~10^{13} possible sequences) through its convergently evolved C-type lectin (CLec)-fold, which organizes these 12 variable residues into a remarkably static binding site. The static nature of these sites is stabilized by trimeric assembly and two unique insert regions (McMahon et al., 2005), and may function to decrease the incidence of protein misfolding upon variation. Evidence suggests that proteins of other DGRs accommodate variation through the CLec-fold as well (Doulatov, 2004; McMahon et al., 2005). The scale of Mtd variation is paralleled only by the adaptive immune system of vertebrates, with the immunoglobulin (Ig)-fold of antibodies and T cell receptors (TCRs) accommodating ~10^{14}-10^{16} sequences (Davis and Bjorkman, 1988), and recently recognized leucine-rich repeat (LRR)-fold of variable lymphocyte receptors in agnathous fish accommodating ~10^{14} sequences (Alder et al., 2005).

We sought to understand how DGR-encoded variable proteins recognize their targets and focused on evolution of Mtd variants through three generations, beginning with Mtd-P1 (Fig. 3.1). Mtd-P1 is expressed by BPP-1 phage and has pertactin as its receptor (Liu, 2002; McMahon et al., 2005). We had previously noted that Mtd-M1, a direct descendant of Mtd-P1, while not being able to use pertactin as a receptor nevertheless has affinity for pertactin (Liu, 2002; McMahon et al., 2005). This affinity is somewhat weaker than that of Mtd-P1 and because it does not lead to infection, is

viewed as non-specific. To understand how a modest difference in affinity is translated into all-or-none phage infectivity, we compared how Mtd-P1, Mtd-M1, and a closely related descendant of Mtd-M1 (i.e., Mtd-P6) interact with pertactin (Fig. 3.1). Study of the evolution of these three Mtd variants leads us to conclude that avidity is critical to Mtd function. Avidity not only provides Mtd with interaction strength but also specificity, and we surmise that these principles apply generally to the immune system and other sequence variable repertoires.

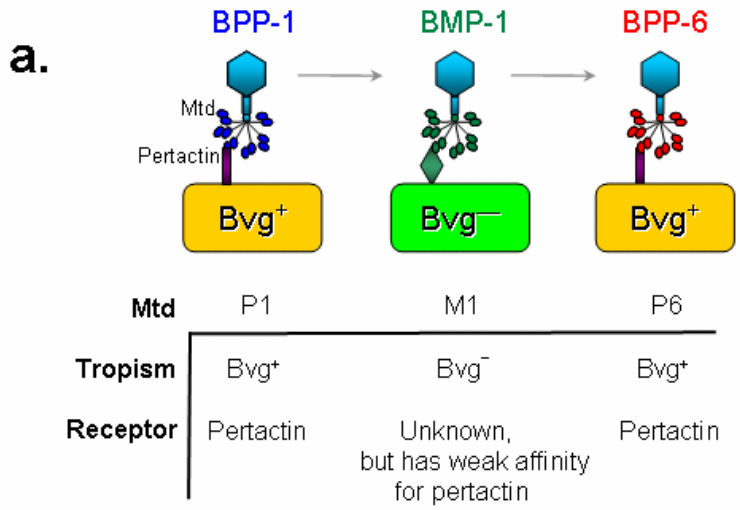


Figure 3.1. Tropism and receptor specificities of Bordetella bacteriophages. (a) The Mtd P1 variant is expressed by BPP 1 phage. The Mtd M1 variant is a direct descendant of Mtd M1 and is expressed by BMP 1 phage. The Mtd P6 variant is a direct descendant of Mtd M1 and is expressed by BPP-6 phage. (b) Variable residues of Mtd-P1 are shown in blue. For Mtd M1 variable residues in common with Mtd P1 are shown in blue and those that are unique to Mtd M1 in green. For Mtd P6, residues in common with Mtd M1 or Mtd P1 are shown in green or blue, respectively, and those that are unique to Mtd P6 in red. Mtd has 381 residues in total.

Experimental Procedures

Phage Evolution, DNA Cloning, and Protein Purification

BMP-1(Δ*brt*) lysogens were induced, as previously described (Liu, 2002), from *B. bronchiseptica* RB50 expressing Brt ectopically from the *B. bronchiseptica* FHA promoter on the pBBR1MCS-derived, medium-copy plasmid pMin1 (M. Xu and J.F. Miller, manuscript in preparation). Resulting phage were propagated on *B. bronchiseptica* RB54 (Bvg^-), and then selected for plaque formation on *B. bronchiseptica* RB53 (Bvg^+) (Cotter and Miller, 1997). These BPP phage were screened for dependency on pertactin for infection, as assayed by their inability to form plaques on *B. bronchiseptica* RB50Δ*prn* (Bvg^+). Pertactin-dependent phage were then lysogenized and the VRs of single colony lysogens were sequenced before reinduction of phage.

Mtd variants were cloned, expressed, and purified as described previously, as was His-tagged Prn-E (residues 38-640) (McMahon et al., 2005). Loop deletions of Prn-E (Δ190-199 and Δ399-407) were constructed with an N-terminal His-tag (MGSSHHHHHHSSGLVPRGSHMAS) by strand-overlap extension PCR (Higuchi et al., 1988) and expressed using pET28b (Novagen). These mutant proteins were refolded from inclusion bodies and purified as previously described (McMahon et al., 2005). Size-exclusion chromatography (Superdex 200) was used to confirm the non-aggregated and monomeric state of these mutant proteins. *In vivo* biotinylation of Prn-E was carried out through expression of pertactin residues 38-640 carrying an

N-terminal biotinylation sequence (MSGLNDIFEAQKIEWHEGAPELE) from pAN-5 (AviTag) in *E. coli* AVB101. Biotinylated Prn-E was expressed as inclusion bodies, and refolded and purified as above.

Crystallization and Structure Determination

Mtd-P1/Prn-E complexes were formed by mixing ~1.6-fold molar excess of purified Mtd-P1 trimers with His-tagged Prn-E, and separating complexes from unbound proteins using successive Ni^{2+}-chelation and size exclusion (Superdex 200) chromatographies. Protein crystals were grown by the sitting drop, vapor diffusion method at 25^{o} C by mixing equal volumes of 10 mg/mL Mtd-P1/Prn-E ($\varepsilon_{280calc}$ 254050 M^{-1} cm^{-1}) and precipitant (2 M NaCl, 100 mM Tris, pH 8.5, 200 mM Li_2SO_4).

X-ray diffraction data were collected at 100 K using synchrotron radiation at the Advanced Light Source (λ = 1.000 Å; oscillation range 0.35^{o}; beamline 5.0.1) from crystals that had been cryoprotected by soaking in precipitant solution containing 25% glycerol. Crystals of Mtd-P1/Prn-E belong to space group $P6_1$ and contain two Mtd-P1/Prn-E complexes in the asymmetric unit. Diffraction data were processed and scaled using HKL2000 (Otwinowski and Minor, 1997), and the structure was determined by the molecular replacement technique using Amore (Winn, 2003) and Mtd-P1 (PDB Id: 1YU0) as the search model. The model was built and refined using O (Jones et al., 1991) and Refmac 5.2 (Winn, 2003), respectively. Non-crystallographic symmetry (NCS) restraints were used in initial stages of refinement but removed in later stages, whereupon TLS refinement was used with each

polypeptide chain set to an independent TLS group and all atomic B-factors set to the Wilson B-factor value of 81.5 $Å^2$. This was followed by restrained positional and B-factor refinement (without prior phase information restraints).

Table 3.1. Data collection and refinement statistics for Mtd-P1/Prn-E (Molecular Replacement). * Data collected from single crystal. **Highest resolution shell is shown in parenthesis.

	Mtd-P1/Prn-E*
Data collection	
Space group	$P6_1$
Cell dimensions	
a, *b*, *c* (Å)	413.5, 413.5, 98.9
α, β, γ (°)	90.0, 90.0, 120.0
Resolution (Å)	50.00-3.156 (3.26-3.156)**
R_{merge}	15.6 (78.7)
$I / \sigma I$	13.8 (2.0)
Completeness (%)	100.0 (100.0)
Redundancy	9.2 (8.3)
Refinement	
Resolution (Å)	49.63-3.156 (3.238-3.156)
No. reflections	157,861
R_{work} / R_{free}	22.4/24.9
No. atoms	
Protein	23,896
Ligand/ion	4
Water	0
B-factors	
Protein	69.5
Ligand/ion	52.3
Water	0
R.m.s. deviations	
Bond lengths (Å)	0.008
Bond angles (°)	1.141

Electron density for loops 265-291 (the RGD loop) and 349-354 of Prn-E were not visible and these residues were not modeled. Except for these loops, no breaks in the main chain were evident for Mtd-P1 or Prn-E. No residues are in the disallowed region of the Ramachandran plot, with 81.7% (of the 2170 residues other than glycine or proline) being in the most favored region and 17% in the additional allowed region.

Molecular graphics were made using PyMol (http://pymol.sourceforge.net).

Coprecipitation assay

For assays with Prn-E loop deletion mutants, 50 μL of His-tagged Prn-E, PrnEΔ190-199) and Prn-E(Δ399-407) (8 μM) were incubated (5 min, 25 ^{o}C) with 30 μL Ni^{2+}-NTA agarose beads (Sigma). For preparation of beads, 50 μL of Ni^{2+}-NTA agarose bead slurry (50% suspension in 30% ethanol) were centrifuged, and 20 μL of the overlying solution were removed. The Prn-E constructs were then added to the beads, and unbound Prn-E was removed by centrifugation and removal of overlying solution. Mtd (60 μM trimer) was added to the protein-bead mixture, incubated for 5 min at 25 °C, and washed with binding buffer (150 mM NaCl and 50 mM Tris, pH 8.0) three times. Bound protein was eluted with binding buffer supplemented with 500 mM imidazole, then visualized by SDS-PAGE. The final wash was devoid of protein. For quantification of binding in the coprecipitation assay, the intensity of Mtd bound to His-tagged Prn-E, PrnEΔ190-199), or Prn-E(Δ399-407) was measured using Kodak 1D imaging software (ver.3.5.0), and the intensity of Mtd

bound in the absence of His-tagged Prn-E was taken as background and subtracted from this value.

Comparison of Mtd-P6 and Mtd-M1 affinity for Prn-E was carried out as above, except that varying concentrations of Mtd were incubated with 80 μM Prn-E (5 min, 25 °C) prior to addition of beads Ni^{2+}-NTA agarose beads. Beads (30 μL) were then added and samples handled as above. Intensities of protein bands were quantified from Coomassie stained SDS-PAGE gels using Kodak 1D imaging software (ver.3.5.0). A standard concentration curve of known Mtd quantities was used to ensure linearity of measurements. The apparent K_d of Mtd-M1 for Prn-E was estimated by identifying total concentrations of Mtd-P6 and Mtd-M1 that yield equivalent amounts of bound Mtd. The K_d for Mtd-P6, as determined by SPR, was then used in the Langmuir isotherm to calculate the concentration of bound and free Mtd-P6. The concentration of bound Mtd-P6 was applied to Mtd-M1 in the Langmuir isotherm to determine its apparent K_d. Two equivalent concentrations of Mtd-P6 and Mtd-M1 (3 and 25 μM of Mtd-P6 and Mtd-M1, respectively, and 8 and 63 μM of Mtd-P6 and Mtd-M1) were used to estimate an apparent K_d of ~500 mM for Mtd-M1.

Surface Plasmon Resonance (SPR)

SPR measurements were performed using a Biacore 3000 at 25 ^{o}C. Approximately 100 response units (RU) of biotinylated Prn-E were immobilized on the surface of a streptavidin sensor chip in running buffer (150 mM NaCl, 50 mM Tris, pH 8.0, 0.005% Tween 20); a streptavidin sensor chip without any immobilized protein was used as a blank. Mtd-P1, -M1, and -P6 proteins and BPP-1, BMP-1, BIP-1, and BPP-6 phage (in running buffer) were injected at 30 µL/min for 2 min with collection data rate set to high. A two minute wash delay was used for complete dissociation of Mtd from Prn-E and accurate fitting of dissociation data. The data were analyzed using BIA evaluation 3.0 software for non-linear curve fitting using a single site model.

For SPR assays using phage, cultures of induced phage were treated with chloroform and centrifuged twice (8000 x *g*, 10 min) to remove cell debris. Phage were then concentrated by centrifugation (13,000 x *g*, 3 hr) in Corex tubes, and pellets (which contain phage) were resuspended in SM buffer (50 mM Tris, pH 7.5, 0.1 M NaCl, 8 mM $MgSO_4$, and 0.01% gelatin). Phage were then further concentrated by ultrafiltration (Amicon, YM-30) into running buffer. The number of plaque forming units per unit volume was determined from this final phage stock (on the appropriate *B. bronchiseptica* background), and this value was converted to molarity using Avogadro's number.

BIAsimulation was used to estimate a K_d limit for Mtd-M1. The estimation is based on the observations that Mtd-P6 produces a binding event with an R_{MAX} of 66.6 RU, and the baseline noise level of detection is 1 RU. If it is assumed that either the on-rate or off-rate of Mtd-M1 is the same as that for Mtd-P6, then the fact that 2 μM Mtd-M1 yields no detectable binding signal is consistent with an apparent K_d of ≥ 450 μM.

BIAsimulation was also used to estimate an apparent K_d limit for BMP-1. The off-rate of BMP-1 was set to that of BPP-6 (1.78×10^{-3} sec^{-1}), and the R_{MAX} of BMP-1 was set to that of BPP-6 (2.96×10^{6} RU). The fact that a concentration of BMP-1 of 8.50×10^{-12} M produces a binding event of 10 RU is consistent with an apparent K_d of ~500 nM.

Accession codes

Protein Data Bank: Coordinates have been deposited with accession code 2IOU.

Results

Recognition of pertactin by Mtd-P1

We isolated, crystallized, and determined the structure of Mtd-P1 bound to the ectodomain of *B. bronchiseptica* pertactin (Prn-E, residues 38-640) (McMahon et al., 2005). Structure determination to 3.16 Å resolution reveals that a single Mtd-P1 trimer (120 kDa in total), although having three receptor-binding sites, associates with just one pertactin ectodomain (60 kDa), agreeing with the 3:1 stoichiometry determined in solution (McMahon et al., 2005) (Fig. 3.2). Two of these ~180 kDa complexes occupy the asymmetric unit of the crystal (Table 3.1), and except for certain unbound loops of pertactin, molecular details are unambiguous and nearly identical in the two independent complexes in the crystal.

A total of ~2300 $Å^2$ is buried at the Mtd-pertactin interface, with 66% being contributed by hydrophobic residues, in agreement with prior indication of an entropically driven association (McMahon et al., 2005). This compares favorably with the 1680 $Å^2$ of an average antibody-antigen interface (Lo Conte et al., 1999). About three-fourths of the extensive buried surface area comes from residues that are in atomic contact ($\leq$ 4 Å), and the remaining from residues that are not in contact but close enough to exclude water (Fig. 3.2, Tables 3.2 and 3.3).

Figure 3.2. Structure of Mtd-P1 bound to Prn-E. (a) The complex is depicted in molecular surface representation, with each protomer in the Mtd-P1 trimer colored different shades of gray and Prn-E colored blue. (Top, right) Arrows designate 90° rotations that expose the interacting surfaces to view (green, hydrophobic contacts; red, hydrophilic contacts; yellow, hydrophobic buried; purple, hydrophilic buried). Contact residues are those having an interatomic Mtd-P1/Prn-E distance of $\leq$ 4 Å, and buried residues are those having an interatomic distance of > 4 Å but excluding water due to association. Two of the three receptor-binding sites on trimeric Mtd-P1 are occupied (major site binds pertactin loop 399-407, sequence TELPPIPGA, and minor site binds pertactin loop 190-199, sequence TLQPLQPEDL), and third one is empty (exposed in 90° rotated view at bottom left). (b) Ribbon representation of Mtd-P1 (protomers in red, pink, and purple) bound to Prn-E (blue). Prn-E loops that contact Mtd-P1 are colored green and labeled. For reference, the RGD loop is shown in light gray in the conformation observed for *B. pertussis* pertactin (Emsley et al., 1996), but is not modeled in the structure of the Mtd-P1/Prn-E complex.

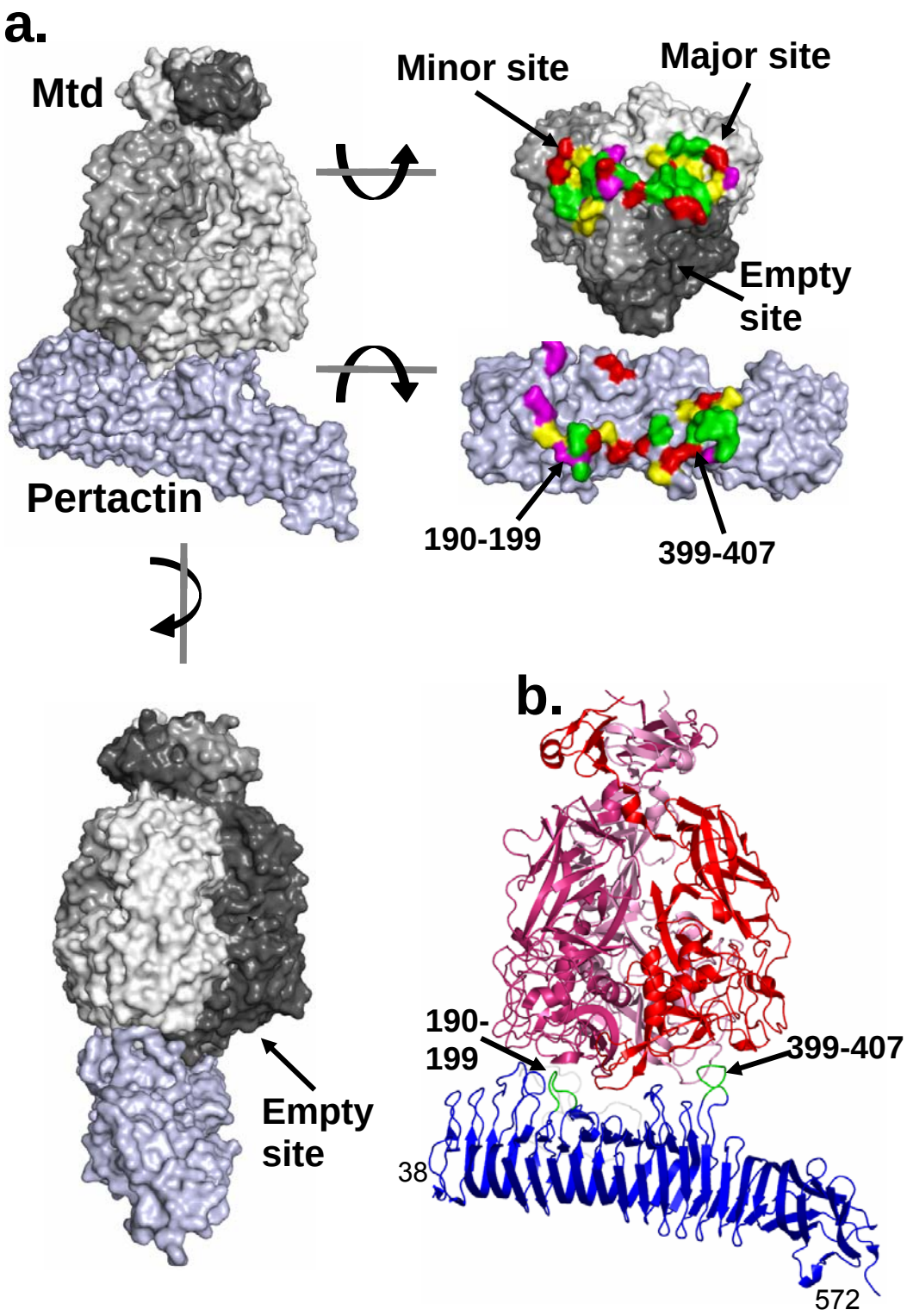

a.
Mtd
Minor site
Major site
Empty site
Pertactin
190-199
399-407
Empty site
b.
190-199
399-407
38
572

Table 3.2. Mtd-P1 residues buried by contact with Prn-E. [a]The average buried surface area of the two Mtd-P1/Prn-E complexes in the asymmetric unit of the crystal is reported. The probe radius used was 1.4 Å. [b]The major pertactin loop is composed of residues 399-407 and the minor pertactin loop of residues 190-199. The pertactin loop is indicated if it is contacted (≤ 4 Å) by the Mtd residue. [c] + denotes that residue is ≤ 4 Å of pertactin. — denotes that residue is > 4 Å of pertactin. [d]Residues in blue are variable.

Mtd residue (Chains)	Buried surface area (Å^2)[a]	Pertactin loop contacted[b]	Contact[c]
Tyr359 (B, D)[d]	92.0	Major	+
Tyr322 (C, E)	56.5	Major	+
Phe368 (B, D)	48.2	Major	+
Phe366 (B, D)	45.4	Major	+
Ala319 (B, D)	43.7	Minor	+
Tyr333 (C, E)	42.6	Major	+
Glu321 (C, E)	41.8	Major	+
Tyr333 (B, D)	40.3	Minor	+
Asn346 (B, D)	37.3	Major	+
Asn346 (A, F)	37.2	Minor	+
Phe366 (A, F)	36.2	Minor	+
Phe364 (B, D)	35.0		—
Ser320 (C, E)	34.0	Major	+
Leu350 (B, D)	33.5		—
Tyr359 (A, F)	32.1	Minor	+
Glu321 (B, D)	31.1	Minor	+
Glu321 (A, F)	30.8		—
Tyr322 (B, D)	29.8	Minor	+
Val316 (B, D)	28.4	Minor	+
Phe368 (A, F)	28.1	Minor	+
Ser320 (B, D)	24.2		—
Leu357 (B, D)	19.4	Major	+
Leu350 (A, F)	18.5	Minor	+
Asn317 (B, D)	18.2	Minor	+
Ser320 (A, F)	16.6		—
Thr348 (A, F)	15.8	Minor	+
Gly318 (B, D)	15.4	Minor	+
Leu357 (A, F)	14.6	Minor	+
Thr323 (B, D)	13.3		—
Phe369 (B, D)	9.8		—
Phe369 (A, F)	8.3	Minor	+
Gly347 (B, D)	5.3		—
Trp307 (A, F)	4.0		—
Ala324 (A, F)	2.6		—
Ser360 (B, D)	2.5	Major	+
Gly347 (A, F)	0.8		—

Table 3.3. Prn-E residues buried by contact with Mtd-P1. [a]The average buried surface area of the two Mtd-P1/Prn-E complexes in the asymmetric unit of the crystal is reported. [b]The Mtd binding site is indicated if it is contacted (≤ 4 Å) by the pertactin residue. The major Mtd the site is composed of variable residues from chain B and conserved residues from the β2β3 (residues 316-333) loop of chain C in one of the Mtd-P1/Prn-E complexes in the asymmetric unit of the crystal. In the other complex in the asymmetric unit, the major Mtd site is composed of variable residues from chain D and conserved residues from the β2β3 loop of chain E. The minor Mtd site is composed of variable residues from chain A and conserved residues from the β2β3 loop of chain B in one complex, and variable residues from chain F and conserved residues from the β2β3 loop of chain D in the other complex. The Mtd binding sites formed by variable residues of chains C and E are unoccupied.

Pertactin residue	Buried surface area (Å^2)[a]	Mtd Binding Site[b]	Contact[c]
Leu194	137.0	Minor	+
Ile404	117.0	Major	+
Pro405	101.0	Major	+
Pro403	85.1	Major	+
Thr399	84.3	Major	+
Gln192	82.2	Minor	+
Ala348	70.5	Minor	+
Leu401	70.1		—
Arg242	55.1		—
Arg350	40.3	Major	+
Leu191	37.9		—
Pro196	36.8		—
Asp250	35.2	Minor	+
Arg323	33.2	Minor	+
Ala407	32.9		—
Gln195	29.6		—
Val377	26.2		—
Gly347	24.1	Minor	+
Pro379	21.4	Major	+
Gly406	18.4	Major	+
Phe351	13.3		—
Arg349	12.9		—
Arg110	10.6		—
Val109	9.8		—
Pro402	7.5	Major	+
Tyr375	5.9		—
Gly346	5.5		—
Met249	3.4		—
Gly218	3.0		—

Pertactin contacts Mtd through two loops that emanate from its β-helix scaffold (Emsley et al., 1996) and nestle separately into two of the three receptor-binding sites of Mtd. The third receptor-binding site of Mtd is unoccupied and sterically occluded from binding another pertactin molecule. One of the contacting pertactin loops, composed of residues 399-407, forms the majority of interactions with a total of ~870 $Å^2$ being buried at this interface. The other loop, composed of residues 190-199, makes less extensive but nevertheless essential interactions (see below) with a total of ~585 $Å^2$ being buried at this interface. Based on buried surface area, pertactin loop 399-407 is termed the 'major loop' and the Mtd site to which it binds the 'major site', and likewise, loop 190-199 the 'minor loop' and the Mtd site to which it binds the 'minor site'.

The structure provides direct evidence that the variable residues of Mtd constitute receptor-binding sites. In Mtd-P1 these sites take the form of shallow pockets containing a central cluster of aromatic residues edged by a stripe of polar residues (Figs. 3.3). The aromatic residues dominate interactions, with the variable residue Tyr-359 and the invariant residues Tyr-322 and Tyr-333 being particularly significant in terms of buried surface area in both binding sites, and the variable residues Phe-366 and Phe-368 playing significant roles in the major site (Table 3.2). The invariant Tyr-322 and Tyr-333 are located on the β2β3 loop, which infiltrates into the binding site from a neighboring protomer, reinforcing the notion that the trimeric nature of Mtd is essential for function (McMahon et al., 2005). Although the three binding-sites of Mtd-P1 interact with different chemical environments, they are nearly

indistinguishable in conformation from one another and from unbound Mtd-P1 (Fig. 3.4). The static, lock-and-key nature of these sites is likely to lessen the entropic cost of binding. This mode differs considerably from the flexible binding sites of antibodies and TCRs, which are capable of induced fit (James et al., 2003). Despite their static nature, the Mtd sites are remarkably adaptable to different epitopes, as evidenced by their pseudosymmetric mode of interaction with pertactin.

In the major site, pertactin buries mostly hydrophobic residues (Leu401, Pro403, Ile404, Pro405) from loop 399-407 within the Mtd aromatic cluster, and in the minor site, it buries Leu-191, Gln-192, and Leu-194 from loop 190-199 (Figs. 3.3a and 3.3b, Table 3.3). The hydrogen bonding capacity of Gln-192 is satisfied by Mtd Tyr-322 and Tyr-333. These two side chain hydrogen bonds in the minor site and the two hydrogen bonds to main chain carbonyl atoms in the major site (Fig. 3.3c-d) are the only ones observed, in contrast to the ~9 hydrogen bonds of average antibody-antigen complexes (Lo Conte et al., 1999). The pertactin loops are highly flexible in their unbound state as demonstrated by high B-factors for *B. pertussis* pertactin (Emsley et al., 1996), and are likely to differ in conformation between bound and unbound states (Fig. 3.5). Residues interacting with Mtd-P1 are conserved between *B. bronchiseptica* and *B. pertussis*, except for Pro-403 which is substituted by Ser. The substitution is tolerated, as Mtd-P1 is found to bind *B. pertussis* Prn-E (data not shown). Unlike other regions of pertactin, neither loop is antigenically variant (Mooi, 1998).

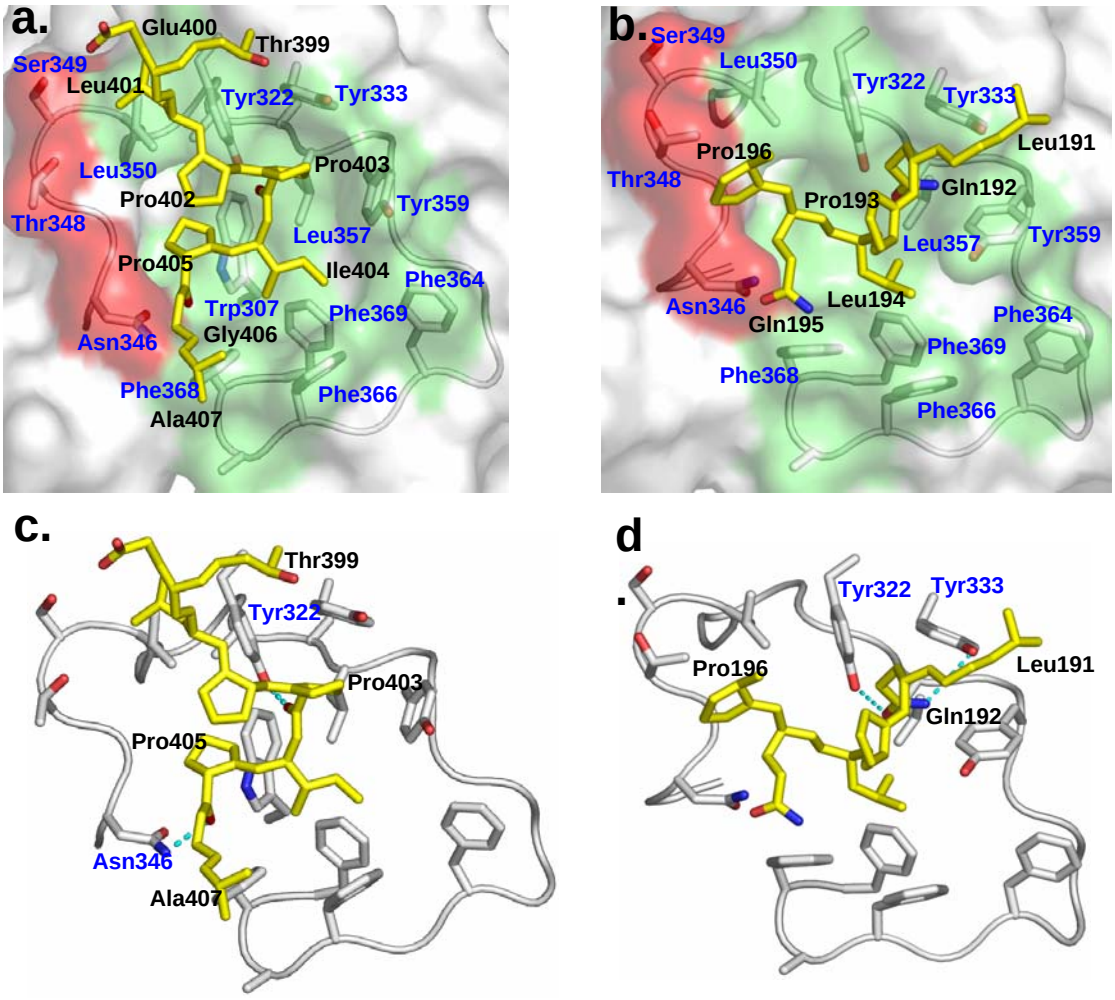


Figure 3.3. Interactions at major (a) and minor (b) binding sites. The molecular surface of Mtd is shown (green, hydrophobic residues; red, hydrophilic residues), with underlying backbone in gray and side chain carbons, oxygens, and nitrogens in gray, red, and blue, respectively. Mtd residues are labeled in blue. Pertactin residues (labeled in black) are shown in yellow stick representation. Hydrogen bonds at major (c) and minor (d) sites are shown as dashed lines. Atom coloring is as in panels a and b.

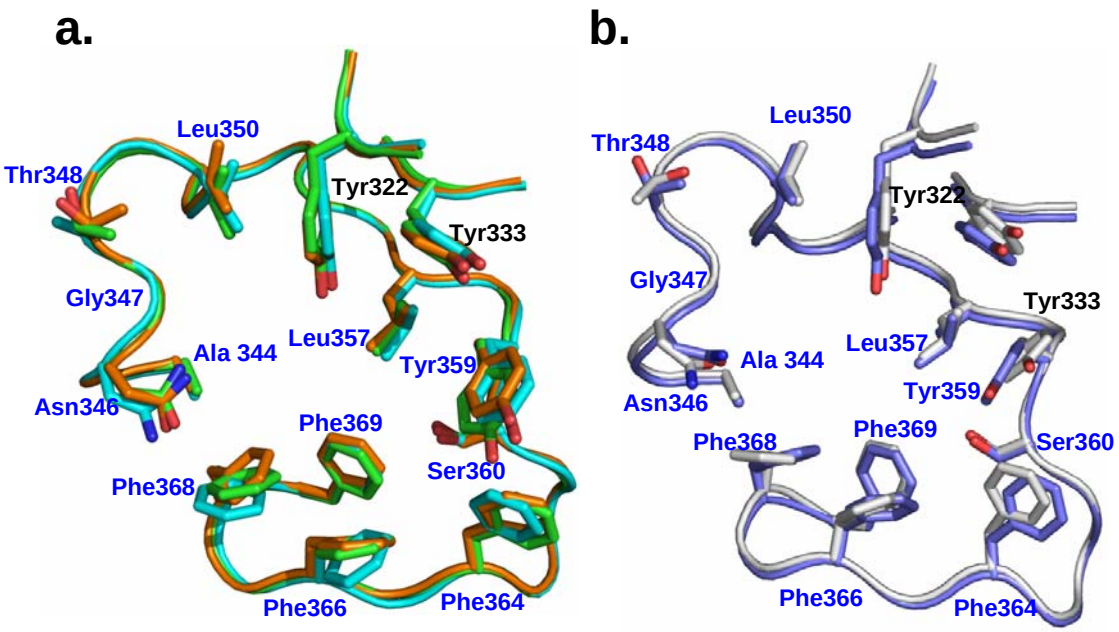
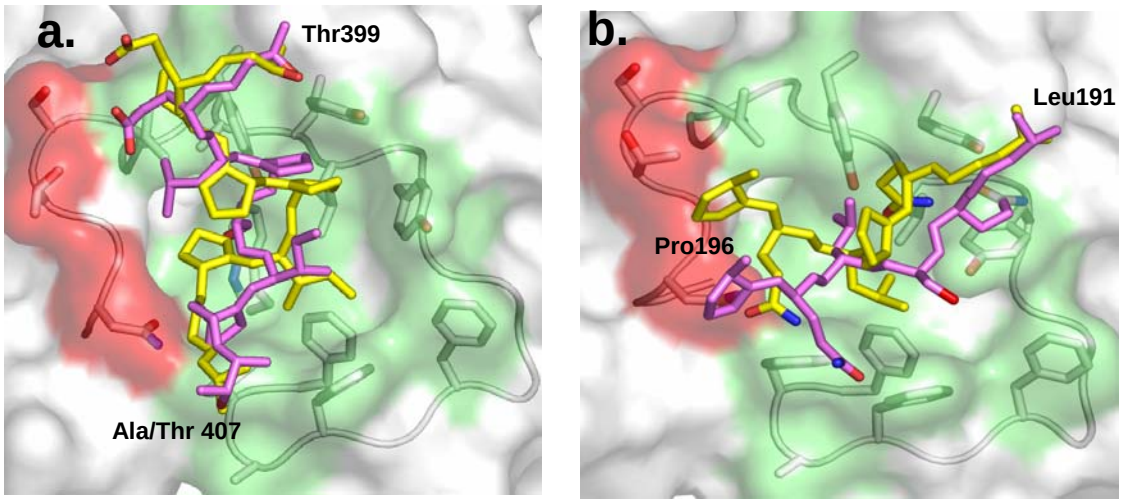


Figure 3.4. Superposition of Mtd-P1 variable regions. (a) The Mtd-P1 major (orange), minor (green), and empty (cyan) receptor-binding sites from the Mtd-P1/Prn-E complex are superposed (average r.m.s.d. 0.303 Å for 376 Cα atoms). (b) The major receptor-binding site (gray) of Mtd-P1 from the Mtd-P1/Prn-E complex is superposed with a receptor-binding site from unbound Mtd-P1 (blue) (r.m.s.d. 0.450 Å for 376 Cα atoms). The backbone is shown in curvilinear representation, and side chains of variable residues, along with the invariant residues Tyr-322 and Tyr-333, are shown in stick representation. Variable residues are labeled in blue.

Notably, the fit between Mtd and pertactin closely resembles that of an antibody-antigen complex, as quantified by surface complementarity (SC) (Lawrence and Colman, 1993). Complexes that have undergone evolutionary fine-tuning have high SC values of 0.70-0.74 (e.g., 0.77 for the interprotomer interfaces of Mtd, where 1.0 is perfect complementarity), whereas antibody-antigen interfaces have lower SC values of 0.64-0.68, reflecting their random and non-coevolved fit. The major and minor Mtd-pertactin sites are found to have SC values of 0.67 and 0.62, respectively, distinguishing these interactions as antibody-like. The small adjustments seen in side

chain positions upon binding are important for achieving this shape fit (Fig. 3.4b), as documented by much lower SC values of 0.55 and 0.56 at major and minor sites, respectively, for unbound Mtd-P1 modeled in complex with pertactin.




```
B.b  399 TELPPIPGA          B.b  190 TLQPLQPEDL
              :::: :::                    :: ::::::
B.p. 399 TELPSIPGT          B.p. 190 ALQSLQPEDL
```

Figure 3.5. Interactions at major (a) and minor (b) binding sites shown in the same representation as in Figure 3.3. Along with the conformation of major and minor loops from bound *B. bronchiseptica* pertactin (yellow sticks), the conformation of these same loops from unbound *B. pertussis* pertactin (Emsley et al., 1996) is shown (in purple sticks). Sequence alignment between *B. bronchiseptica* (*B.b.*) and *B. pertussis* (*B.p.*) pertussis in this region is shown below. The superposition is based on 492 Cα atoms of bound *B. bronchiseptica* pertactin and unbound *B. pertussis* pertactin (r.m.s.d. 0.83 Å).

Mtd-P1 and Mtd-M1 differ in pertactin epitopes

To test whether both pertactin loops are necessary for association with Mtd-P1, individual loop deletions were constructed. Mtd-P1 was observed to be severely attenuated in binding His-tagged Prn-E mutants lacking either major (Prn-E Δ399-407) or minor (Prn-E Δ190-199) loop as compared to wild-type Prn-E in an *in vitro* coprecipitation assay (Fig. 3.6). This indicates that both loops are necessary for association with Mtd, and conversely, that each of the Mtd sites confers a portion of the affinity. The requirement of two equivalent sites in Mtd for binding pertactin is a further difference from antibodies and TCRs (Amit et al., 1986; Garboczi et al., 1996), which have single and sufficient antigen-binding sites. Whether a single binding site in Mtd suffices to recognize other receptors requires further study.

The above structural and biochemical evidence that pertactin loops 190-199 and 399-407 are required for association with Mtd-P1 allowed us to ask whether Mtd-M1, a direct descendant of Mtd-P1 that binds pertactin non-productively (McMahon et al., 2005), retains the same epitope as Mtd-P1. Nine of 12 variable residues differ between Mtd-P1 and Mtd-M1, with most of the aromatic cluster in Mtd-P1 being substituted by polar residues, as shown by direct structural comparison (McMahon et al., 2005) (Fig. 3.1b, Fig. 3.7). These nine changes are sufficient to switch specificity from pertactin to an unknown receptor expressed by Bvg$^-$ *Bordetella* (Liu, 2002).

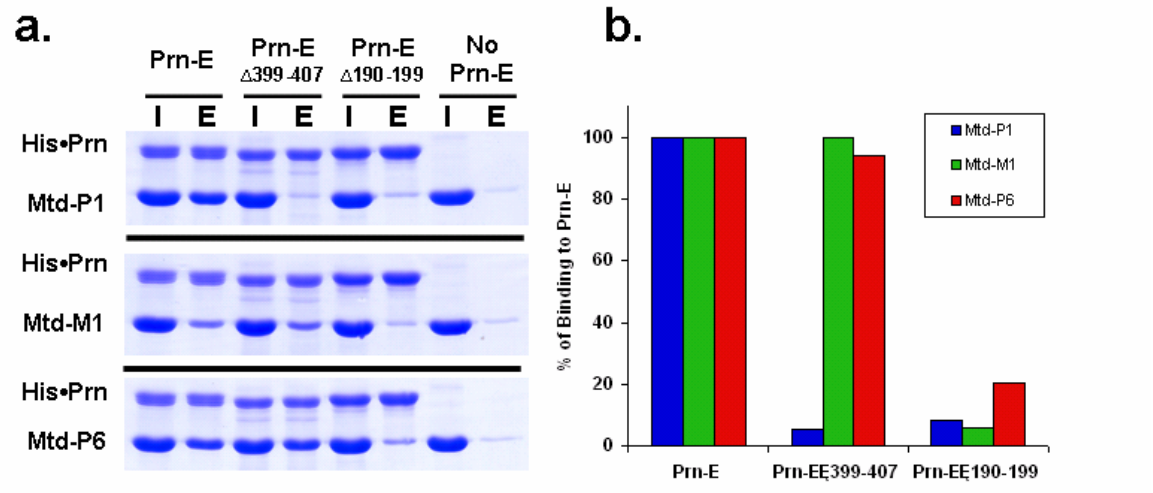


Figure 3.6. Dependency of Mtd interactions on pertactin loops. (a) Left, Coprecipitation of 60 μM (trimer concentration) Mtd-P1 (top), Mtd-M1 (middle), and Mtd-P6 with His-tagged Prn-E, Prn-EΔ399-407, Prn-EΔ190-199, or no Prn-E using Ni^{2+}-NTA beads and visualized by SDS-PAGE and Coomassie staining. I, relative amount incubated with Ni^{2+}-NTA beads; E, relative amount eluted from beads. Right, quantification of binding of Mtd-P1 (blue), Mtd-M1 (green), and Mtd-P6 (red) to Prn-EΔ399-407 and Prn-EΔ190-199, normalized by binding to Prn-E. Values were corrected for background binding (no Prn-E) to Ni^{2+}-NTA beads.

In contrast to Mtd-P1, Mtd-M1 binds Prn-E Δ399-407 equally well as Prn-E (Fig. 3.6). Loop 190-199 appears to form at least part of the epitope for Mtd-M1, as deletion of this loop reduces binding substantially (Fig. 3.6). However, we note that interaction between Mtd-M1 and pertactin is weak, and that partial loss of binding would be difficult to differentiate from complete loss in this assay. These results indicate that Mtd-M1 uses a different pertactin epitope than its progenitor Mtd-P1, and that the epitope for Mtd-M1 does not involve loop 399-407 but does require loop 190-199, at least in part.

Evolution of productive interactions

The observation of weak and non-productive interaction of Mtd-M1 with pertactin led us to ask whether phage genetics could be used to convert this interaction into a productive one. Descendants of BMP-1, the phage that expresses Mtd-M1, were selected for pertactin dependency of infection. One such progeny phage (BPP-6) was found to express an Mtd variant, called Mtd-P6, that differs at only five of 12 variable residues from Mtd-M1 (Fig. 3.1). These changes are predicted to enlargen the hydrophobic portion of the binding site and segregate hydrophobic from polar areas (Fig. 3.7). Interestingly, we find that deletion of loop 399-407 in Prn-E does not abrogate association with Mtd-P6, and deletion of loop 190-199 reduces association substantially but not fully (Fig. 3.6). As with Mtd-M1, the epitope for Mtd-P6 does not involve loop 399-407 but does require loop 190-199 at least in part, suggesting strongly that Mtd-P6 retains the same pertactin binding epitope as Mtd-M1 and does not revert to the Mtd-P1 epitope. These results provide formal evidence that Mtd is capable of recognizing more than one pertactin epitope in a productive manner.

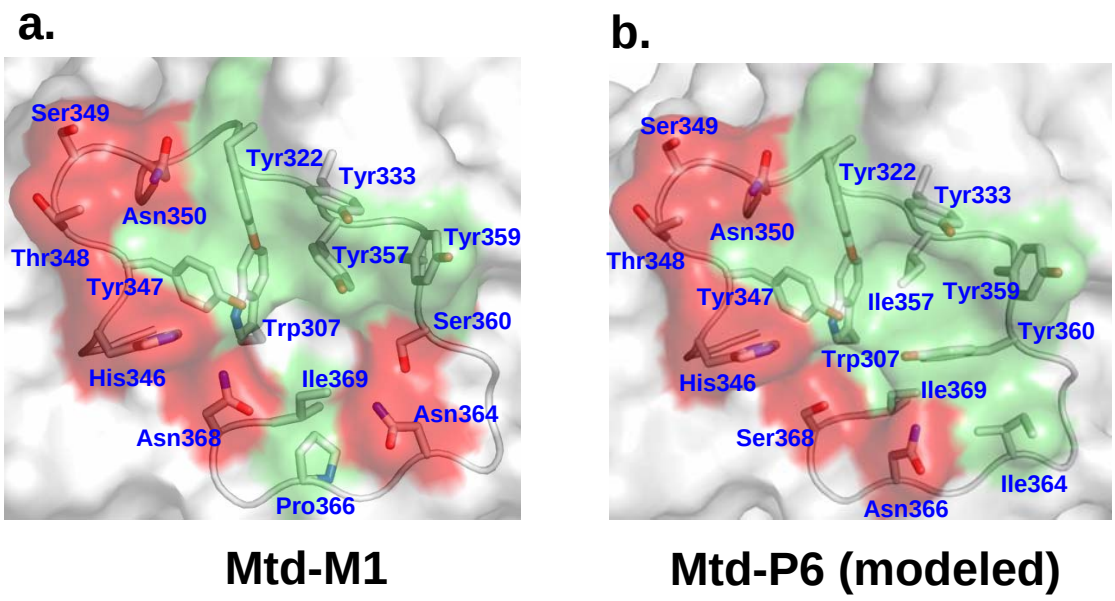


Figure 3.7. Molecular surface representation of Mtd-M1 (experimental data) and Mtd-P6 (modeled) receptor-binding sites. The surface of Mtd is shown (green, hydrophobic residues; red, hydrophilic residues), with underlying backbone in gray and side chain carbons, oxygens, and nitrogens in gray, red, and blue, respectively. Mtd residues are labeled in blue. The Mtd-P6 model is based on Mtd-M1, with the five Mtd-P6 residues differing from Mtd-M1 modeled in rotamer conformations that avoid steric clashes with neighboring atoms.

Surface plasmon resonance (SPR) was used to more precisely characterize the interaction between Mtd-P6 and pertactin. First, we validated the SPR experiment by finding that the dissociation constant (K_d) between soluble Mtd-P1 and biotinylated Prn-E immobilized on the surface of a streptavidin chip is 3.53 μM (Table 3.4; Fig. 3.8a), agreeing almost exactly with prior isothermal titration calorimetry (ITC) measurements (McMahon et al., 2005). The affinity of Mtd-M1 for pertactin was too low to be detected by SPR, as had also been the case for ITC, but a coprecipitation assay provides an estimate of ~500 μM for its apparent K_d (Fig. 3.9). A similar value is obtained by considering limits of detection in the SPR experiment (see methods). By comparison, the K_d of Mtd-P6 for pertactin is found to be 7.36 μM by SPR (Table 3.4; Fig. 3.8b), demonstrating that some or all of the five differences between Mtd-M1 and Mtd-P6 result in ~10^2-fold increase in affinity for a pertactin epitope likely shared between the two.

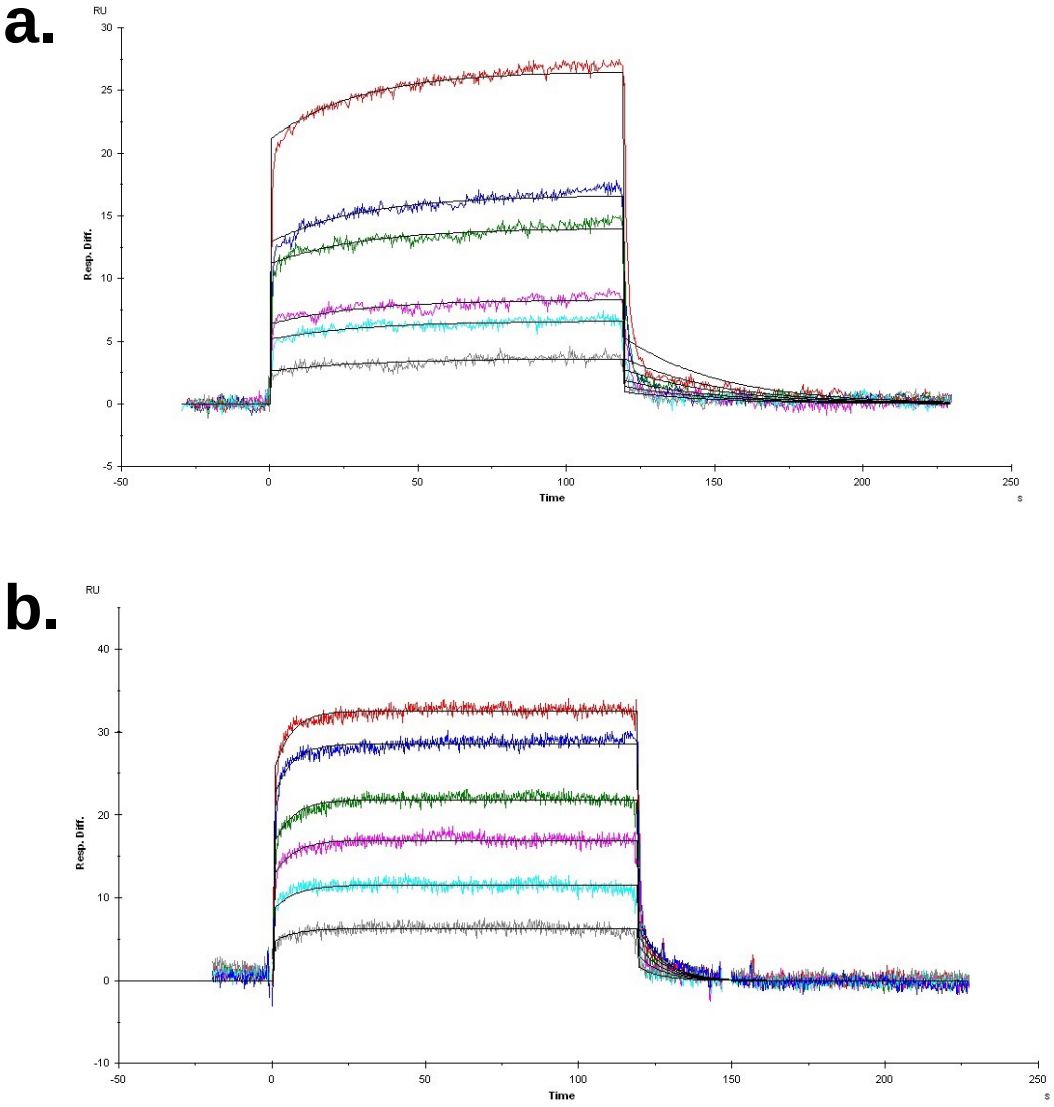


Figure 3.8. Surface plasmon resonance sensorgrams showing association of Mtd with biotinylated Prn-E immobilized on the surface of a streptavidin chip. (a) Concentrations of Mtd-P1: red, 124 x 10^{-9} M; blue, 83 x 10^{-9} M; green, 62 x 10^{-9} M; pink, 42 x 10^{-9} M; cyan, 31 x 10^{-9} M; gray, 21 x 10^{-9} M. (b) Concentrations of Mtd-P6: red, 1.50 x 10^{-6} M; blue, 1.25 x 10^{-6} M; green, 1.00 x 10^{-6} M; pink, 750 x 10^{-9} M; cyan, 500 x 10^{-9} M; gray, 250 x 10^{-9} M. Global fitting (black lines) was used to analyze association and dissociation curves.

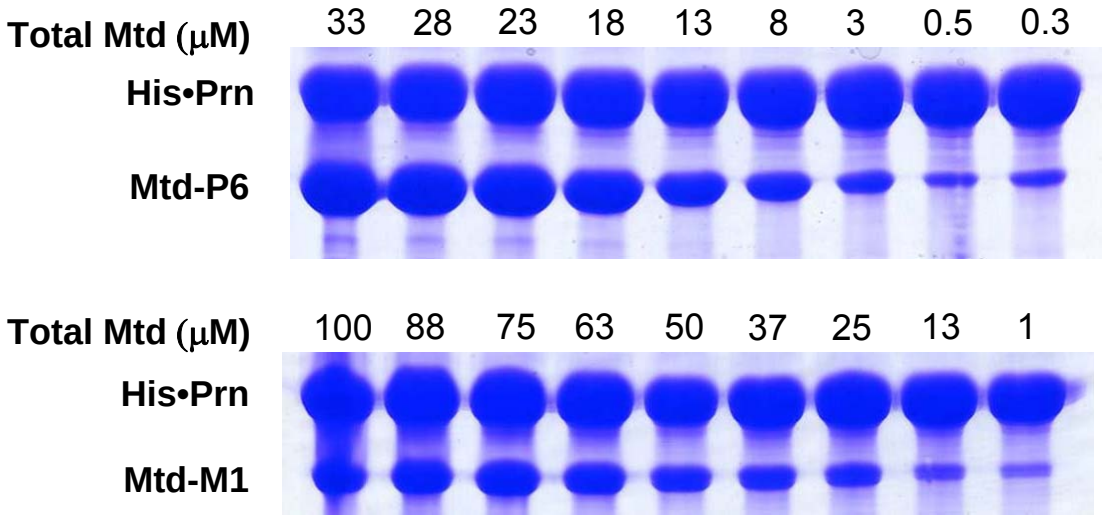


Figure 3.9. Association of Mtd-P6 and Mtd-M1 with pertactin. Coprecipitation of varying concentrations of Mtd-P6 and Mtd-M1 (total concentrations indicated above lanes) with 80 μM Prn-E using Ni^{2+}-NTA beads and visualized by SDS-PAGE and Coomassie staining. Concentrations of Mtd/Prn E complexes appear equivalent when 3 and 25 μM of Mtd P6 and Mtd M1, respectively, and 8 and 63 μM of Mtd P6 and Mtd M1, respectively, are incubated with Prn E.

Avidity amplifies differentially

While these measurements look at individual Mtd trimers, the situation is vastly different for the phage. The phage has six tail fibers and the best evidence indicates that each fiber carries two Mtd trimers (Liu, 2002; Liu, 2004) (A.H. and Z.H. Zhou , unpublished results), raising the possibility that avidity is an essential feature of interaction. We used SPR to test this and found that, under our experimental conditions, the low micromolar affinities of individual Mtd-P1 and Mtd-P6 molecules are dramatically amplified in the phage through avidity to apparent K_d's of 8.1 and 9.2 pM, respectively (Fig. 3.10, BPP-1 and BPP-6). For both Mtd-P1 and Mtd-P6, the ~10^6 enhancement in phage affinity partitions as a ~10^4-fold increase in on-rate and a ~10^2-fold decrease in off-rate (Table 3.4). The more prominent gains in on-rate suggest a complex association mechanism for both variants. The binding of these phage to pertactin is specific, as seen by the lack of binding by BIP-1 phage (Fig. 3.11), which expresses an Mtd variant (Mtd-I1) that has previously been shown to have no detectable affinity for pertactin (McMahon et al., 2005).

Most strikingly, we observed no interaction of BMP-1 phage with pertactin, despite the fact that the Mtd-M1 has demonstrable albeit weak affinity for pertactin (Fig. 3.11). However, at a ~10^4-fold higher concentration of BMP-1 than used for BPP-6, binding was observed (Fig. 3.10c; Fig. 3.11). Binding by highly concentrated BMP-1 has a much slower on-rate than observed for either of the other two phages but an off-rate that is apparently similar. A reliable fit to the data could not be achieved for BMP-1, but the 10^4-fold greater concentration required to detect binding suggests

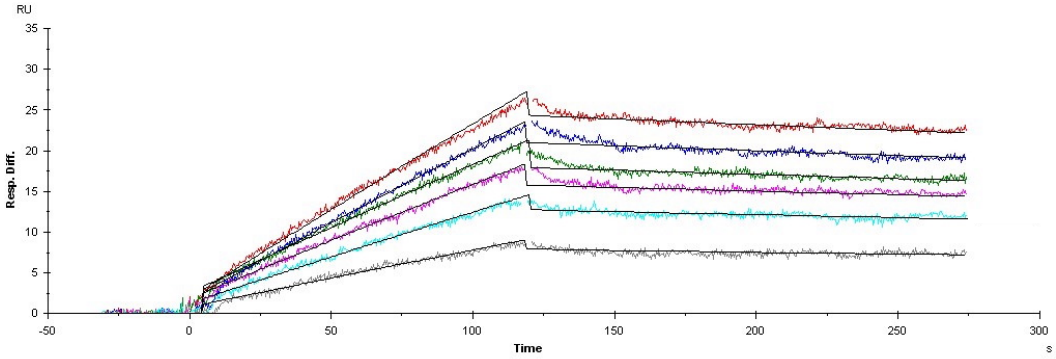
an apparent K_d of ~500 nM (see methods). Therefore, it appears that rather than the 10^6-fold amplification in affinity provided Mtd-P1 and Mtd-P6 (from 3-7 μM to 8-9 pM), avidity provides only a ~10^3-fold amplification to Mtd-M1 (from ~500 μM to ~500 nM). With differences being prominent in on-rates, it is likely that differential amplification stems from the on-rates of individual Mtd-M1 binding events being too slow for cooperativity to have a dramatic effect on association.

Table 3.4. Kinetics and thermodynamics of Mtd-pertactin interactions (SPR)

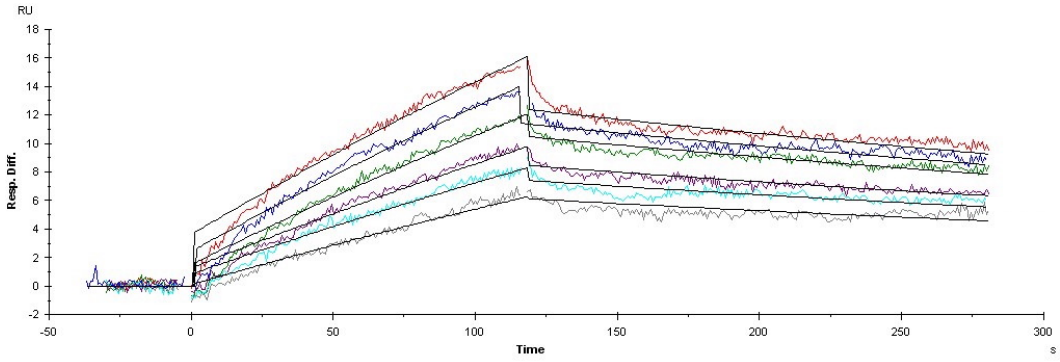
	k_{on} (1/Ms)	k_{off} (1/s)	K_a (1/M)	K_d (M)	χ^2
Mtd-P1	7.61×10^{3}	2.69×10^{-2}	2.83×10^{5}	3.53×10^{-6}	0.246
BPP-1	7.78×10^{7}	6.34×10^{-4}	1.23×10^{11}	8.14×10^{-12}	0.182
Mtd-P6	1.65×10^{4}	1.22×10^{-1}	1.36×10^{5}	7.36×10^{-6}	0.353
BPP-6	1.93×10^{8}	1.78×10^{-3}	1.08×10^{11}	9.26×10^{-12}	0.176
Mtd-M1	ND	ND			
BMP-1	ND	ND			
10,000x BMP-1	No fit	No fit			
BIP-1	ND	ND			

Figure 3.10. Surface plasmon resonance sensorgrams showing association of phage with biotinylated Prn-E immobilized on the surface of a streptavidin chip. (a) Concentrations of BPP-1 phage: red, 3.68 x 10^{-15} M; blue, 3.18 x 10^{-15} M; green, 2.70 x 10^{-15} M; pink, 2.40 x 10^{-15} M; cyan, 1.83 x 10^{-15} M; gray, 1.20 x 10^{-15} M. (b) Concentrations of BPP-6 phage: red, 2.04 x 10^{-16} M; blue, 1.93 x 10^{-16} M; green, 1.73 x 10^{-16} M; pink, 1.40 x 10^{-16} M; cyan, 1.22 x 10^{-16} M; gray, 1.01 x 10^{-16} nM. (e) Concentrations of BMP-1 phage: red, 8.50 x 10^{-12} M; blue, 7.50 x 10^{-12} M; green, 5.85 x 10^{-12} M; pink, 4.90 x 10^{-12} M; cyan, 3.62 x 10^{-12} M. Global fitting (black lines) was used to analyze association and dissociation curves, except for BMP-1, which could not be fit. No interactions were detected for Mtd-M1 in the concentration range 21 x 10^{-9} M to 2.1 x 10^{-6} M and for BIP-1 phage in the concentration range 8.3 x 10^{-16} M to 8.3 x 10^{-15} M. Concentrations of BMP-1 equivalent to those used for BPP-1 and BPP-6 (e.g., 3.68 x 10^{-15} M) did not yield detectable binding.

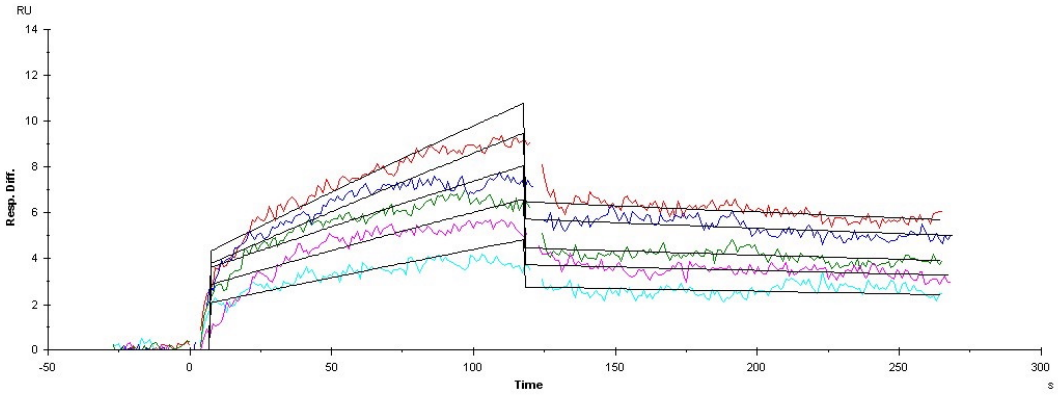
a.



b.



c.



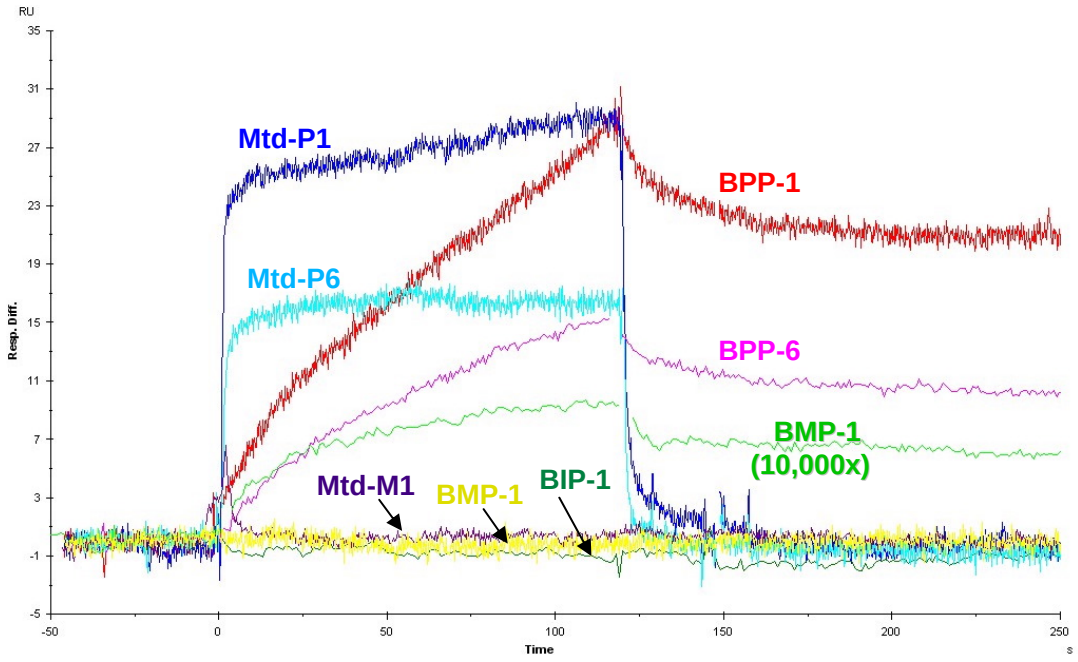


Figure 3.11. Affinity of Mtd and avidity of phage. Surface Plasmon Resonance sensorgrams showing association of Mtd-P1 (blue), Mtd-P6 (cyan), and Mtd-M1 (purple) molecules and of phage BPP-1 (red), BPP-6 (pink), BIP-1 (thin green), BMP-1 (yellow), and BMP-1 10,000x (thick green) with biotinylated Prn-E immobilized on a streptavidin chip. Curves displayed were chosen for clarity, and represent one of 6 (or 5, in the case of BMP-1 10,000x) different concentrations measured.

Discussion

Our results suggest that avidity is essential to sequence variable repertoires in amplifying affinity to the strengths required for biological results (e.g., infection). This likely stems from the fact that even with diversities as large as ~10^{13}-10^{16}, sequence variable repertoires typically produce micromolar affinity binders at best, as seen with Mtd and also in the adaptive immune system. Antibodies (from the primary response) and T cell receptors (TCR) have affinities for antigens and MHC-peptide complexes, respectively, that are usually in the micromolar range (Batista and Neuberger, 1998; Foote and Milstein, 1991; Van der Merwe and Davis, 2003). For both DGRs and the immune system, these low affinities reflect the fact that interactions with targets are random rather than co-evolved, in agreement with observations on surface complementarity. It should be noted that while high affinity (nanomolar) binders are rare, they are not unobserved; an example has been noted in the primary antibody response (Roost et al., 1995). In contrast to these micromolar affinities, biological processes require much stronger interactions. *Bordetella* bacteriophage infectivity occurs in the picomolar range of phage-receptor interactions, and protective antibody responses have been shown to require ~20-50 nM interactions (Bachmann et al., 1997).

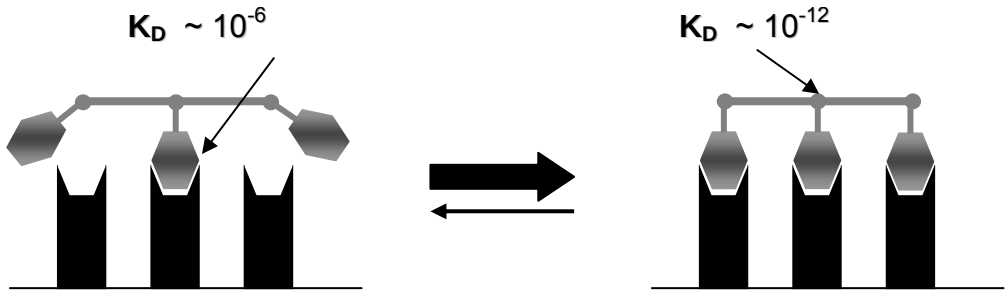
The leap from micromolar to picomolar comes from avidity, as shown by the dramatically stronger interaction of phage with pertactin as compared to individual Mtd molecules. This strength derives from multiple copies of Mtd trimers (most

likely 12 per phage) interacting with surface-localized pertactin. Cooperativity of these interactions results in high local concentrations of Mtd, accounting for the marked increases in on-rates (10^4) and decreases in off-rates (10^2) for the phage (Fig. 3.12). Of note, autotransporter proteins have been documented to cluster to the poles of bacterial cells (Jain et al., 2006), raising the possibility that pertactin is likewise clustered. Such clustering would facilitate avidity and provide an explanation for why pertactin is a preferred receptor for Mtd (A.H. and J.F.M., unpublished results). Consistent with our observations on Mtd, DGR-encoded variable proteins of *Bacteriodes thetaiotamicron* and *Treponema denticola* are predicted to be lipoproteins localized to the bacterial outer membrane and therefore available to interact cooperatively with host cell surface-bound targets (Doulatov, 2004). In the immune system, avidity has been observed in interactions of antibodies with viruses (Bachmann et al., 1997) as well as of IgG-bound antigens with cell surface-localized B cell receptors (Batista and Neuberger, 1998). Avidity has also been suggested to occur in intercellular interactions of T cell receptor with MHC-peptide complexes (Schamel et al., 2005).

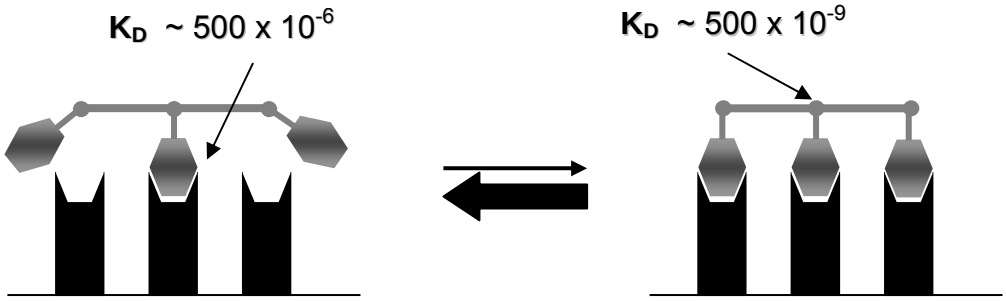
Lastly, our results indicate that the amplification provided by avidity is differential, and this has implications for specificity. It has been a puzzle as to how the immune system responds specifically and robustly to low micromolar affinities, but not to ones that are just somewhat weaker. For example, a 0.3 μM affinity interaction between an antigen and a B cell receptor was shown to trigger a specific B cell response, but one having a ~1.4 μM affinity was found to be inert in eliciting a

specific response (Batista and Neuberger, 1998). Interestingly, avidity was shown to convert the ~1.4 μM affinity interaction into a productive one, but the affinity boundary beyond which avidity has no productive effect was undetermined. We have found such a boundary in the ~500 μM interaction of Mtd-M1 with pertactin. Although values of such boundaries are likely to depend on the detailed nature of interactions, our results indicate that binding events in the micromolar range are sensitive to the differential amplification of avidity.

Avidity is seen to amplify ~10^2-fold differences in the micromolar range into ~10^5-fold differences in the picomolar range, clearly distinguishing interactions that support infectivity from those that do not. The micromolar range is likely key to this differential effect, since lower affinities are amplified ineffectively (as shown here) and much higher affinities (e.g., nanomolar or better) are expected to be above thresholds required for productive interaction or amplified indiscriminately by avidity above such thresholds. This latter point is consistent with the observation of an affinity ceiling for antibodies (Batista and Neuberger, 1998). The differential aspect of avidity is seen to endow specificity by limiting the number of productive interactions available to any particular Mtd variant. This limit is likely beneficial to the immune system in enhancing distinctions between stimulatory (non-self) and non-stimulatory (self) antigens, and is also likely to dictate functional interactions of other DGR-encoded variable proteins with their targets.



Multiple simultaneous interactions increase on-rate and decrease off-rate for the phage, resulting in 10^6-fold amplification



Multiple simultaneous interactions do not occur frequently, resulting in a lower amplification in the phage of 10^3-fold

Figure 3.12. Differential aspect of avidity. Top, slow off rates and fast on rates, corresponding to micromolar dissociation constants in Mtd, enables multiple simultaneous contacts to be made. Cooperativity of these interactions amplifies individual binding events to yield a 10^6-fold greater affinity for the phage. Bottom, faster off rates and slower on rates, corresponding to a dissociation constant of ~500 μM, results in cooperative interactions occurring less frequently than in the top panel. This results in an amplification of only 10^3-fold in the right panel. A valency of only three is shown for simplicity.

Bibliography

Akerley, B. J., Cotter, P. A., and Miller, J. F. (1995). Ectopic expression of the flagellar regulon alters development of the Bordetella-host interaction. Cell *80*, 611-620.

Alder, M. N., Rogozin, I. B., Iyer, L. M., Glazko, G. V., Cooper, M. D., and Pancer, Z. (2005). Diversity and Function of Adaptive Immune Receptors in a Jawless Vertebrate 10.1126/science.1119420. Science *310*, 1970-1973.

Amit, A. G., Mariuzza, R. A., Phillips, S. E., and Polijak, R. J. (1986). Three-dimensional structure of an antigen-antibody complex at 2.8 A resolution. Science *233*, 747-753.

Bachmann, M. F., Kalinke, U., Althage, A., Freer, G., Burkhart, C., Roost, H., Aquet, M., Hengartner, H., and Zinkernagel, R. M. (1997). The role of antibody concentration and avidity in antiviral protection. Science *276*, 2024-2027.

Batista, F. D., and Neuberger, M. S. (1998). Affinity dependence of the B cell response to antigen: a threshold, a ceiling, and the importance of off-rate. Immunity *8*, 751-759.

Cotter, P. A., and Miller, J. F. (1997). A mutation in the Bordetella bronchiseptica bvgS gene results in reduced virulence and increased resistance to starvation, and identifies a new class of Bvg-regulated antigens. Mol Microbiol *24*, 671-685.

Davis, M. M., and Bjorkman, P. J. (1988). T-cell antigen receptor genes and T-cell recognition. Nature *334*, 395-402.

Doulatov, S. (2004). Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. Nature *431*, 476-481.

Emsley, P., Charles, I. G., Fairweather, N. F., and Isaacs, N. W. (1996). Structure of Bordetella pertussis virulence factor P.69 pertactin. Nature *381*, 90-92.

Foote, J., and Milstein, C. (1991). Kinetic maturation of an immune response. Nature *352*, 530-532.

Garboczi, D. N., Ghosh, P., Utz, U., Fan, Q. R., Biddison, W. E., and Wiley, D. C. (1996). Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *384*, 134-141.

Higuchi, R., Krummel, B., and Saiki, R. K. (1988). A general method of in vitro preparation and specific mutagenesis of DNA fragments: study of protein and DNA interactions. Nucleic Acids Res *16*, 7351-7361.

Jain, S., van Ulsen, P., Benz, I., Schmidt, M. A., Fernandez, R., Tommassen, J., and Goldberg, M. B. (2006). Polar localization of the autotransporter family of large bacterial virulence proteins. J Bacteriol *188*, 4841-4850.

James, L. C., Roversi, P., and Tawfik, D. S. (2003). Antibody multispecificity mediated by conformational diversity. Science *299*, 1362-1367.

Jones, T. A., Zou, J.-Y., Cowan, S. W., and Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr A *47*, 110-119.

Lawrence, M. C., and Colman, P. M. (1993). Shape complementarity at protein/protein interfaces. J Mol Biol *234*, 946-950.

Liu, M. (2002). Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage. Science *295*, 2091-2094.

Liu, M. (2004). Genomic and genetic analysis of Bordetella bacteriophages encoding reverse transcriptase-mediated tropism-switching cassettes. J Bacteriol *186*, 1503-1517.

Lo Conte, L., Chothia, C., and Janin, J. (1999). The atomic structure of protein-protein recognitions sites. J Mol Biol *285*, 2177-2198.

Mattoo, S., and Cherry, J. D. (2005). Molecular pathogenesis, epidemiology, and clinical manifestations of respiratory infections due to Bordetella pertussis and other Bordetella subspecies. Clin Microbiol Rev *18*, 326-382.

McMahon, S. A., Miller, J. L., Lawton, J. A., Kerkow, D. E., Hodes, A., Marti-Renom, M. A., Doulatov, S., Narayanan, E., Sali, A., Miller, J. F., and Ghosh, P. (2005). The C-type lectin fold as an evolutionary solution for massive sequence variation. *12*, 886-892.

Mooi, F. R. (1998). Polymorphism in the Bordetella pertussis virulence factors P.69/pertactin and pertussis toxin in The Netherlands: temporal trends and evidence for vaccine-driven evolution. Infect Immun *66*, 670-675.

Otwinowski, Z., and Minor, W. (1997). Processing of x-ray diffraction data collected in oscillation mode. Methods Enzymol *276*, 307-326.

Roost, H., Bachmann, M., Haag, A., Kalinke, U., Pliska, V., Hengartner, H., and Zinkernagel, R. (1995). Early High-Affinity Neutralizing Anti-Viral IgG Responses without Further Overall Improvements of Affinity 10.1073/pnas.92.5.1257. PNAS *92*, 1257-1261.

Schamel, W. W. A., Arechaga, I., Risueno, R. M., van Santen, H. M., Cabezas, P., Risco, C., Valpuesta, J. M., and Alarcon, B. (2005). Coexistence of multivalent and monovalent TCRs explains high sensitivity and wide range of response 10.1084/jem.20042155. J Exp Med *202*, 493-503.

Uhl, M. A., and Miller, J. F. (1996). Integration of multiple domains in a two-component sensor protein: the Bordetella pertussis BvgAS phosphorelay. EMBO J *15*, 1028-1036.

Van der Merwe, P. A., and Davis, S. J. (2003). Molecular interactions mediating T-cell antigen recognition. Annu Rev Immun *21*, 659-684.

Winn, M. D. (2003). An overview of the CCP4 project in protein crystallography: an example of a collaborative project. J Synchrotron Radiat *10*, 23-25.

IV.

Discussion

Discussion

Variable proteins are found widespread throughout nature's macromolecular design, from eukaryotic to prokaryotic species and even viruses. Protein variability allows organisms to adapt to their surroundings, either for evasion of predatory proteins or recognition. In nature, it is the latter which requires greater adaptability. Whether it is for antigen recognition by T-cell receptors and antibodies of the eukaryotic immune system or receptor recognition by the *Bordetella* phage major tropism determinant (Mtd), these proteins must accommodate massive variability in order to bind almost any ligand it is presented. The end result of this adaptation is the ability to generate approximately 10^{14}-10^{16} unique sequences for antibodies and T-cell receptors (Davis and Bjorkman, 1988) or nearly 10^{13} sequence variants for Mtd (Liu, 2002). For proteins that are preyed upon, only slight changes in sequence are sufficient to evade recognition. As a result, prey proteins tend to have sequence diversities that reach only ~10^{3} sequences.

The C-type lectin fold (CLec) stabilizes the variable region (VR) of Mtd, unlike the Ig-fold that forms the scaffolding for antibodies and T-cell receptors variability. This is the first observation of a protein utilizing the CLec as a platform for sequence variability and its homology with other diversity generating retroelement (DGR) proteins identifies a new class of CLec stabilized variable proteins (Doulatov, 2004; Liu, 2004; McMahon et al., 2005). Unlike variable Ig-fold proteins, the number of Mtd VR residues is fixed and are positioned on a static backbone. A potential benefit of this design is a decreased chance of misfolding due to unstable sequence

combinations. Alternatively, the ability to present desirable structural conformations for receptor recognition is limited with a scaffold that does not allow sequence length or conformation variability.

Despite these differences, Mtd binds its receptor with antibody-like shape complementarity and the affinity of a primary antibody/antigen interaction (Foote and Milstein, 1991; Lawrence and Colman, 1993). Both Mtd and primary antibodies bind their specific targets with low micromolar affinities, near the approximate 1 μM threshold determined for a B-cell response (Batista and Neuberger, 1998). Although the similarity in binding affinities falls within the limits for a productive interaction observed in Ig-fold proteins during a primary immune response, the evolution of binding specificity for a specific antigen or receptor occurs by uniquely different methods. Specificity between antibody-antigen interactions is improved during affinity maturation where antibody affinities for their respective antigen are refined to the nanomolar range, thus initiating protective antibody responses (Foote and Milstein, 1991). For *Bordetella* phage infectivity, specificity is enhanced by the avidity created from multiple Mtd trimers attached to the tails of the phage. In this case, Mtd affinity for its receptor need only reach the low micromolar threshold for a productive interaction, then avidity acts as a differential amplifier to improve specificity.

The use of avidity for enhanced ligand specificity is not entirely unique to *Bordetella* phage Mtd. Avid effects have been observed between interactions of IgG-bound antigens with B-cell surface receptors as well as implicated in T-cell receptor/MHC-peptide complexes (Batista and Neuberger, 1998; Schamel et al., 2005). Additionally, other DGR-encoded variable proteins are predicted to localize to

the bacterial cell surface where they could potentially act cooperatively to bind host antigens. Avidity is a powerful tool for specific enhancement; however, receptor discrimination occurs at the individual protein level. Why is it that a low micromolar affinity (3-7 μM) translate to an apparent picomolar binding event resulting in infectious phage, but a high micromolar affinity (500 μM) fails to produce infectious phage? These data show that there is a thermodynamically driven amplification where high micromolar affinities are only amplified by a factor of 10^3, whereas low micromolar affinities are amplified by a factor of 10^6 by *Bordetella* phage. While this difference in amplification provides specificity to a particular receptor, it also limits the number of receptors available to the phage for binding. For better or worse, phage receptor recognition is governed by this differential amplifier that discriminates specific interactions from non-specific ones.

Future Directions

The massive variability accommodated by the Mtd CLec-fold (~10^{13}) is rivaled only by the Ig-fold of antibodies and T-cell receptors (~10^{14}-10^{16}) and the leucine-rich repeat (LRR) fold (~10^{14}) of the lymphocyte receptors found in jawless hagfish (Alder et al., 2005). The addition of just one mutable codon (AAC) to the Mtd TR increases its potential amino acid sequence diversity (10^{14}) to that seen in hagfish LRR receptors and the Ig-fold proteins of the immune system. Further addition of 2 more mutable condons extends Mtd variability to the limits of the Ig-fold (10^{16}). Structural analysis

of the Mtd VR indicates several locations where amino acid mutations or insertions may occur without disrupting the stable scaffold of the CLec. Both Thr348 and Ala367 side chains are positioned toward the periphery of the Mtd trimer, away from potential inter- or intra-molecular contacts, and are solvent exposed. Ala367 is an invariant residue located near the C-terminal end of the VR that could be directly mutated to code for 15 variable amino acids by switching its TR codon to AAC. Thr347 is a variable residue stabilized by the 3_{10} helix located on the VR; however, its TR nucleotide sequence (CAC) dictates only 4 possible amino acid variations. Mutation to the more variable AAC codon would increase its diversity nearly 4-fold.

The suggested mutations of the variable region could theorectically increase the structural diversity of the Mtd VR 50-fold (5×10^{14}), however, a loop insertion of just 2 amino acids could raise this projected variation to nearly 10^{16} sequences (7.8×10^{15}). Since most of the Mtd VR backbone is stabilized by a network of hydrogen bonds, there are limited positions where insertions would be tolerated without potentially disrupting the CLec. One exception can be found at the loop region between residues 365-368. Even though structural alignment indicates virtually no variation in backbone conformation at this region, it is without stabilization by molecular contacts and positioned with complete solvent accessibility. Both residues 366 and 368 are variable amino acids in Mtd; however, variable loops could be introduced between 365 and 366 or 366 and 367. I propose that this stretch of the variable loop region is the most likely candidate for amino acid insertion due to its lack of molecular contacts and exposure to the surrounding environment, thereby increasing the variability and size of the Mtd binding surface.

Stabilization of Mtd variability is accomplished by the CLec and trimeric association of the Mtd protomers. We have identified approximately 20 other DGR variable proteins, many of which are predicted to be stabilized by the CLec. Some of these are believed to have VR diversity comparable to Mtd or possible greater. Currently, no structure of a DGR protein other than Mtd has been solved while new homologs are identified on an annual basis. Preliminary studies with several Mtd protein homologs from *Nostoc* species (N. PCC 1 and N. PCC 2B) have yield two soluble targets for crystallization and structure determination. In addition, initial protein expression trials with an Mtd homolog discovered in *Bacteroides thetaiotaomicron* resulted in no soluble protein production for crystallization. Finally, I have cloned the DGR-encoded protein from *Treponema denticola* for bacterial overexpression. The protein is of particular interest since it contains a predicted N-terminal lipoprotein signal sequence to direct expressed protein to the bacterial cell surface. It is possible that *Treponema* use this protein for host-receptor recognition. Structural characterization of these DGR related proteins would give valuable insight into the diversity tolerated by the CLec or other protein folds.

The potential development of large libraries of ligands using Mtd as a template for design is an attractive possibility. Similar to antibodies, the Mtd CLec platform can accommodate massive sequence variation and exhibits antibody-like binding affinity. Unfortunately the components necessary for DNA sequence variation and reinsertion into the Mtd locus is poorly understood. It is likely the *Bordetella* reverse-transcriptase is responsible for the mutagenesis of template RNA, however, the mechanism for reinsertion of the mutated DNA copy into the VR remains a mystery.

Recent studies have identified the *atd* coding region located within the DGR cassette as an essential gene for phage tropism switching; however, further experiments are required to identify the exact role of this protein in Mtd variation. Understanding the mechanism of Mtd variation is essential for its use in the future development of ligands for biochemical applications. Additionally, *Bordetella* phage could serve as a useful vector for therapeutic delivery in patients.

Bibliography

Alder, M. N., Rogozin, I. B., Iyer, L. M., Glazko, G. V., Cooper, M. D., and Pancer, Z. (2005). Diversity and Function of Adaptive Immune Receptors in a Jawless Vertebrate 10.1126/science.1119420. Science *310*, 1970-1973.

Batista, F. D., and Neuberger, M. S. (1998). Affinity dependence of the B cell response to antigen: a threshold, a ceiling, and the importance of off-rate. Immunity *8*, 751-759.

Davis, M. M., and Bjorkman, P. J. (1988). T-cell antigen receptor genes and T-cell recognition. Nature *334*, 395-402.

Doulatov, S. (2004). Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. Nature *431*, 476-481.

Foote, J., and Milstein, C. (1991). Kinetic maturation of an immune response. Nature *352*, 530-532.

Lawrence, M. C., and Colman, P. M. (1993). Shape complementarity at protein/protein interfaces. J Mol Biol *234*, 946-950.

Liu, M. (2002). Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage. Science *295*, 2091-2094.

Liu, M. (2004). Genomic and genetic analysis of Bordetella bacteriophages encoding reverse transcriptase-mediated tropism-switching cassettes. J Bacteriol *186*, 1503-1517.

McMahon, S. A., Miller, J. L., Lawton, J. A., Kerkow, D. E., Hodes, A., Marti-Renom, M. A., Doulatov, S., Narayanan, E., Sali, A., Miller, J. F., and Ghosh, P. (2005). The C-type lectin fold as an evolutionary solution for massive sequence variation. *12*, 886-892.

Schamel, W. W. A., Arechaga, I., Risueno, R. M., van Santen, H. M., Cabezas, P., Risco, C., Valpuesta, J. M., and Alarcon, B. (2005). Coexistence of multivalent and monovalent TCRs explains high sensitivity and wide range of response 10.1084/jem.20042155. J Exp Med *202*, 493-503.

V.

Appendix A.

Directed Evolution of

Helicobacter pylori VacA

Abstract

Helicobacter pylori, which infects over half the worlds population, colonizes the mucosal layer of the stomach epithelial cells (Suerbaum and Michetti, 2002). *H. pylori* secretes a vacuolating toxin, VacA (88 kDa), which is associated with the destruction of the gastric epithelium in mice and cell death in cultured epithelial cells (Fujikawa et al., 2003; Smoot et al., 1996). The vacuolation activity of VacA has been localized to a 44 kDa N-terminal fragment (residues 34-455, called VacA-455) (Ye and Blanke, 2000). Unfortunately neither VacA nor VacA-455 is produced by either *H. pylori* native expression or recombinant *Escherichia coli* expression at levels suitable for biochemical and structural experiments. While VacA-455 has been overexpressed at high levels in *E. coli* as inclusion bodies, it has not been successfully refolded despite numerous attempts in other labs. To address this major impediment in understanding the mechanism of action of VacA, I attempted to produce soluble forms of VacA-455 using the technique of directed evolution. Directed evolution has been used successfully in generating soluble variants for a number of proteins, some of which have been crystallized (Pedelacq et al., 2002). My preliminary results show variations of established directed evolution protocols that achieve larger libraries for screening, and evidence for initial mutations that slightly increase the solubility of VacA-455. Although solubility of several VacA-455 mutants increased, studies were hampered by proteolytic degradation of solubilized protein constructs.

Introduction

VacA Function

Helicobacter pylori is a Gram-negative bacterium which colonizes the gastric mucosa of the human stomach. Once established, it secretes a vacuolating cytotoxin (VacA) that is known to be an essential virulence factor in *H. pylori* pathogenesis (Montecucco et al., 1999). Infection of *H. pylori* is associated with chronic gastritis, peptic ulcer disease, and gastric carcinoma (Telford et al., 1994). Introduction of VacA into cultured mammalian cells results in the induction of cytoplasmic vacuolues and accelerated rates of apoptosis (Peek Jr et al., 1999).

The *vacA* gene codes for a 140 kDa protoxin, which is processed by *H. pylori* to form an 88 kDa mature toxin (Fig. 1). During secretion, an amino-terminal signal sequence (residues 1-33) and a C-terminal 40 kDa fragment are removed by proteolysis. The 40 kDa fragment is likely to form a porin-like β-barrel pore in the bacterial outer membrane, through which mature VacA is secreted. Mature VacA is often found nicked by a protease at a sensitive loop connecting an N-terminal p37 fragment and a C-terminal p58 fragment. It has been shown that only the N-terminal 422 residues (34-455, 44 kDa) are required for vacuolation activity in mammalian cells (Ye et al., 1999)(Fig. 5.1).

VacA exists as a relatively inactive oligomer (hexamers mainly) in solution, but is known to disassociate into highly active monomers upon exposure to acidic or basic conditions (McClain et al., 2000). The current mechanism of cellular entry

begins with direct membrane association of VacA, possibly binding to the EGF receptor (Seto et al., 1998), after which it is endocytosed and translocated into the cytoplasm in an undefined way. VacA also interacts with a protein tyrosine phosphatase receptor on the cell surface (Fujikawa et al., 2003), although this interaction is not involved in toxin internalization. Rather, this interaction leads to signaling changes that promote tissue damage. Following VacA entry, the late endocytic pathway is altered and vacuolation of cellular endosomes and lysosomes occurs (Fig. 5.2).

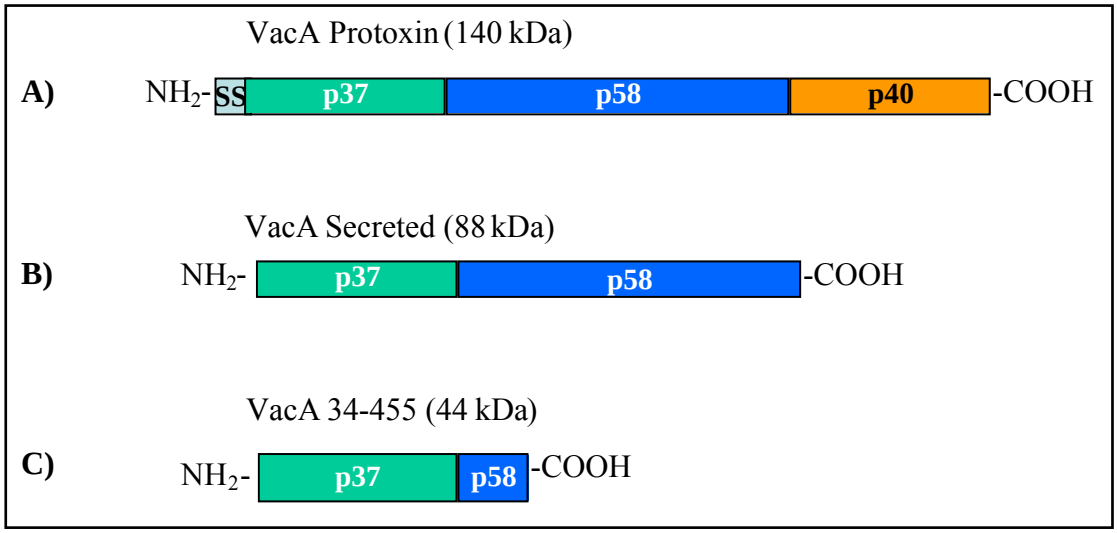


Figure 5.1. Schematic of VacA. A) The *vac*A gene codes for a protoxin that contains a short N-terminal signal sequence (SS) and a C-terminal (p40) domain, which are cleaved during secretion. **B)** Secreted VacA contains a p37 and p58 domain that are connected by a protease sensitive loop. **C)** Minimal fragment of VacA (VacA-455)required for vacuolating activity.

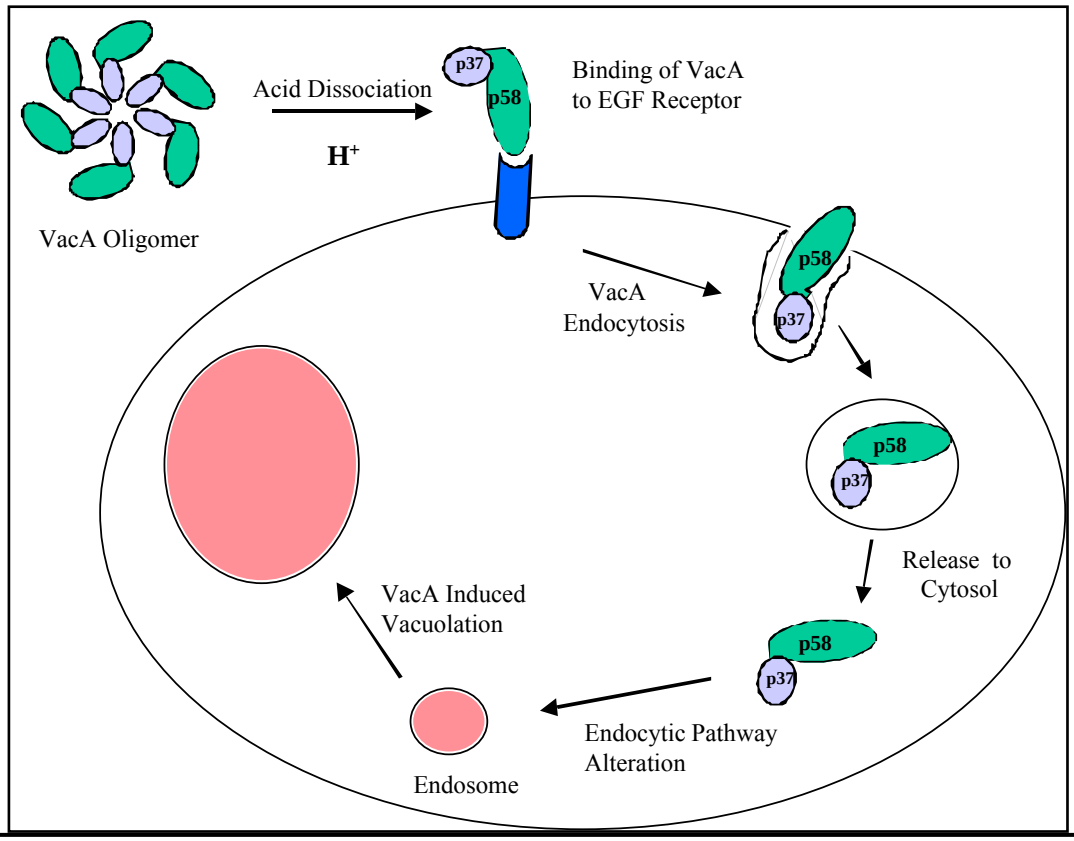


Figure 5.2. Model of VacA intoxication. VacA exists as an inactive oligomer that dissociates into its active components following acidification. Active VacA then binds to a surface receptor followed by the opening of anion-selective channels. This allows VacA to translocate to the cytoplasm, driven by pH grandient changes across the cell membrane. Upon entry, VacA alters the late endocytic pathways causing induced vacuolation of endosomes by an unknown mechanism

The mechanism by which VacA promotes formation of vacuoles is unkown. A major impediment to investigating this is the lack of soluble VacA in quantities sufficient for biochemical and structural studies. Currently, VacA is purified from *H. pylori*, which demands stringent growth conditions and produces small amounts of the mature toxin. Overexpression of mature VacA in *E. coli* (as carried out in our laboratory) results in the formation of inclusion bodies that have proven resistant to

solubilization by refolding. One report indicates mature VacA can be produced in soluble form in *E. coli*, but expression is detectable only by Western blot (McClain and Cover, 2003). Likewise, the minimal vacuolating fragment VacA-455 has been overexpressed in *E. coli*, but forms insoluble inclusion bodies that have also proven resistant to refolding. Novel approaches to obtaining soluble VacA-455 are required for elucidation of the basis of action of this protein.

Directed Evolution

Directed evolution has been applied in a number of cases to obtain soluble forms of proteins that are otherwise insoluble when overexpressed by recombinant means (Kuk Yang et al., 2003; Waldo, 2003; Waldo et al., 1999). Although random mutagenesis strategies such as DNA shuffling and error-prone PCR are capable of producing large libraries of mutant sequences (Arnold and Zhao, 1997; Stemmer, 1994a; Stemmer, 1994b), efficient methods for screening for mutants that increase solubility are required. Several studies suggest that green fluorescent protein (GFP) can be used as a folding reporter. Formation of the GFP fluorophore is inherently dependent upon proper folding of GFP. If a protein sequence which produces unfolded protein is fused upstream of GFP, then the fusion protein is usually unfolded and non-fluorescent (Fig. 5.3). This property allows screening for mutations in the target sequence that lead to folding and hence fluorescence of the fusion protein. The procedure can be applied iteratively to obtain maximal fluorescence and thus maximal solubility. Indeed, such a technique was used to increase the solubility of GFP itself as expressed in E. coli (Crameri et al., 1996).

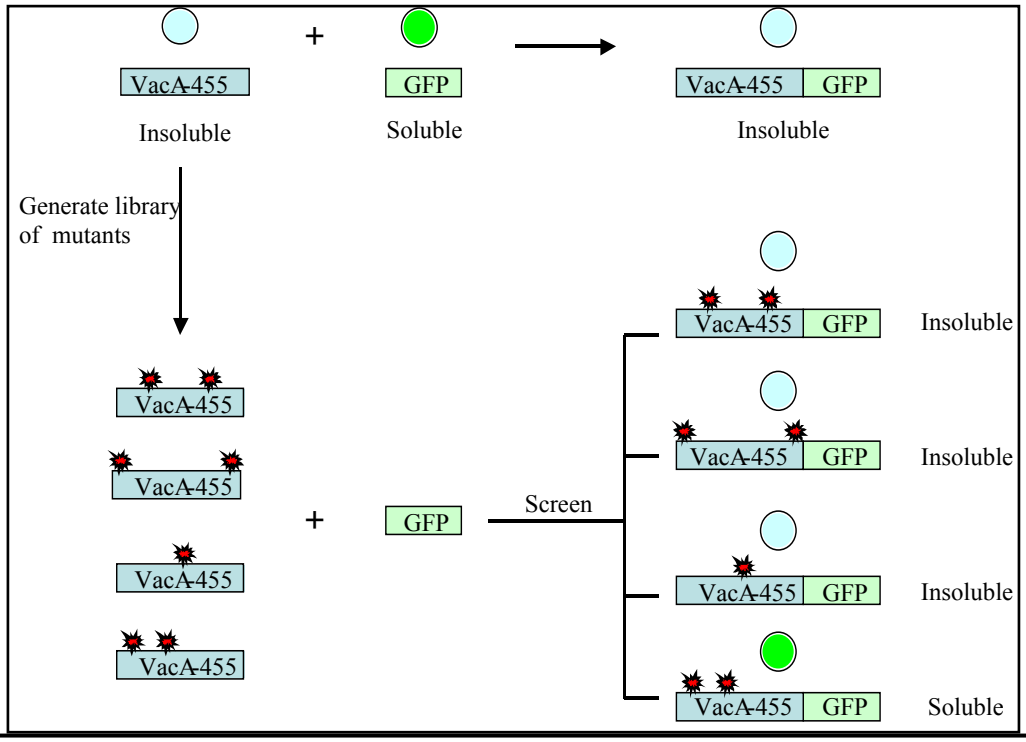


Figure 5.3. Colony screening of soluble VacA/GFP. VacA-455 expressed in *E. coli* produces insoluble protein and non-fluorescent colonies (light blue), while GFP produces soluble protein and fluorescent colonies(bright green). A library of VacA-455 mutants are fused to GFP, and the fluorescence of colonies containing fusion proteins is assessed. Fluorescent colonies contain mutants in VacA-455, which increase the solubility of the protein, and thus allow the VacA-455/GFP to fold and GFP to form its fluorophore.

Experimental Procedures & Results

Directed Evolution of VacA-455

Evolution of the VacA-455 toxin is achieved by DNA shuffling, which introduces point mutations at a frequency of 0.7%, similar to frequencies achieved by error-prone PCR (Stemmer, 1994). In a typical shuffling reaction, the templates is digested using DNaseI to small fragments, shuffled by carrying out PCR in the absence of primers, and reassembled by carrying out PCR using primers. The mutation frequency can be regulated by the metal ion cofactor used during DNaseI digestion and the type of DNA polymerase used in gene reassembly. In these experiments, 2mM Co^{2+} was used as the cofactor during digestion, and Taq polymerase and Pfu polymerase were used in the first and second PCR steps, respectively.

VacA-455 DNA (2 μg) in 50 mM Tris pH 7.5 was digested with DNase I (20 μg/mL) for 15 minutes at 15 ^{o}C and purified using Centrisep columns (Princeton Separations) (Fig 5.4). The purified products were concentrated, and shuffled and reassembled according to the protocol outlined by W. Stemmer, 1994. This final product was digested with XhoI and BamHI, gel purified, and ligated into pUC19/GFP for expression (as described below).

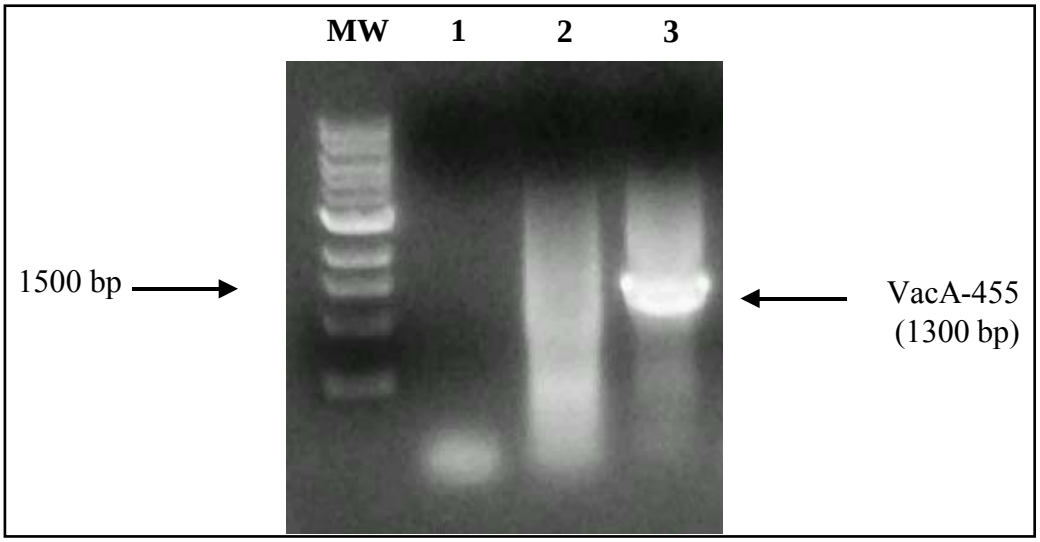


Figure 5.4. DNA Shuffling of VacA. VacA-455 was amplified by polymerase chain reaction, then (1) digested with DNAse I, (2) shuffled with Taq DNA polymerase, and (3) reassembled with Pfu polymerase

Expression Vectors and Fluorescent Colony Screening: 10^4 colonies screened in 3.5 days

In the past, directed evolution protocols using GFP reporters have utilized the pET system for expression in *E. coli* BL21 (DE3) (Kuk Yang et al., 2003; Waldo, 2003; Waldo et al., 1999). This system has several flaws that decrease the efficiency of protein evolution. BL21 (DE3) cells have low transformation efficiencies, and therefore, evolved genes must first undergo transformation into a highly competent *E. coli* strain first (such as Top10). Second, plasmids encoding evolved genes must then be extracted from the colonies and retransformed as supercoiled DNA into BL21

(DE3). This results in losses through manipulation and increases the time required to carry out the procedure. Additionally, transformed BL21 (DE3) must first be grown without induction, as overexpression from the T7 promoter inhibits bacterial growth. This again prolongs the time required for screening. Although this system is capable of identifying increased solubility mutants, I created a reporter system that shortens the time required for screening and increases the library size.

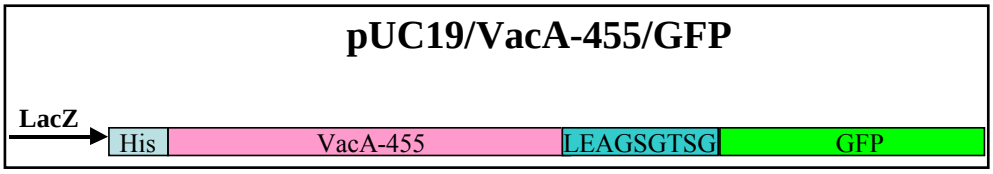


Figure 5.5. pU19/GFP cloning system. VacA-455 is expressed as a fusion with GFP under control of the lac promoter. A short amino acid liker joins the fused protein while a polyhistidine tag precedes VacA-455.

By exploiting the uninduced, leaky expression of GFP from the lac promoter in pUC19, I developed a cloning system that permits high efficiency transformation into electrocompetent Top10 cells (Fig. 5.5). GFP was cloned from pGFPuv (Clontech) into pBAD/HisA (Invitrogen) with the addition of a short N-terminal amino acid linker, LEAGSGTSG. The linker-GFP sequence was removed from pBAD/HisA and cloned into pUC19, such that the poly-histidine tag was included. Thus, pUC19/GFP was constructed to include an N-terminal His-tag followed by restriction sites XhoI/BamHI for VacA ligation, which precedes the linker-GFP. Overnight growth

from transformed colonies results in sufficient low-level expression of fusion proteins to screen for fluorescent colonies using a hand-held UV lamp.

Cloning efficiencies were optimized to obtain 10^6 cfu per μg of insert DNA used in ligation. Transformations are performed using freshly prepared electrocompetent Top 10 *E. coli* cells (Invitrogen) and transformants are spread on large diameter (150 mm) Kirby-Bauer selective media LB plates with 50 μg/mL carbenicillin. Approximately 2,000 colonies are screened per plate when cloned in this manner. The typical yield is 10^4 colonies that are screened per round of mutagenesis, with 3.5 days required from start to final screening.

A number of colonies displaying increased fluorescence were sequenced. Shuffling produced approximately four point mutations per ~1350 nucleotides, approximately half of the 0.7% reported. However, as noted above, the frequency of mutation may be modulated by experimental conditions. Figure 5.6 shows two of the most fluorescent mutants. Sequencing showed that mutant #1 contains two amino acid changes, T50S and W115R and mutant #2 contains three amino acid changes, G59R, Q128R, V285M.

Analysis of Solubility of Mutants #1 and #2

In order to measure the effect these mutations had on solubility, VacA-455 was expressed as a GFP fusion in Top 10 cells overnight. The expression of VacA 455/GFP is too low in pUC19/GFP to be seen clearly by Coomassie staining on SDS-PAGE, and therefore, a Western Blot using anti-GFP antibodies was used to visualize the solubility characteristics of mutants #1 and #2. Colonies expressing evolved,

fluorescent VacA 455/GFP are grown in 30 mL Terrific Broth with 50 μg/mL carbenicillin at 37 ^{o}C, induced using 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG) at an OD600 of 0.4, and grown for a further 12 hours at 25 ^{o}C. The cultures are then pelleted by centrifugation and resuspended in phosphate buffered saline (PBS) containing 0.5 mM lysozyme. The cells are incubated on ice for 15 min and then lysed by sonication. The crude supernatant is separated from the lysed pellet by centrifugation (20 minutes at 30,000 x g) and analyzed by 10% SDS-polyacrylamide gel electrophoresis (Fig 5.7). Proteins resolved by SDS-PAGE are transferred to PVDF membranes for Western blot analysis using a polyclonal GFP antiserum (Clontech).

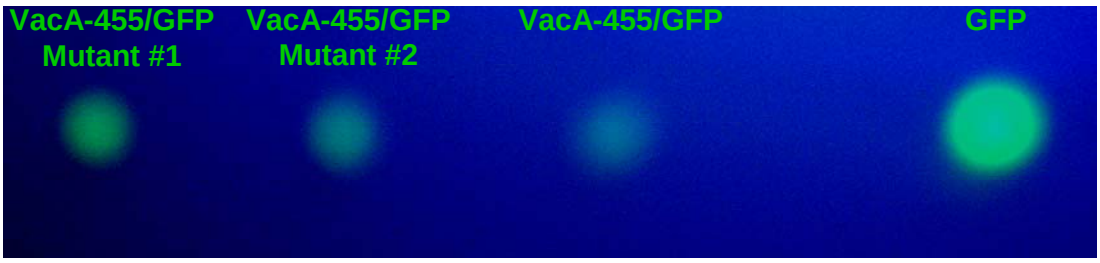


Figure 5.6. Increased fluorescence of evolved VacA-455/GFP. Following mutagenesis, the evolved VacA-455 sequences are inserted into a pUC19/GFP cloning vector and transformed into Top 10 E. coli cells. Following overnight growth, the relative cell fluorescence was compared to cells expressing wild-type VacA-455/GFP fusion protein and GFP alone.

Figure 5.7 shows that the mutations increase the quantity of VacA-455/GFP (72 kDa) localizing to the soluble fraction as compared with wild-type VacA-455/GFP. Clearly, the bulk of VacA-455/GFP even in the two mutants remains in the insoluble fraction, again indicating that further rounds of mutagenesis are warranted to

maximize solubility. Further inspection reveals significant proteolysis of the VacA-455/GFP fusion protein in the insoluble and soluble fractions, with the majority of GFP detected near or below the control GFP marker. The presence of free GFP in the soluble fraction may prove misleading upon detection of fluorescent clones. The resulting cell fluorescence may be a consequence of free GFP released upon proteolysis of VacA.

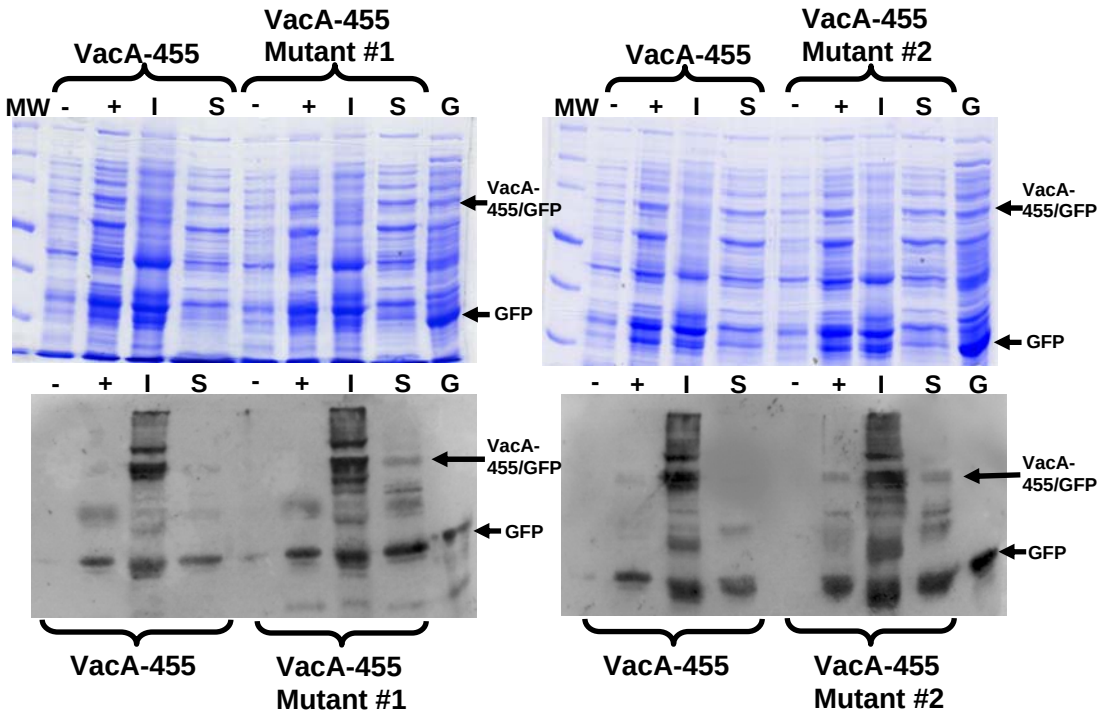


Figure 5.7. SDS-PAGE and Anti-GFP Western Blot of evolved VacA-455. VacA was expressed as a fusion with GFP in pUC19, and products were run on 10% SDS-PAGE, then stained with Coomassie Blue. VacA-455 Mutant #1 & #2 were expressed in parallel for solubility comparison. The lanes are labeled as follows: preinduction (-), postinduction (+), insoluble fraction (after lysis) (I), and soluble fraction (S). Cell lysate with expressed GFP (G) was used as a control for nonspecific binding by the anti-GFP antibody. VacA-455/GFP runs at approximately 72 kDa and GFP at 27 kDa.

Discussion

VacA is an important determinant of pathogenicity in the infectious bacterial pathogen *H. pylori*. Biochemical and structural studies of VacA are essential to understanding its mechanism of action and to devising strategies for designing therapeutics to treat diseases resulting from *H. pylori* infection. Methods to produce soluble, cytotoxic VacA in sufficient quantities for biochemical assays remain elusive. The oligomeric state of VacA likely promotes the tendency for insoluble aggregation of the cytotoxin, however, this oligomeric conformation also appears essential for host membrane channel pore formation (El-Bez et al., 2005). In addition, deletion mapping studies have identified the first 27 amino acids of the N-terminus (VacAΔ6-27), containing the GXXXG motif, as essential for vacuolation and cytotoxicity (McClain et al., 2003; Torres et al., 2006). Low resolution structural studies of VacA/VacAΔ6-27 dominant-negative hybrid protein show the amino terminal residues play an important functional role in the formation of the central pore region of the VacA oligomer. Although neither VacA-455 mutant evolved in this study contained substitutions near the amino terminus, it is also likely that other regions of VacA are important for oligomeric stabilization (Torres et al., 2006).

Experimental evidence suggests it is the acid dissociated VacA monomer that is essential for vacuolation activity (McClain et al., 2000), however, more recent data point toward oligomerization as a structural mainstay not only for host cell entry but for intracellular vacuolation activity (Torres et al., 2006; Willhite et al., 2002). It is

the soluble monomeric form of VacA that we hoped to evolve by directed evolution, however, our observations indicate that improved solubility increases VacA susceptibility to proteolytic degradation. Evolving VacA to soluble monomers may expose regions that are normally protected from proteolysis in its oligomeric form. Our data further magnifies the importance of VacA to self-associate into higher order molecular species.

Directed evolution continues to be a useful tool in the modification and improvement of enzymes and protein design. Recent advances in the GFP folding reporter technology have improved methods for detection of protein solubility while minimizing false positives (Cabantous et al., 2005; Cabantous and Waldo, 2006; Pedelacq et al., 2006). Directed evolution combined with the GFP folding reporter may prove beneficial for enhancing protein solubility, but they do not account for importance of existing protein structure. Clearly, the quaternary structure of VacA may not be beneficial for large scale protein production or high resolution structure determination, yet the oligomeric state serves a specific purpose for cytotoxic functionality.

Bibliography

Arnold, F. H., and Zhao, H. (1997). Optimization of DNA shuffling for high fidelity recombination. Nucleic Acids Research *25*, 1307-1308.

Cabantous, S., Pedelacq, J. D., Mark, B. L., Naranjo, C., Terwilliger, T. C., and Waldo, G. S. (2005). Recent advances in GFP folding reporter and split-GFP solubility reporter technologies. Application to improving the folding and solubility of recalcitrant proteins from Mycobacterium tuberculosis. J Struct Funct Genomics *6*, 113-119.

Cabantous, S., and Waldo, G. S. (2006). In vivo and in vitro protein solubility assays using split GFP. Nat Methods *3*, 845-854.

Crameri, A., Whitehorn, E. A., Tate, E., and Stemmer, W. P. C. (1996). Improved Green Fluorescent Protein by Molecular Evolution Using DNA Shuffling. Nature Biotechnology *14*, 315-319.

El-Bez, C., Adrian, M., Dubochet, J., and Cover, T. L. (2005). High resolution structural analysis of Helicobacter pylori VacA toxin oligomers by cryo-negative staining electron microscopy. J Struct Biol *151*, 215-228.

Fujikawa, A., Shirasaka, D., Yamamoto, S., Ota, H., Yahiro, K., Fukada, M., Shintani, T., Wada, A., Aoyama, N., Hirayama, T.*, et al.* (2003). Mice deficient in protein tyrosine phosphatase receptor type Z are resistant to gastric ulcer induction by VacA of *Helicobacter pylori*. Nature Genetics *33*, 375-381.

Kuk Yang, J., Park, M. S., Waldo, G. S., and Won Suh, S. (2003). Directed evolution approach to a structural genomics project: Rv2002 from *Mycobacterium tuberculosis*. Proceedings of the National Academy of Sciences of the United States of America *100*, 455-460.

McClain, M. S., and Cover, T. L. (2003). Expression of *Helicobacter pylori* Vacuolating Toxin in *Escherichia coli*. Infection and Immunity *71*, 2266-2271.

McClain, M. S., Iwamoto, H., Cao, P., Vinion-Dubiel, A. D., Li, Y., Szabo, G., Shao, Z., and Cover, T. L. (2003). Essential role of a GXXXG motif for membrane channel formation by Helicobacter pylori vacuolating toxin. J Biol Chem *278*.

McClain, M. S., Schraw, W., Ricci, V., Boquet, P., and Cover, T. L. (2000). Acid activation of *Helicobacter pylori* vacuolating cytotoxin (VacA) results in toxin internalization by eukaryotic cells. Molecular Microbiology *37*, 433-442.

Montecucco, C., Papini, E., de Bernard, M., Telford, J. L., and Rappuoli, R. (1999). *Helicobacter pylori* vacuolating cytotoxin and associated pathogenic factors. The Comprehensive Sourcebook of Bacterial Protein Toxins, 264-275.

Pedelacq, J. D., Cabantous, S., Tran, T., Terwilliger, T. C., and Waldo, G. S. (2006). Engineering and characterization of a superfolder green fluorescent protein. Nat Biotechnol *24*, 79-88.

Pedelacq, J. D., Piltch, E., Liong, E. C., Berendzen, J., Kim, C. Y., Rho, B. S., Park, M. S., Terwilliger, T. C., and Waldo, G. S. (2002). Engineering soluble proteins for structural genomics. Nature Biotechnology *20*, 927-932.

Peek Jr, R. M., Blaser, M. J., Mays, D. J., Forsyth, M. H., Cover, T. L., Song, S. Y., Krishna, U., and Pietenpol, J. A. (1999). *Helicobacter pylori* strain-specific genotypes and modulation of the gastric epithelial cell cycle. Cancer Research *59*, 6124-6131.

Seto, K., Hayashi-Kuwabara, Y., Yoneta, T., Suda, H., and Tamaki, H. (1998). Vacuolation induced by cytotoxin from Helicobacter pylori is mediated by the EGF receptor in HeLa cells. Federation of European Biochemical Societies Letters *431*, 347-350.

Smoot, D. T., Resau, J. H., Earlington, M. H., Simpson, M., and Cover, T. L. (1996). Effects of *Helicobacter pylori* vacuolating cytotoxin on primary cultures of human gastric epithelial cells. Gut *39*, 795-799.

Stemmer, W. P. C. (1994). DNA shuffling by random fragmentation and reassembly: *In vitro* recombination for molecular evolution. Proceedings of the National Academy of Sciences of the United States of America *91*, 10747-10751.

Stemmer, W. P. C. (1994b). Rapid evolution of a protein *in vitro* by DNA shuffling. Nature *370*, 389-391.

Suerbaum, S., and Michetti, P. (2002). Helicobacter pylori infection. N Engl J Med *347*, 1175-1186.

Telford, J. L., Ghiara, P., Dell'Orco, M., Comanducci, M., Burroni, D., Bugnoli, M., Tecce, M. F., Censini, S., Covacci, A., Xiang, Z., *et al.* (1994). Gene structure of the *Helicobacter pylori* cytotoxin and evidence of its key role in gastric disease. Journal of Experimental Medicine *179*, 1653-1658.

Torres, V. J., McClain, M. S., and Cover, T. L. (2006). Mapping of a domain required for protein-protein interactions and inhibitory activity of a Helicobacter pylori dominant-negative VacA mutant protein. Infection and Immunity *74*, 2093-2101.

Waldo, G. S. (2003). Genetic screens and directed evolution for protein solubility. Current Opinion in Chemical Biology *7*, 33-38.

Waldo, G. S., Standish, B. M., Berendzen, J., and Terwilliger, T. C. (1999). Rapid protein-folding assay using green fluorescent protein. Nature Biotechnology *17*, 691-695.

Willhite, D. C., Ye, D., and Blanke, S. R. (2002). Fluorescence Resonance Energy Transfer Microscopy of the Helicobacter pylori Vacuolating Cytotoxin within Mammalian Cells. Infection and Immunity *70*, 3824-3832.

Ye, D., and Blanke, S. R. (2000). Mutational Analysis of the *Helicobacter pylori* Vacuolating Toxin Amino Terminus: Identification of Amino Acids Essential for Cellular Vacuolation. Infection and Immunity *68*, 4354-4357.

Ye, D., Willhite, D. C., and Blanke, S. R. (1999). Identification of the Minimal Intracellular Vacuolating Domain of the *Helicobacter pylori* Vacuolating Toxin. The Journal of Biological Chemistry *274*, 9277-9282.

VI.

Appendix B.

Purification of Taq Polymerase

Abstract

Directed evolution requires the use of a low fidelity DNA polymerase isolated from thermophilic bacteria for DNA shuffling and error-prone PCR. Since Pfu DNA polymerase maintains a low error rate of nucleotide-misincorporation due to an inherent proof-reading mechanism, DNA polymerase from *Thermus aquaticus* (Taq) is an ideal choice for error-prone template reassembly. Commercially distributed Taq DNA polymerase (TaqPol) can be costly during the production of large gene libraries produced from DNA shuffling or error-prone PCR; alternatively, the expression and purification of TaqPol is well characterized and can produce high yields of soluble, active protein for mutagenesis assays. Here, I describe the methods for cloning, purification, and activity testing for TaqPol.

Experimental Procedures & Results

Growth of *Thermus aquaticus* and cloning DNA polymerase

Freeze-dried *T. aquaticus* (strain Y-VII-51B, ATCC# 25105) were resuspended in #461 broth (ATCC) containing Castenholz TYE medium, then grown on #461 broth/agar slants at 70 ^{o}C. Once stable growth was established, a 10 mL #461 broth culture was inoculated with a single colony of Taq for overnight growth at 70 ^{o}C. Genomic DNA extraction was performed following the protocol described for yeast DNA preparation (Hoffman and Winston, 1987).

The genomic DNA was used in a polymerase chain reaction (PCR) using Pfu DNA polymerase for amplification with the forward primer (CACGAATTCGGGGATGCTGCCCCTCTTTGAGCCCAAG) and reverse primer (GTGAGATCTATCACTCCTTGGCGGAGAGCCAGTC). The PCR product was cloned into pUC19 by blunt-ended ligation and sequenced to confirm intact DNA polymerase amplification. Sequence analysis of the cloned TaqPol reveals only an 85% amino acid identity to the published TaqPol sequence (Acces #: J04639, gi: 155128) from strain YT-1 (Table 6.1).

TaqPol was excised from pUC19 with EcoRI (from forward primer) and HindIII (from pUC19 MCS) and re-ligated into pET28b and transformed into Top 10 *E. coli* for plasmid DNA propagation. Cloning into pET28b adds a 36 residue N-terminal histine-tag (MGSSHHHHHHSSGLVPRGSHMASMTGGQQMGRDPNS). The Taq/pET28b plasmid was transformed into BL-21 (DE3) *E. coli* cells for overexpression.

Table 6.1. TaqPol protein sequence alignment between strains YT-1 and Y-VII-51B.

```
        10        20        30        40        50        60
YT-1    GMLPLFEPKGRVLLVDGHHLAYRTFHALKGLTTSRGEPVQAVYGFAKSLLKALKEDGDAV
        :::::::::::::::::::.::::::: :::::::::::::::::::::::::::.:.::.:
Y-VII   GMLPLFEPKGRVLLVDGYHLAYRTFFALKGLTTSRGEPVQAVYGFAKSLLKALRENGDVV
                10        20        30        40        50        60

                70        80        90       100       110       120
YT-1    IVVFDAKAPSFRHEAYGGYKAGRAPTPEDFPRQLALIKELVDLLGLARLEVPGYEADDVL
        :::::::::::::..: .:::::::::::::::::::::::.:::::: ::::::.::::::
Y-VII   IVVFDAKAPSFRHQTYEAYKAGRAPTPEDFPRQLALIKEMVDLLGLERLEVPGFEADDVL
                70        80        90       100       110       120

               130       140       150       160       170       180
YT-1    ASLAKKAEKEGYEVRILTADKDLYQLLSDRIHVLHPEGYLITPAWLWEKYGLRPDQWADY
        :.:::::::::::::::::::.:::::::::: .:::::::::: :::::::::.:.::.::
Y-VII   ATLAKKAEKEGYEVRILTADRDLYQLLSDRISILHPEGYLITPEWLWEKYGLKPSQWVDY
               130       140       150       160       170       180

               190       200       210       220        230
YT-1    RALTGDESDNLPGVKGIGEKTARKLLEEWGSLEALLKNLDRLKPA-IREKILAHMDDLKL
        :::.:: :::.:::::::::::: ::..:::::: :::.:...::: .::::::.::.::::
Y-VII   RALAGDPSDNIPGVKGIGEKTAAKLIREWGSLENLLKHLEQVKPASVREKILSHMEDLKL
               190       200       210       220       230       240

     240       250       260       270       280       290
YT-1    SWDLAKVRTDLPLEVDFAKRREPDRERLRAFLERLEFGSLLHEFGLLESPKALEEAPWPP
        : .:..: :::::.::::.::::::: :.:::::::::::::::::::: : :::::::
Y-VII   SLELSRVYTDLPLQVDFARRREPDREGLKAFLERLEFGSLLHEFGLLESPVAAEEAPWPP
               250       260       270       280       290       300

     300       310       320       330       340       350
YT-1    PEGAFVGFVLSRKEPMWADLLALAAARGGRVHRAPEPYKALRDLKEARGLLAKDLSVLAL
        :::::::.:::: ::::::.: :::::  :::.:: .: .::: :.:.:::::::.::::
Y-VII   PEGAFVGYVLSRPEPMWAELNALAAAWEGRVYRAEDPLEALRGLREVRGLLAKDLAVLAL
               310       320       330       340       350       360

     360       370       380       390       400       410
YT-1    REGLGLPPGDDPMLLAYLLDPSNTTPEGVARRYGGEWTEEAGERAALSERLFANLWGRLE
        :::..: :::::::::::::::::.:::::::::::: :::::: ::::::.: :  ::.
Y-VII   REGIALAPGDDPMLLAYLLDPSNTAPEGVARRYGGEWTGEAGERALLSERLYAALLERLK
               370       380       390       400       410       420
```

Table 6.1 TaqPol protein sequence alignment between strains YT-1 and Y-VII-51B, continued.

```
     420       430       440       450       460       470
YT-1   GEERLLWLYREVERPLSAVLAHMEATGVRLDVAYLRALSLEVAEEIARLEAEVFRLAGHP
       ::::::::::.:::.::: ::::::::::::::::::.:::::::  :. ::: :: ::::::
Y-VII  GEERLLWLYEEVEKPLSRVLAHMEATGVRLDVAYLKALSLEVEAELRRLEEEVHRLAGHP
              430       440       450       460       470       480

     480       490       500       510       520       530
YT-1   FNLNSRDQLERVLFDELGLPAIGKTEKTGKRSTSAAVLEALREAHPIVEKILQYRELTKL
       :::::::::::::::::::::::::::::::::::::::::::::::..::::::::.::
Y-VII  FNLNSRDQLERVLFDELGLPAIGKTEKTGKRSTSAAVLEALREAHPIVDRILQYRELAKL
              490       500       510       520       530       540

     540       550       560       570       580       590
YT-1   KSTYIDPLPDLIHPRTGRLHTRFNQTATATGRLSSSDPNLQNIPVRTPLGQRIRRAFIAE
       :.:::::::: :.::.:.:::::::::::::::::::::::::::::: ::::::::::.::
Y-VII  KGTYIDPLPALVHPKTNRLHTRFNQTATATGRLSSSDPNLQNIPVGTPLGQRIRRAFVAE
              550       560       570       580       590       600

     600       610       620       630       640       650
YT-1   EGWLLVALDYSQIELRVLAHLSGDENLIRVFQEGRDIHTETASWMFGVPREAVDPLMRRA
       ::: ::.:::::::::::::::::::::::::::.::::.::::::::: :::: :::::
Y-VII  EGWRLVVLDYSQIELRVLAHLSGDENLIRVFQEGQDIHTQTASWMFGVPPEAVDSLMRRA
              610       620       630       640       650       660

     660       670       680       690       700       710
YT-1   AKTINFGVLYGMSAHRLSQELAIPYEEAQAFIERYFQSFPKVRAWIEKTLEEGRRRGYVE
       :::::::::::::::::: ::::::::: ::::::::::.:::::::::: :::.:::::
Y-VII  AKTINFGVLYGMSAHRLSGELAIPYEEAVAFIERYFQSYPKVRAWIEKTLAEGRERGYVE
              670       680       690       700       710       720

     720       730       740       750       760       770
YT-1   TLFGRRRYVPDLEARVKSVREAAERMAFNMPVQGTAADLMKLAMVKLFPRLEEMGARMLL
       :::::::::::: .:::::.::::::::::::::::::::::::::::.:.::::::
Y-VII  TLFGRRRYVPDLASRVKSIREAAERMAFNMPVQGTAADLMKLAMVKLFPRLQELGARMLL
              730       740       750       760       770       780

     780       790       800       810       820       830
YT-1   QVHDELVLEAPKERAEAVARLAKEVMEGVYPLAVPLEVEVGIGEDWLSAKE
       ::::::::::::.:: ::. ::..:: :.:: :::::::::::::::::
Y-VII  QVHDELVLEVPKEQAEEVAQEAKRTMEEVWPLKVPLEVEVGIGEDWLSAKE
              790       800       810       820       830
```

Purifcation and Activity Assays

Purification of TaqPol was performed using published protocol (Desai and Pfaffle, 1995). Briefly, *E. coli* expressing TaqPol were lysed by the addition of lysozyme and several freeze-thaw cycles. TaqPol was further purified by heat treatment for 1 hour at 70 ^{o}C. The thermostability of TaqPol prevents denaturation at high temperature, and the insoluble protein was centrifuged from the heated solution. Protein expression and purity was assessed by SDS-PAGE (Fig 6.1). The soluble protein that remains after heat treatment was used for PCR activity analysis (Fig. 6.2).

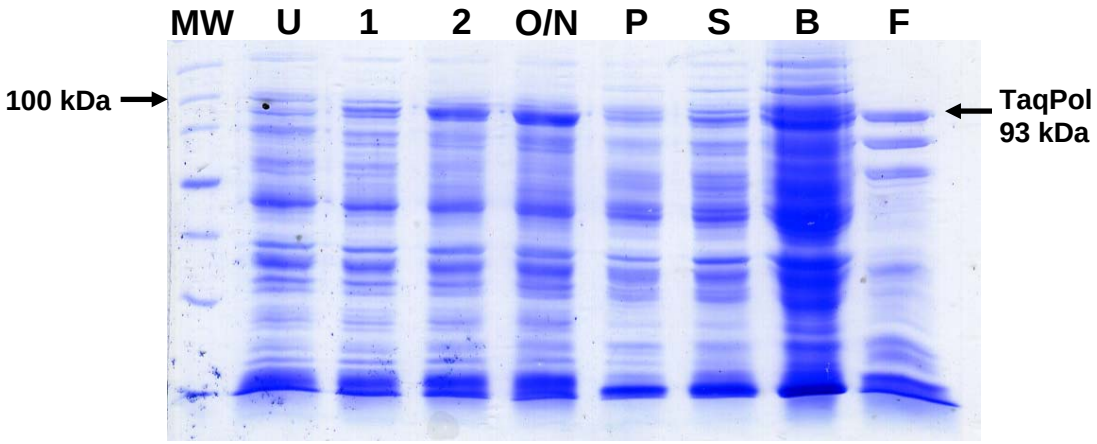


Figure 6.1. TaqPol expression and purification. TaqPol (93 kDa) samples were run on 10% SDS-PAGE and visualized by Coomassie Blue staining. Sample labels are as follows: MW, molecular weight markers; U, uninduced; 1, one hour induction; 2, two hour induction; O/N, overnight induction; P, insoluble pellet after lysis; S, soluble lysate; B, precipitated protein after boiling; F, final product used for activity measurement.

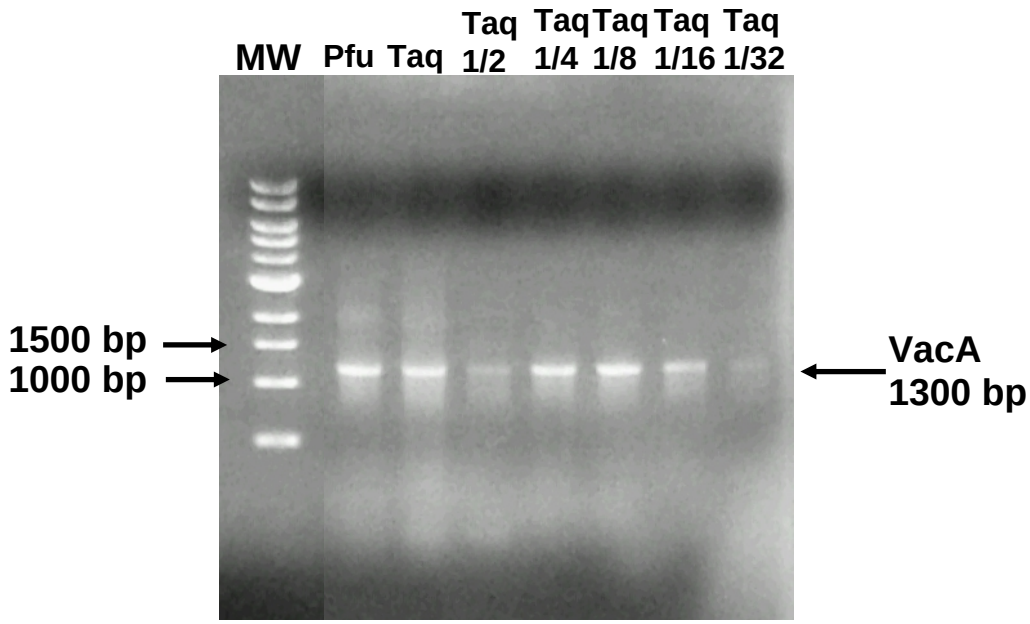


Figure 6.2. TaqPol PCR. pUC19/VacA DNA vector was used as a template for PCR amplification. Sample labels are as follows: MW, molecular weight standards; Pfu, pfu polymerase; Taq, purified taq polymerase undiluted. Lanes 4-8: Taq activity was tested after dilutions of 1/2, 1/4, 1/8, 1/16, 1/32.

Although SDS-PAGE shows protein impurities still present after heat treatment TaqPol remained active up to a 1/16 dilution. Inexplicably, the undiluted purified TaqPol amplified the VacA DNA fragment, but 1/2 diluted TaqPol produced very little DNA amplification. By design, TaqPol could be purified to homogeneity using the His-tag for Ni^{2+} affinity purification; however, initial experiments using this technique resulted in inactive TaqPol. Purification by this technique or other methods should be further tested in order to eliminate remaining DNA or protein contamination.

Bibliography

Desai, U. J., and Pfaffle, P. K. (1995). Single-step purification of a thermostable DNA polymerase expressed in Escherichia coli. Biotechniques *19*, 780-782.

Hoffman, C. S., and Winston, F. (1987). A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of Escherichia coli. Gene *57*, 267-272.