nature structural & molecular biology

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

Stephen A McMahon[1,5,6], Jason L Miller[1,6], Jeffrey A Lawton[1,5], Donald E Kerkow[2,5], Asher Hodes[3], Marc A Marti-Renom[4], Sergei Doulatov[3,5], Eswar Narayanan[4], Andrej Sali[4], Jeff F Miller[3] & Partho Ghosh[1,2]

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 fold. We now add to this the C-type lectin (CLec) fold, as found in the 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* spp. Mtd is an intertwined, pyramid-shaped trimer, with variable residues organized by its CLec fold into discrete receptor-binding sites. The CLec fold provides a highly static scaffold for combinatorial display of variable residues, probably reflecting a different evolutionary solution for balancing diversity against stability from that in the immunoglobulin fold. Mtd variants are biased toward the receptor pertactin, and there is evidence that the CLec fold is used broadly for sequence variation by related retroelements.

The major tropism determinant (Mtd, 40 kDa), the receptor-binding protein of *Bordetella* bacteriophage, varies greatly in sequence[1,2]. Variation in Mtd depends on a phage-encoded retroelement that belongs to a family of retroelements implicated in generating sequence diversity in various phage and bacterial genomes[3]. 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 fold has been the sole paradigm for toleration of massive sequence variation in proteins[4,5]. Lack of similarity of Mtd to other proteins suggests that it represents a previously unobserved 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 an infectious cycle. The 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*[6,7]. 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[8]. 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 using 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. 1a**). The process relies on two nearly identical direct repeats, called the variable region (VR) and template region (TR)[1,2]. In brief, sequence information in the variable region, which encodes the C-terminal 45 residues of Mtd, is replaced by sequence information from the template region, which remains unchanged by this process. The mutagenic aspect derives 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[1]. As a consequence, variability in Mtd is focused to 12 adenine-encoded amino acids that are scattered across its C-terminal variable region (**Fig. 1a**).

RESULTS

Overall structure of Mtd

We determined crystal structures of four Mtd variants, P1, M1, P3c and I1 (**Fig. 1a**), that differ in receptor specificity and host tropism (resolution limits 1.56–2.52 Å, **Supplementary Tables 1** and **2** online). Mtd-P1 confers phage infectivity against *Bordetella* in the pathogenic (Bvg$^+$) but not environmental (Bvg$^-$) phase because it uses the Bvg$^+$-specific surface protein and virulence factor pertactin as a receptor[3,9].

[1]Department of Chemistry & Biochemistry and [2]Section of Molecular Biology, University of California at San Diego, La Jolla, California 92093, USA. [3]Department of Microbiology, Immunology and Molecular Genetics, David Geffen School of Medicine and the Molecular Biology Institute, University of California at Los Angeles, Los Angeles, California 90095, USA. [4]Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, Mission Bay Genentech Hall, University of California at San Francisco, San Francisco, California 94143, USA. [5]Present addresses: Centre for Biomolecular Sciences, University of St. Andrews, Fife KY16 9ST, UK (S.A.M.), Department of Chemistry, Eastern University, 1300 Eagle Road, St. Davids, Pennsylvania 19087, USA (J.A.L.), Scripps Research Institute, 10550 N. Torrey Pines, Mail Stop MB33, La Jolla, California 92037, USA (D.E.K.) and Department of Microbiology and Medical Genetics, University of Toronto, Toronto, Ontario M5G 2C1, Canada (S.D.). [6]These authors contributed equally to this work. Correspondence should be addressed to P.G. (pghosh@ucsd.edu).

Received 29 June; accepted 2 August; published online 18 September 2005; doi:10.1038/nsmb992

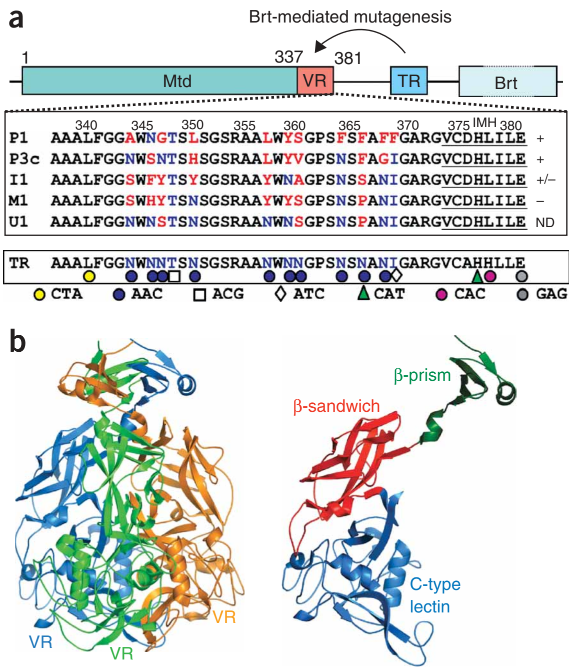

a
Brt-mediated mutagenesis
1
337
381
Mtd
VR
TR
Brt
340 345 350 355 360 365 370 375 IMH 380
P1 AAALFGGAWNGTSLSGSRAALWYSGPSFSFAFFGARGVCDHLILE +
P3c AAALFGGNWSNTSHSGSRAALWYVGPSNSFAGIGARGVCDHLILE +
I1 AAALFGGSWFYTSYSGSRAAYWNAGPSNSSANIGARGVCDHLILE +/–
M1 AAALFGGSWHYTSNSGSRAAYWYSGPSNSPANIGARGVCDHLILE –
U1 AAALFGGNWNSTSNSGSRAANWNSGPSNSPANIGARGVCDHLILE ND
TR AAALFGGNWNNTSNSGSRAANWNNGPSNSNANIGARGVCAHHLLE
● CTA
● AAC
□ ACG
◇ ATC
▲ CAT
● CAC
● GAG
b
β-prism
β-sandwich
C-type lectin
VR
VR
VR

Figure 1 Mtd variation and structure. (**a**) 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 those that are identical in blue). Region corresponding to the initiation of mutagenic homing (IMH) sequence is underlined. Tropism is indicated to the right of sequences: +, Bvg$^+$; –, Bvg$^-$; +/–, either Bvg$^+$ or Bvg$^-$; ND, not determined. Adenine-containing codons in TR are denoted by colored symbols. Adenine-specific mutagenesis of codons (blue or white) results in nonsynonymous substitutions; the three adenine-containing codons at the end of the sequence are part of the IMH element and invariant. (**b**) Left, structure of Mtd trimer (height ~90 Å, base ~50 Å) with protomers colored blue, green and gold and locations of VR indicated. Mtd-P1 is depicted, but other Mtd variants are identical (average r.m.s. deviation 0.33 Å, 376 Cα atoms). Right, Mtd protomer. Green, β-prism domain; red, β-sandwich; blue, C-type lectin. Molecular graphics were made with PyMOL (http://pymol.sourceforge.net).

In contrast, Mtd-M1 confers infectivity against *Bordetella* in the Bvg$^-$ phase only, indicating that it binds an unknown minus-specific receptor, possibly glycosyl groups of the O-antigen as has been suggested by indirect evidence (A.H. and J.F.M., unpublished data). Mtd-P3c confers infectivity against the plus phase only but uses an unknown receptor other than pertactin, and Mtd-I1 confers infectivity indiscriminately against both plus and minus phases, 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 that no large conformational changes result from sequence variation. Mtd is an intertwined, pyramid-shaped trimer (**Fig. 1b**), corresponding in size and shape to knobs seen at the ends of phage tail fibers[2]. A largely hydrophobic interface buries >4,500 Å^2 of surface area in each protomer, consistent with obligatory trimerization (**Supplementary Data** online). Hydrogen bonds and a shared cation are also involved in trimerization. Mtd protomers are composed of N-terminal, intermediate and C-terminal domains (**Fig. 1b** and **Supplementary Fig. 1** online). At the apex of the pyramid, trimerization of the N-terminal domain (residues 1–48) of Mtd forms a three-fold symmetric β-prism. This resembles the pseudo three-fold–symmetric β-prisms of monocot lectins (r.m.s. deviation 2.4 Å, 60 Cα atoms), but lacks residues in these lectins identified as binding carbohydrates (**Supplementary Data**)[10,11]. The intermediate domain forms a β-sandwich containing three- and four-stranded antiparallel sheets with a nearly right-angle turn in the middle, and it has a novel fold (**Supplementary Data**). 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 the 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 do (**Fig. 1b**).

C-type lectin fold in the variable domain

The C-terminal domain has a C-type lectin (CLec) fold[12], as identified by structural homology searches[13] (**Fig. 2a**). The CLec fold of Mtd is related to those of divergently and convergently evolved CLec proteins, such as macrophage mannose receptor (r.m.s. deviation 2.7 Å, 101 Cα, $Z = 6.5$)[14] and intimin (r.m.s. deviation 3.0 Å, 89 Cα, $Z = 5.8$)[15,16], respectively. Notably, the CLec fold is not functionally restricted to calcium-dependent carbohydrate binding but instead constitutes a general ligand (including protein)-binding motif[17].

The typical topological features of the ~110–130-residue CLec fold, 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. 2a**). 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; it has different core residues than do divergently evolved CLec proteins and lacks residues required for calcium- or carbohydrate-binding[12]. Likewise, none of the four disulfide bond–forming cysteines seen 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).

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. 2b**), indicative of 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[18] and intimin[15,16].

All but two of the variable residues in the template region are encoded by AAC (**Fig. 1a**). Adenine-specific mutagenesis of AAC-encoded asparagine permits substitution by 14 other amino acids covering the gamut of chemical character. For example, tryptophan cannot be encoded, but phenylalanine and tyrosine can, and likewise glutamic acid and lysine cannot be encoded, but aspartic acid and arginine (and also histidine) can. Notably, use of the AAC codon precludes the introduction of a nonsense codon. Residues 348 and 369 are encoded by ACG (threonine) and ATC (isoleucine), respectively, in the template region, and adenine-specific mutagenesis of these permits substitution by three other amino acids (serine, proline and alanine for 348; valine, leucine and phenylalanine for 369). There does not seem to be a structural need for residue 348 to be small, but 369 may need to be hydrophobic to pack between the invariant residues Trp307 and Trp309.

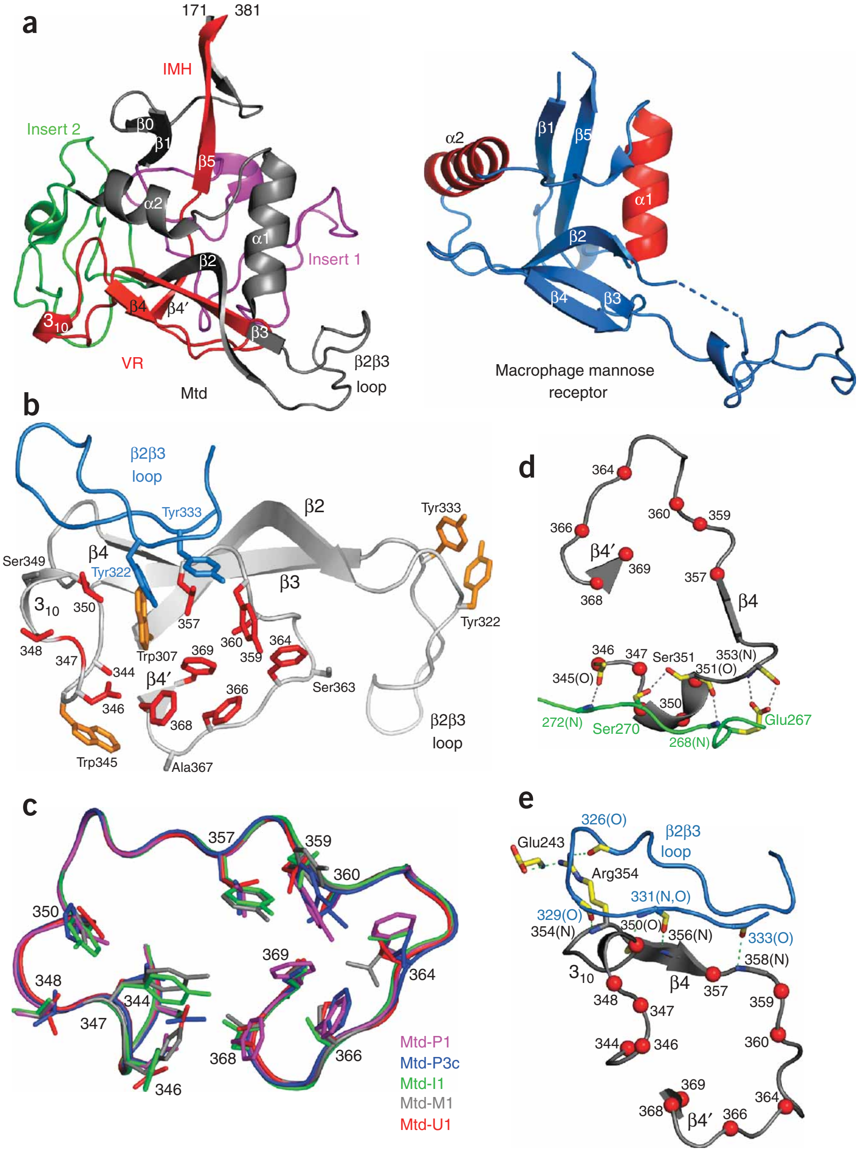


Figure 2 Variable region (**a**) Left, C-type lectin domain of Mtd. Red, variable region (VR); pink, insert 1 (residues 200–236, between β1 and α1); green, insert 2 (residues 264–305, between α2 and β2); gray, other regions. Secondary structure elements are labeled, including short 3_{10}-helix in VR. 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 (residues 639–763) of mannose macrophage receptor. Red, α-helices; blue, β-strands and loops. (**b**) Receptor-binding site (residues 307–369) of Mtd. Gray, main chain and nonaromatic invariant side chains; orange, invariant aromatic side chains; red, variable side chains; blue, β2β3 loop from a neighboring protomer and its invariant Tyr322 and Tyr333. (**c**) Superposition of variable regions of Mtd variants. (**d**) Stabilization of VR by insert 2. Main chain and side chains (Glu267 and Ser270) of insert 2 (green) form hydrogen bonds to main chain and side chains (Ser351 and Ser353) of VR (gray). Red spheres, Cα positions of variable residues; dashed lines, hydrogen bonds; yellow atoms, carbons; red atoms, oxygens. Main chain atoms are indicated in parentheses. (**e**) 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) of VR. Coloration is as in **d**.

The binding site also contains two invariant, solvent-exposed aromatic residues, Trp307 and Trp345, which are likely to partake in receptor interactions (**Fig. 2b**). Two additional aromatic residues, Tyr322 and Tyr333, are contributed by the β2β3 loop from a neighboring protomer that intertwines into the binding site (**Fig. 2b**). The β2β3 loop varies greatly in structure among CLec proteins and is responsible for calcium-dependent carbohydrate binding among a subset of them[19]. The 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 at the 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[3]. It is therefore invariant despite containing adenine-encoded residues and is also noteworthy in containing codons that differ between the variable and template regions at bases other than adenine (those encoding residues 376–379). Invariance of IMH at the nucleotide level is reflected at the protein level, with β5 positioned in the central core of trimer, making close intra- and intermolecular 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 despite large differences in sequence (**Fig. 2c**, r.m.s. deviation 0.27 Å, 45 Cα). This is made more notable by the fact that more than half the variable residues (344, 346, 347, 359, 360, 364 and 366) are located on loops (**Fig. 2b**). Providing stabilization to these loops are the two inserts in the CLec domain along with the trimeric assembly. The inserts form hydrogen bonds to the main chain and invariant side chains (such as Ser351, 353 and 365) of the variable region (**Fig. 2d**). 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. 2e**). The β2β3 loop has the same intertwining conformation in all Mtd variants examined, being positioned exclusively over invariant residues (351–356) (**Fig. 2b,c**). 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. Mtd-P1 and Mtd-I1 have highly hydrophobic binding sites, and the continuity of the hydrophobic surface decreases successively for Mtd-P3c, Mtd-M1 and Mtd-U1 (**Fig. 3**), with Mtd-U1 having nine template region–encoded mostly hydrophilic residues (**Fig. 1a**). The binding sites of Mtd-P1 and Mtd-I1 accommodate four or five large, exposed hydrophobic residues, and although a preponderance of exposed hydrophobic surface is typically

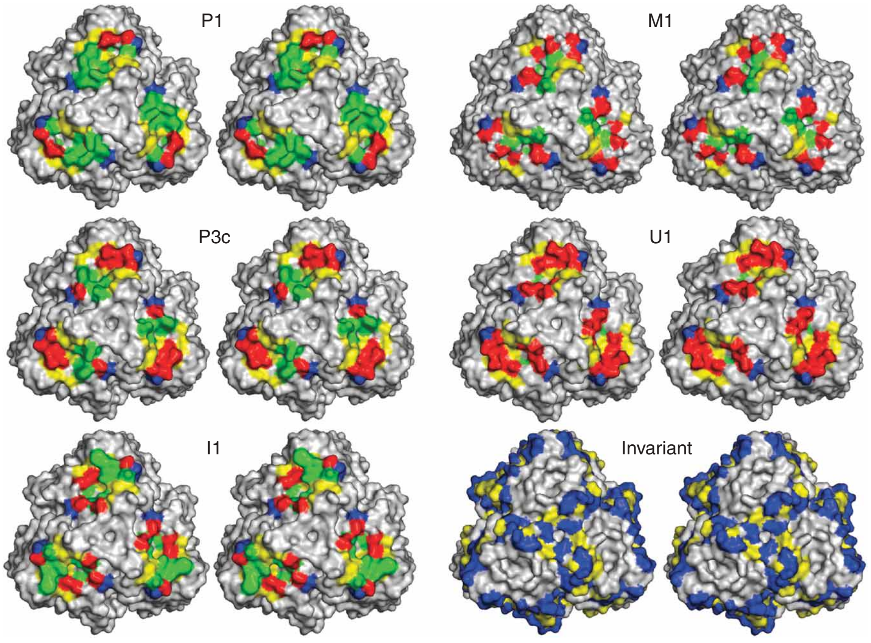


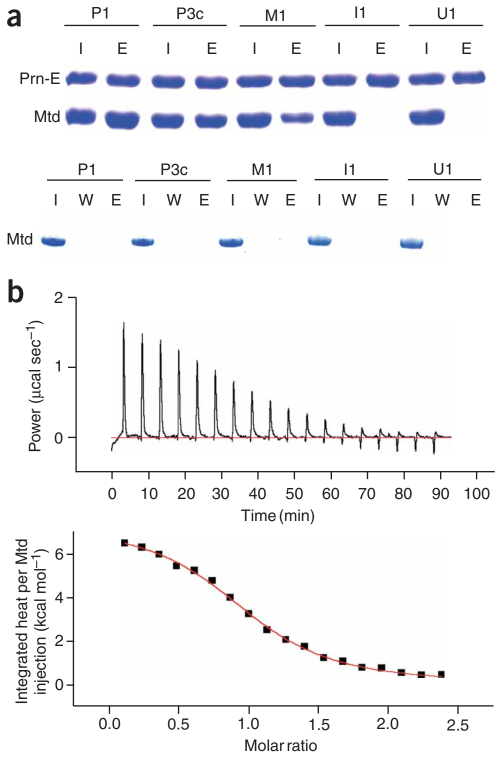
Figure 3 Receptor-binding site. Molecular surface in stereo of receptor-binding sites of Mtd variants. Residue coloration: green, variable hydrophobic (Ala, Val, Leu, Ile, Phe, Tyr, Trp and Met); yellow, invariant hydrophobic; red, variable hydrophilic (Ser, Thr, Asn, Gln, Asp, Glu, His, Lys, Arg and Cys); blue, invariant hydrophilic. 'Invariant' shows surface surrounding the receptor-binding site, with binding site residues uncolored and all others colored as above.

correlated with protein instability[20], both Mtd-P1 and Mtd-I1 are highly stable proteins. Protein stability is probably aided by the hydrophilic surface formed by invariant residues surrounding the binding site.

Pertactin interactions

Direct association between Mtd-P1 and its *Bordetella* receptor pertactin[3] 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 Mtd-U1 (**Fig. 4a**). Notably, pertactin also associated with Mtd-P3c and, weakly, with Mtd-M1, both direct descendants of Mtd-P1 (ref. 1). Tyr359 is the only variable residue common to these three pertactin-binding Mtd variants (**Fig. 1a**) and is also highly conserved among plus-tropic phages[1]. Although an explanation for the importance of Tyr359 requires further investigation, our results suggest that this residue is a dominant pertactin-binding determinant. The existence of pertactin cross-reactivity, conferred by the single residue Tyr359, along with the observation that the variable region is usually replaced in short patches rather than *en bloc*[3] explain why phage tropism is vastly skewed toward the plus phase. This skew is evident in the fact that plus-tropic phages convert to minus-tropic or indiscriminate (infecting both plus and minus *Bordetella)* phages at a frequency of 10^{-6}, in contrast to the thousand-fold greater frequency of 10^{-3} at which minus-tropic and indiscriminate phages convert to plus-tropic phages[1]. It is possible that this skew affects *Bordetella* variation between its pathogenic (plus) and environmental (minus) phases.

Potentially trivalent Mtd-P1 was found by static light scattering to associate with only a single pertactin molecule (**Supplementary Fig. 2** online). This suggests that pertactin sterically occludes other binding sites, associates with Mtd pseudo symmetrically 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 provided further evidence for 3:1 Mtd/pertactin stoichiometry (**Fig. 4b**). As *Bordetella* bacteriophage seems to have



Temperature (°C)	K_d (μM)	ΔH (kcal mol^{-1})	N	ΔG (kcal mol^{-1})	$T\Delta S$ (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 4 Pertactin association. (**a**) Top, coprecipitations showing association of Mtd variants with the His-tagged ectodomain of pertactin (Prn-E), visualized using SDS-PAGE and Coomassie staining. I, relative amount of each variant incubated; E, relative amount eluted. Bottom, coprecipitation of Mtd variants in the absence of Prn-E, visualized using SDS-PAGE and Coomassie staining. W, relative amount of each variant upon the last of three washes. (**b**) Isothermal titration calorimetry of binding between Mtd-P1 and Prn-E. Top, measurements of the endothermic reaction resulting from successive 15-μl injections of Mtd-P1 into a solution of Prn-E. Middle, integrated heats of injection plotted against molar ratio of Mtd-P1 trimer to Prn-E. Bottom, summary of data from experiments carried out at 15.5 and 23 °C, with free energy and entropy changes calculated using the relationships $\Delta G = -RT\ln(1/K_d)$ and $\Delta G = \Delta H - T\Delta S$. N is the number of Mtd trimers calculated to bind Prn-E.

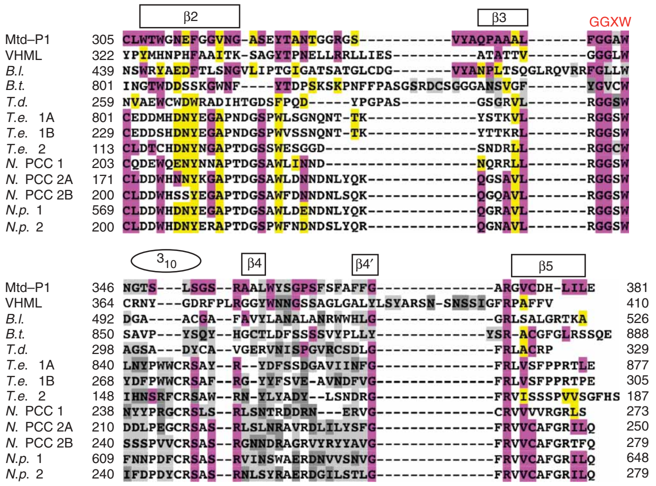

β2
β3
GGXW
Mtd–P1 305 CLWTWGNEFGGVNG-ASEYTANTGGRGS------------VYAQPAAAL-------FGGAW
VHML 322 YPYMHNPHFAAITK-SAGYTPNELLRRLLIES----------ATATTV-------GGGLW
B.l. 439 NSWRYAEDFTLSNGVLIPTGIGATSATGLCDG--------VYANPLTSQGLRQVRRFGLLW
B.t. 801 INGTWDDSSKGWNF----YTDPSKSKPNFFPASGSRDCSGGGANSVGF-------YGVCW
T.d. 259 NVAEWCWDWRADIHTGDSFPQD-----YPGPAS---------GSGRVL-------RGGSW
T.e. 1A 801 CEDDMHDNYEGAPNDGSPWLSGNQNT-TK--------------YSTKVL-------RGGSW
T.e. 1B 229 CEDDSHDNYEGAPNDGSPWVSSNQNT-TK--------------YTTKRL-------RGGSW
T.e. 2 113 CLDTCHDNYNGAPTDGSSWESGGD-----------------SNDRLL-------RGGCW
N. PCC 1 203 CQDEWQENYNNAPTDGSAWLINND-----------------NQRRLL-------RGGSW
N. PCC 2A 171 CLDDWHNNYKGAPTDGSAWLDNNDNLYQK-------------QGSAVL-------RGGSW
N. PCC 2B 200 CLDDWHSSYEGAPTDGSAWFDNNDNLSQK-------------QGQAVL-------RGGSW
N.p. 1 569 CLDDWHDNYEGAPTDGSAWLDENDNLYQK-------------QGRAVL-------RGGSW
N.p. 2 200 CLDDWHDNYERAPTDGSPWFNDNDSLYQR-------------QGNAVL-------RGGSW
3_{10}
β4
β4′
β5
Mtd–P1 346 NGTS---LSGS--RAALWYSGPSFSFAFFG---------------ARGVCDH-LILE 381
VHML 364 CRNY---GDRFPLRGGYWNNGSSAGLGALYLSYARSN-SNSSIGFRPAFFV 410
B.l. 492 DGA----ACGA--FAVYLANALANRWWHLG---------------GRLSALGRTKA 526
B.t. 850 SAVP---YSQY--HGCTLDFSSSSVYPLLY--------------YSR-ACGFGLRSSQE 888
T.d. 298 AGSA---DYCA--VGERVNISPGVRCSDLG---------------FRLACRP 329
T.e. 1A 840 LNYPWWCRSAY--R--YDFSSDGAVIINFG---------------FRLVSFPPRTLE 877
T.e. 1B 268 YDFPWWCRSAF--RG-YYFSVE-AVNDFVG---------------FRLVSFPPRTPE 305
T.e. 2 148 IHNSRFCRSAW--RN-YLYADY--LSNDRG---------------FRVISSSPVVSGFHS 187
N. PCC 1 238 NYYPRGCRSLS--RLSNTRDDRN---ERVG---------------CRVVVVRGRLS 273
N. PCC 2A 210 DDLPEGCRSAS--RLSLNRAVRDLILYSFG---------------FRVVCAFGRILQ 250
N. PCC 2B 240 SSSPVVCRSAS--RGNNDRAGRVYRYYAVG---------------FRVVCAFGRTFQ 279
N.p. 1 609 FNNPDFCRSAS--RVINSWAERDNVVSNVG---------------FRVVCAFGRILQ 648
N.p. 2 240 IFDPDYCRSAS--RNLSYRAERDGILSTLG---------------FRVVCAFGRILQ 279

Figure 5 Structure-based sequence alignment of Mtd with potentially variable proteins of related retroelements. Pink, identical residues; yellow, chemically conserved residues (grouped as follows: Trp, Phe, Tyr; Ala, Val, Leu, Ile; Ser, Thr; Asn, Gln, Asp, Glu; Arg, Lys); light gray, variable residues in Mtd and residues that differ between VR and TR in genomic sequences of potentially variable proteins; dark gray, 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, β-strand; oval, 3_{10}- helix) for Mtd and the GGXW motif are denoted above the alignment. 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. Nine of these sequences are identified as belonging to the DUF-323 family (**Supplementary Methods**), suggesting that DUF-323 proteins are likely to have C-type lectin folds.

12 Mtd trimers in total[1,2], 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 ≥40 μM can be estimated (see Methods), distinguishing between infectious (3 μM) and noninfectious (≥40 μM) interactions.

Variable proteins of other retroelements

Nine retroelements having reverse transcriptases similar to the one in the *Bordetella* bacteriophage retroelement, and similar adenine-specific differences between the template and variable regions, have been identified in phage and bacterial genomes[3]. 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. 5**). In particular, the invariant Mtd binding site residue Trp345 is present in a highly conserved GGXW motif. Invariant residues (Ser351, Ser353 and Arg354) involved in loop stabilization, trimeric contacts or both (**Fig. 2d**,**e**) are also generally conserved. As in Mtd, residues differing between variable and template regions, or 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 show statistically significant matches (*e*-values $< 10^{-8}$) between the C terminus of Mtd and C termini of potentially variable proteins from *Treponema denticola*, *Vibrio harveyi* ML phage, two *Trichodesmium* species and two *Nostoc* species (**Supplementary Methods** online). An additional piece of evidence is that variable regions and IMH elements of related retroelements are consistently located at the extreme C terminus of potentially variable proteins[3]. This property is not necessitated by genetic mechanisms of variation, but it is consistent with features of the CLec fold of Mtd (**Fig. 2**). Together, these results provide evidence that the CLec fold is used broadly as a scaffold for sequence variation by related retroelements.

In contrast to the CLec domain, neither the β-prism nor the β-sandwich domain of Mtd is conserved in proteins of related retroelements. Indeed, some of these putatively variable proteins (such as *Trichodesmium erythraeum* 2) are quite short and seem to be composed of only the CLec domain, whereas others (such as *Bacteroides thetaiotaomicron* and *Trichodesmium erythraeum* 1A) are quite large 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 *Bifidobacterium 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 folds may have 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 of, for example, an antibody binding 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. Notably, pertactin itself is antigenically variant[21], requiring Mtd to keep pace with pertactin variation driven by selection pressure from antibodies.

Although predator proteins are extremely rare, a fairly large number of antigenically variable prey proteins are known, with well-characterized examples being trypanosomal variable-surface glycoproteins[22] and gonococcal pilins[23]. The selection pressure for diversity in antigenically variable proteins is not nearly as great as for predator proteins. For example, the number of possible variable-surface glycoprotein sequences[24] is estimated to be $\sim 10^3$, considerably lower than the $\sim 10^{14}$–10^{16} possible sequences of immune system proteins[5] and $\sim 10^{13}$ possible Mtd sequences[1]. 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, whereas the binding site of a predatory protein must accommodate almost any sequence in order to have sufficient diversity to bind almost any ligand. This latter demand has been met successfully by the immunoglobulin 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[25]. Our work provides the first evidence that evolution has made use not only of the immunoglobulin fold but also of 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 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 immunoglobulin fold being continuous in primary sequence rather than dispersed as in the CLec fold. It is reasonable to expect that the immunoglobulin fold provides greater binding diversity because of its combined sequence and conformational variations. However, there seems to be a cost for greater diversity, as certain conformations of the immunoglobulin fold are deleterious to protein stability and promote fibril formation, as in the case of light chain amyloidosis[26–29]. It seems likely that the CLec fold, although more limited in diversity, is 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 immunoglobulin folds.

METHODS

Expression and purification. Coding sequences of Mtd (1–381) variants and Prn-E (38–640) were amplified by PCR from *B. bronchiseptica* RB50 or phage lysates. Mtd variants were expressed in *Escherichia coli* BL21 (DE3) as glutathione *S*-transferase fusion or His-tagged proteins and purified by standard procedures, including proteolytic removal of fusions or tags. Prn-E was expressed similarly, with an N-terminal His-tag, and refolded and purified from inclusion bodies[30] without removal of the tag (**Supplementary Methods**).

Structure determination. Crystals of Mtd variants were grown by vapor diffusion at 25 °C by mixing equal volumes of protein with various precipitants (**Supplementary Methods**). For phase determination by MAD, methionine substitutions were introduced in Mtd-P1 at Leu52 and Val53 via QuikChange (Stratagene), and selenomethionine 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 ∼110 K from cryoprotected crystals (19-ID, Advanced Photon Source) and processed and scaled using HKL2000 (ref. 31) (**Supplementary Tables 1** and **2**). Positions of three selenomethionines (Met100, Met182 and Met245) and initial phases were determined using SOLVE[32], and solvent flattening was performed using DM[33]. Automated model building with ARP/wARP[34] produced a trace of residues 5–380. This model was used for molecular replacement with Amore[33] of native Mtd-P1, Mtd-P3c, Mtd-I1, Mtd-M1 and Mtd-U1. O[35] was used for manual model building and REFMAC5[33] for refinement of 95% of data (the remaining 5% were used for a 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 were within electron density in $2F_o$–F_c maps calculated with model phases, except for His346 in Mtd-M1 and Phe346 in Mtd-I1, for which density was seen only to the Cβ atom.

Coprecipitation assay. Prn-E (80 μM, His tagged) was incubated (5 min, 25 °C) with Mtd variants (42 μM trimer) in binding buffer (150 mM NaCl and 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 devoid of protein.

Isothermal titration calorimetry. ITC measurements were performed at 15 and 23 °C and data were analyzed with Origin (MicroCal MCS). 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 and 50 mM Tris (pH 8.0). Estimation of the lower bound for the 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 ($c = M_{tot}/K_d$, where M_{tot} is the concentration of Prn-E in the cell)[36] for Mtd-M1 is likely to be <1, as values of $1 < c < 500$ yield interpretable binding isotherms. Therefore, the binding affinity can be estimated to have a $K_d \geq 40$ 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.

Accession codes. Protein Data Bank: Coordinates have been deposited with the accession codes 1YU0, 1YU1, 1YU2, 1YU3 and 1YU4. BIND identifiers (http://bind.ca): 330939, 330940.

Note: Supplementary information is available on the Nature Structural & Molecular Biology website.

ACKNOWLEDGMENTS
This work was supported by a W.M. Keck Distinguished Young Scholars in Medicine Award to P.G., a University of California Discovery Grant to J.F.M. and US National Institutes of Health grants T32GM008326 and F31AI061840 to J.L.M. and F32AI49695 to J.A.L. Discussions with F. Stevens on antibody stability are appreciated, as is help with ITC from S. Bergqvist.

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/nsmb/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Liu, M. *et al.* Reverse transcriptase-mediated tropism switching in *Bordetella* bacteriophage. *Science* **295**, 2091–2094 (2002).
2. Liu, M. *et al.* Genomic and genetic analysis of *Bordetella* bacteriophages encoding reverse transcriptase-mediated tropism-switching cassettes. *J. Bacteriol.* **186**, 1503–1517 (2004).
3. Doulatov, S. *et al.* Tropism switching in *Bordetella* bacteriophage defines a family of diversity-generating retroelements. *Nature* **431**, 476–481 (2004).
4. Chothia, C. & Lesk, A.M. Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196**, 901–917 (1987).
5. Davis, M.M. & Bjorkman, P.J. T-cell antigen receptor genes and T-cell recognition. *Nature* **334**, 395–402 (1988).
6. Uhl, M.A. & Miller, J.F. Integration of multiple domains in a two-component sensor protein: the *Bordetella pertussis* BvgAS phosphorelay. *EMBO J.* **15**, 1028–1036 (1996).
7. Akerley, B.J., Cotter, P.A. & Miller, J.F. Ectopic expression of the flagellar regulon alters development of the *Bordetella*-host interaction. *Cell* **80**, 611–620 (1995).
8. Cotter, P.A. & Miller, J.F. 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 (1997).
9. Emsley, P., Charles, I.G., Fairweather, N.F. & Isaacs, N.W. Structure of *Bordetella pertussis* virulence factor P.69 pertactin. *Nature* **381**, 90–92 (1996).
10. Hester, G., Kaku, H., Goldstein, I.J. & Wright, C.S. Structure of mannose-specific snowdrop (*Galanthus nivalis*) lectin is representative of a new plant lectin family. *Nat. Struct. Biol.* **2**, 472–479 (1995).
11. Sauerborn, M.K., Wright, L.M., Reynolds, C.D., Grossmann, J.G. & Rizkallah, P.J. Insights into carbohydrate recognition by *Narcissus pseudonarcissus* lectin: the crystal structure at 2 A resolution in complex with alpha1–3 mannobiose. *J. Mol. Biol.* **290**, 185–199 (1999).
12. Weis, W.I., Kahn, R., Fourme, R., Drickamer, K. & Hendrickson, W.A. Structure of the calcium-dependent lectin domain from a rat mannose-binding protein determined by MAD phasing. *Science* **254**, 1608–1615 (1991).
13. Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138 (1993).
14. Feinberg, H. *et al.* Structure of a C-type carbohydrate recognition domain from the macrophage mannose receptor. *J. Biol. Chem.* **275**, 21539–21548 (2000).
15. Batchelor, M. *et al.* Structural basis for recognition of the translocated intimin receptor (Tir) by intimin from enteropathogenic *Escherichia coli. EMBO J.* **19**, 2452–2464 (2000).
16. Luo, Y. *et al.* Crystal structure of enteropathogenic *Escherichia coli* intimin-receptor complex. *Nature* **405**, 1073–1077 (2000).
17. Drickamer, K. C-type lectin-like domains. *Curr. Opin. Struct. Biol.* **9**, 585–590 (1999).
18. Tormo, J., Natarajan, K., Margulies, D.H. & Mariuzza, R.A. Crystal structure of a lectin-like natural killer cell receptor bound to its MHC class I ligand. *Nature* **402**, 623–631 (1999).
19. Weis, W.I., Drickamer, K. & Hendrickson, W.A. Structure of a C-type mannose-binding protein complexed with an oligosaccharide. *Nature* **360**, 127–134 (1992).
20. Takano, K., Yamagata, Y. & Yutani, K. 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 (1998).
21. Mooi, F.R. *et al.* 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 (1998).
22. Blum, M.L. *et al.* A structural motif in the variant surface glycoproteins of *Trypanosoma brucei. Nature* **362**, 603–609 (1993).
23. Parge, H.E. *et al.* Structure of the fibre-forming protein pilin at 2.6 A resolution. *Nature* **378**, 32–38 (1995).

24. Van der Ploeg, L.H. *et al.* 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 (1982).
25. Pancer, Z. *et al.* Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. *Nature* **430**, 174–180 (2004).
26. Bellotti, V., Mangione, P. & Merlini, G. Review: immunoglobulin light chain amyloidosis–the archetype of structural and pathogenic variability. *J. Struct. Biol.* **130**, 280–289 (2000).
27. Stevens, F.J. Four structural risk factors identify most fibril-forming kappa light chains. *Amyloid* **7**, 200–211 (2000).
28. Pokkuluri, P.R. *et al.* Increasing protein stability by polar surface residues: domain-wide consequences of interactions within a loop. *Biophys. J.* **82**, 391–398 (2002).
29. Helms, L.R. & Wetzel, R. Destabilizing loop swaps in the CDRs of an immunoglobulin VL domain. *Protein Sci.* **4**, 2073–2081 (1995).
30. Emsley, P., McDermott, G., Charles, I.G., Fairweather, N.F. & Isaacs, N.W. Crystallographic characterization of pertactin, a membrane-associated protein from *Bordetella pertussis. J. Mol. Biol.* **235**, 772–773 (1994).
31. Otwinowski, Z. & Minor, W. Processing of x-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
32. Terwilliger, T.C. 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 (1997).
33. Winn, M.D. An overview of the CCP4 project in protein crystallography: an example of a collaborative project. *J. Synchrotron Radiat.* **10**, 23–25 (2003).
34. Perrakis, A. wARP: improvement and extension of crystallographic phases by weighted averaging of multiple-refined dummy atomic models. *Acta Crystallogr. D Biol. Crystallogr.* **53**, 448–455 (1997).
35. Jones, T.A., Zou, J.-Y., Cowan, S.W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
36. Wiseman, T., Williston, S., Brandts, J.F. & Lin, L.N. Rapid measurement of binding constants and heats of binding using a new titration calorimeter. *Anal. Biochem.* **179**, 131–137 (1989).