Title

Crystal structure of a Thermus aquaticus diversity-generating retroelement variable protein

Permalink

https://escholarship.org/uc/item/3ks666c0

Authors

Handa, Sumit
Shaw, Kharissa L
Ghosh, Partho

Publication Date

2018-10-01

DOI

10.1101/432187

Peer reviewed

Crystal structure of a *Thermus aquaticus* diversity-generating retroelement variable protein

Sumit Handa, Kharissa L Shaw, Partho Ghosh*

Department of Chemistry & Biochemistry, University of California San Diego, La Jolla, CA 92093, USA

*Corresponding author

Epmail: pghosh@ucsd.edu (PG)

Abstract

Diversity-generating retroelements (DGRs) are widely distributed in bacteria, archaea, and microbial viruses, and bring about unparalleled levels of sequence variation in target proteins. While DGR variable proteins share low sequence identity, the structures of several such proteins have revealed the C-type lectin (CLec)-fold as a conserved scaffold for accommodating massive sequence variation. This conservation has led to the suggestion that the CLec-fold may be useful in molecular surface display applications. Thermostability is an attractive feature in such applications, and thus we studied the variable protein of a DGR encoded by a prophage of the thermophile *Thermus aquaticus.* We report here the 2.8 Å resolution crystal structure of the variable protein from the *T. aquaticus* DGR, called TaqVP, and confirm that it has a CLec-fold. Remarkably, its variable region is nearly identical in structure to those of several other CLec-fold DGR variable proteins despite low sequence identity among these. TaqVP was found to be thermostable, which appears to be a property shared by several CLec-fold DGR variable proteins. These results provide impetus for the pursuit of the DGR variable protein CLec-fold in molecular display applications.

Introduction

Diversity generating-retroelements (DGRs) are unique and unparalleled generators of massive protein sequence diversity [1, 2]. At least 10^{20} sequences are possible in proteins diversified by these retroelements [3]. This scale of variation exceeds by several orders of magnitude the variation brought about by the adaptive immune systems of jawed and jawless vertebrates, the only other examples of natural massive protein sequence variation [4, 5]. In these immune systems, massive sequence variation of immunoreceptors permits the recognition of novel ligands and is a robust means for adaptation to dynamic environments. Similarly, massive sequence variation by DGRs appears to enable adaptation to dynamic environments for the ecologically diverse microbes – bacteria, archaea, and microbial viruses – that encode them. These include constituents of the human microbiome [2, 3, 6-12] as well as of the microbial 'dark matter' which constitutes a major fraction of microbial life [13-17].

Three fundamental components define DGRs: a reverse transcriptase (RT) that has a unique sequence motif [1, 2, 18], a variable region (*VR*) that forms part of the coding sequence of a DGR diversified protein, and a template region (*TR*) that is nearly identical to *VR* and is located proximally to *VR* (Fig 1). The *TR* serves as an invariant store of amino acid coding information. This information is transferred from *TR* to *VR*, and during this process adenines within *TR* are specifically mutated to other bases. A recent study on the prototypical DGR of *Bordetella* bacteriophage indicates that the DGR RT in association with a second DGR protein, the accessory variability determinant (Avd), is necessary and sufficient for mutagenesis of adenines in *TR* [19]. Adenine-specific mutagenesis culminates in substitutions occurring only at amino acids that have adenines in their codons in *TR*. AAY is the most recurrent adenine-containing codon in the *TR* of a variety of DGRs [2]. As previously noted, adenine-mutagenesis of Asn-encoding AAY can result in the encoding of 14 other amino acids, which cover the gamut of amino acid chemical character, but cannot result in a stop codon [20].

Fig 1. *T. aquaticus* DGR. The *T. aquaticus* prophage DGR contains a gene encoding a variable protein (*taqvp*). The 3' end of *taqvp* contains the variable region (*VR*) and initiation-of-mutagenic homing (IMH) sequence. The DGR also contains an accessory variability determinant (*avd*) gene, followed by an invariant template region (*TR*), which differs from *VR* mainly at adenines. A sequence similar but not identical to the *VR* IMH occurs at the 3' end of the *TR* and is called IMH*. Following these elements is a gene encoding a reverse transcriptase (*rt*). *TR* is transcribed to produce *TR*-RNA, which is reverse transcribed to produce *TR*-cDNA, with adenine-specific mutagenesis of the sequence accompanying reverse transcription. Adenine-mutagenized *TR*-cDNA homes to and replaces *VR* to yield a variant TaqVP.

DGR variable proteins are divergent in sequence (roughly ≤ 20% identity) [2, 20]. To date, structures of three DGR variable proteins have been determined — *Bordetella* bacteriophage Mtd [20, 21], *Treponema denticola* TvpA [3], and *Nanoarchaeota* AvpA [22]. Despite the sequence divergence of these proteins, they all share a C-type lectin (CLec)-fold, which appears to be an evolutionarily conserved scaffold among DGRs for accommodating massive protein sequence variation. The CLec-fold, despite its nomenclature, is a general ligand-binding fold [23]. Variable amino acids in these CLec-fold DGR variable proteins are solvent-exposed and form ligand-binding sites. Some DGR variable proteins, including those that exist in the human virome [6], are predicted instead to have an immunoglobulin (Ig)-fold as a scaffold for variation. The β-strand body of the Ig-fold appears to be the locus for variation in these proteins, rather than loop regions as in antibodies and T cell receptors. Yet other DGR variable proteins have folds that at present cannot be assigned by sequence alone to CLec- or Ig-folds, or to other protein folds [2].

DGR variable proteins with characterized CLec-folds have the potential to add to the existing toolbox of selectable variable protein scaffolds for molecular display purposes, which includes affibodies, fibronectin type III (FN3) domains, and designed ankyrin repeat proteins (DARPins) among others [24, 25]. As thermostability is an attractive property of selectable variable proteins, we focused on a DGR encoded by a prophage of the thermophile *Thermus aquaticus* [2, 26]. The *T. aquaticus* prophage DGR has a genetic structure similar to that of the *Bordetella* bacteriophage DGR (Fig 1). The variable protein of the *T. aquaticus* prophage DGR, which we call TaqVP, is predicted to have a CLec-fold [2, 26]. We determined the 2.8 Å resolution crystal structure of TaqVP and confirmed that it indeed has a CLec-fold. We also characterized the thermostability of TaqVP, along with those of the previously structurally characterized DGR variable proteins Mtd and TvpA, and found that thermostability was a shared property of several DGR CLec-fold proteins.

Materials and Methods

TaqVP expression and purification

The coding sequence of TaqVP (accession no. CP010822.1) was synthesized with codons optimized for expression in *Escherichia coli* (GENEWIZ, Inc.) and cloned into a modified pET28b expression vector encoding an N-terminal His-tag followed by a PreScission protease cleavage site. The integrity of the construct was confirmed by DNA sequencing. TaqVP was expressed in *E. coli* BL21-Gold (DE3). Bacteria were grown with shaking at 37 °C to an OD_{600} of 0.6–0.8 and then cooled to room temperature, followed by induction with 0.5 mM isopropyl β-D-1-thiogalactopyranoside. Bacteria were grown with shaking at room temperature for 4 h further, then harvested by centrifugation (30 min, 5,000 x g, 4 °C); the bacterial pellet was frozen at -80 °C.

Cells were thawed and resuspended in buffer A (250 mM NaCl, 50 mM Tris, pH 8, and 5 mM β-mercaptoethanol; 20 ml/L of bacterial culture) supplemented with 1 mM phenylmethylsulfonyl fluoride. Bacteria were lysed using an Emulsiflex (Manufacturer), and the lysate was centrifuged (30 min, 35,000 x g, 4 °C). The supernatant was incubated at 55 °C for 10 min, and the sample was centrifuged (30 min, 35,000 x g, 4 °C). The supernatant was then applied to a column containing His-Select Nickel affinity gel (Sigma, 1 ml of resin per 20 ml of bacterial lysate), which had been equilibrated with buffer A. The column was washed with 10 column volumes of buffer B (250 mM NaCl, 20 mM Tris, pH 8, and 5 mM β-mercaptoethanol) containing 20 mM imidazole, and the

TaqVP was eluted with buffer B containing 250 mM imidazole. The His-tag was removed by PreScission protease cleavage (1:50 mass ratio TaqVP:protease) overnight at 4 °C. Cleaved TaqVP was separated from the non-cleaved protein by applying the sample to a His-Select Nickel affinity gel column (Sigma) and collecting the flow-through. TaqVP was further purified by gel filtration chromatography (Superdex 75) in 150 mM NaCl, 20 mM Tris, pH 8, and 1 mM dithiothreitol.

Crystallization and structure determination

Selenomethionine (SeMet)-substituted TaqVP was expressed by culturing *E. coli* in synthetic minimal media supplemented with 200 mg/L L(+)-selenomethionine (Sigma) [27]. Purified SeMet-labeled TaqVP was concentrated to 50 mg/mL by ultrafiltration (10 kDa MWCO Amicon, Millipore); the concentration of TaqVP was determined using a calculated molar extinction coefficient at 280 nm of 80,900$M^{-1}cm^{-1}$.

Crystals of SeMet-labeled TaqVP were grown by the hanging drop method at 20 °C by mixing 1 µL of TaqVP (50 mg/mL) and 1 µL of 15 % (w/v) 2-methyl-2,4-pentanediol, 20 mM $CaCl_2$, 100 mM sodium acetate, pH 4.6. Crystals were cryoprotected by soaking in the precipitant solution supplemented with 20% glycerol. Single-wavelength anomalous dispersion (SAD) data were collected at Advanced Photon Source (Argonne, IL) beamline 24-ID-E. Diffraction data were indexed, integrated, and scaled with MOSFLM [28-30]. Se sites were located from SAD data of SeMet-labeled TaqVP, and initial phases were determined using SOLVE [31]. All five methionines (M1, M24, M112, M227, and M320) were located, and the asymmetric unit was found to contain four molecules of TaqVP.

An initial model of TaqVP was built by automatic means using Autobuild (within Phenix). A total of 75 iterative rounds of manual model building and maximum likelihood refinement were carried out with Refine (within Phenix) using default parameters, with each refinement step consisting of 3 cycles [32, 33]. Each refinement step included manual model rebuilding with COOT, guided by σ_A-weighted $2mF_o$-DF_c and mF_o-DF_c difference maps [34]. One round of TLS parameterization with default settings was then used, followed by the addition of water, calcium, and acetate ions into ≥3σ mF_o-DF_c density. Structure validation was carried out with Molprobity [35], and molecular figures were generated with PyMOL (http://www.pymol.org/). The crystal structure and structure factors have been deposited to the Protein Data Bank (accession no. 5VF4).

Structural alignment of VR and equivalent regions

The structure of the VR of TaqVP (amino acids 341–374) was compared to that of the VR of Mtd (amino acids 337–381), TvpA (amino acids 285–329), and AvpA (amino acids 181–210) using FATCAT [36].

CD Spectroscopy

Temperature denaturation scans of 0.4 mg/mL TaqVP, 0.3 mg/mL Mtd-P1, and 0.3 mg/mL TvpA in 40 mM NaF and 10 mM NaP*i* buffer, pH 7.5 were carried out by monitoring the circular dichroism (CD) signal at 216 nm between 25 °C and 110 °C in 1° intervals. A quartz cell with a 1-mm path length and a model 202 spectrometer (Aviv Instruments) equipped with thermoelectric temperature control were used. The wavelength of 216 nm was chosen because it yielded the maximum CD signal based on a scan of these proteins between 195 and 260 nm in 1-nm intervals at 24 °C. The CD signal from buffer alone was subtracted from the data before conversion to mean residue ellipticity. Denaturation profiles represent averages from three independent experiments. The slope at each point of the denaturation profile was determined using a running boxcar average of five points — the point being assessed and the four preceding it (i.e., lower temperature). The temperature at which proteins started to unfold was defined as the temperature immediately prior to one that had a slope of > 0.1 for TaqVP and Mtd-P1 or < -0.1 for TvpA.

Results and Discussion

Structure of TaqVP

TaqVP was overexpressed in *Escherichia coli*, purified, and crystallized. Single-wavelength anomalous dispersion (SAD) data from selenomethionine-labeled crystals of TaqVP were collected and used to determine the structure of TaqVP. The structure was determined and refined to 2.81 Å resolution limit (Table 1), and the entire length of TaqVP, amino acids 1-381, was modeled. TaqVP appeared by gel filtration chromatography to be monomeric in solution and was observed to be monomeric in the crystal (Fig 2a and 2b).

Table 1. Data collection, phasing and refinement statistics

Data collection	
Space group	P 3_2 2 1
Cell dimensions	
a, *b*, *c* (Å)	155 155 202
α, β, γ(°)	90, 90, 120
Wavelength	0.979 Å
Resolution (Å)	134.27-2.81 (3.29-2.81)[a]
R_{merge}	0.18 (1.00)
I / σ_I	12.5 (1.6)
Completeness (%)	100 (100)
Redundancy	7.4 (6.9)
$\mathrm{cc}_{1/2}$	0.98 (0.60)
Refinement	
Resolution (Å)	80.81-2.81 (3.29-2.81)
No. reflections	132403 (12794)
R_{work} / R_{free}	0.18 (0.31)/0.23 (0.33)
No. atoms	
Se	18
Ca	8
S	4
O	2384
N	2032
C	7223
H	90
Average *B*-factors	60.1
R.m.s deviations	
Bond lengths (Å)	0.009
Bond angles (°)	1.22
MolProbity score	2.1 [98th][b]
Ramachandran	
% preferred	93.7
% allowed	5.8
% disallowed	0.5
Clashscore	7.6 [99th][b]

[a] Highest resolution bin in parentheses here and other rows.

[b] Percentile in brackets here and other rows.

Fig 2. Structure of TaqVP. a. TaqVP in ribbon representation (α-helices gold, β-strands blue, loops grey, and VR purple). The amino acid positions of the N- and C-termini of TaqVP are indicated. **b.** TaqVP in ribbon representation with the core elements of the CLec-fold in red (α-helices) and blue (β-strands). Insert 1’, amino acids 129-217, is magenta; insert 2, amino acids 222-296, green; insert 3,

amino acids 301-331, teal. **c.** The core elements of the CLec-fold in TaqVP in ribbon representation (α-helices red, β-strands and loops blue). The inserts are ghosted. **d.** Topology diagram of the CLec-fold in TaqVP. **e.** Superposition of a portion of the TaqVP VR (red) and the catalytic site of *h*FGE (green) in Cα representation. The catalytic triad of *h*FGE is in bonds representation, as are structurally equivalent or near-equivalent amino acids of TaqVP. **f.** Inserts of TaqVP in ribbon representation, with core elements of the CLec-fold are ghosted (color coding same as in panel b).

The structure of TaqVP revealed a single globular CLec-fold domain that strongly resembles the CLec-fold of two previously determined bacterial DGR variable proteins (Fig 2c and 2d). Specifically, TaqVP closely resembles TvpA (2.5 Å rmsd; 109 Cα; Z = 6.0; 15.8% sequence identity) and the CLec domain of Mtd (2.6 Å rmsd; 102 Cα; Z = 5.4, 20.9% sequence identity) [3, 20]. Mtd contains in addition to its C-terminal CLec domain, β-prism and β-sandwich domains which promote its obligatory trimeric state. The other two previously structurally characterized DGR variable proteins (i.e., TvpA and AvpA) are, like TaqVP, single-domain, monomeric proteins.

Like TvpA and Mtd, TaqVP has the formylglycine-generating enzyme (FGE) subtype of the CLec-fold and, like these other two proteins, shares structural homology with human FGE (3.2 Å rmsd; 116 Cα; Z = 6.2; 22% sequence identity) [37]. TaqVP is more distantly related to the CLec-fold of the archaeal DGR variable protein AvpA (3.3 Å rmsd; 114 Cα; Z = 3.7, 14%

sequence identity), which does not belong to the FGE subtype [22].

As noted for TvpA [3], the structural homology in TaqVP to FGE raises the possibility of enzymatic activity. This structural homology in TaqVP extends to the segment containing the catalytic triad of *h*FGE (Fig 2e) (rmsd 1.23 Å; 37 Cα; $p<0.001$; 28.6% sequence identity). While TaqVP lacks the specific amino acids of the *h*FGE catalytic triad, adenine-mutagenesis could supply two of these. These two, *h*FGE Ser 333 and Cys 336 (Fig 2e), superimpose closely with TaqVP D348 and N351, which are variable and through adenine-mutagenesis could be substituted by Ser and Cys, respectively. The third member of the triad, Cys 341, has no close structural equivalent in TaqVP, but TaqVP V354, which is variable and thus could be substituted by Cys, is proximal. Human FGE contains an extensive, sequence-specific binding site for its peptide substrate, and thus TaqVP would not only require the Ser-Cys-Cys catalytic triad but also specific surrounding amino acids that could provide binding affinity for a peptide substrate. Based on these considerations, it seems unlikely that FGE activity in TaqVP or other DGR variable proteins could be easily assembled, but there is no formal reason why other enzymatic activities could not occur. Catalytic activity engineered into antibodies provides precedence for this notion [38].

The FGE-type CLec domain of TaqVP begins at amino acid 112 and extends to its C-terminal amino acid 381. Preceding amino acid 112 is a region composed of short β-strands and α-helices that wraps around the N-terminus of the CLec domain. The characteristic core elements of the CLec-fold, which include the N- and C-termini forming anti-parallel β-strands (β1 and β5) and two roughly perpendicular α–helices (α1 and α2), are found in TaqVP. These core elements also include a four-stranded, anti-parallel β-sheet (β2 β3 β4 β4’). In TaqVP three of the four strands of the sheet are present, with the β4’ strand being replaced by a loop. The sequence between the β3 and β5 strands forms the putative ligand-binding site. TaqVP also has short segments inserted between the secondary elements of the CLec fold (Fig 2f), as also seen in Mtd, TvpA, and AvpA.

Variable Region

The variable region of TaqVP (amino acids 341-381) is located at the very C-terminus of the protein, as in Mtd and TvpA. The TaqVP VR is 40 amino acids in length — between the 45-amino acid VRs of Mtd and TvpA and the 23-amino acid VR of AvpA. As shown previously for all structurally characterized DGR variable proteins, the nine variable amino acids in TaqVP are solvent exposed and form a potential binding site (Fig 3a and 3b). Variable hydrophobic and hydrophilic amino acids are segregated from each other at this site, which has dimensions of ~10 by ~70 Å (Fig 3b). TaqVP has two nonvariable aromatic amino acids (W349 and W371) within the binding site, which might provide a constant hydrophobic contact to ligands. Similarly, Mtd, TvpA, and AvpA have one or two nonvariable aromatic acids within their ligand-binding sites.

Fig 3. Variable region of TaqVP. **a.** VR of TaqVP in ribbon representation. The main chain is in gray, side chains of variable amino acids are in green (sphere is

glycine), and nonvariable aromatic amino acids are in orange. **b.** Surface representation of TaqVP, with the VR facing the viewer. Variable hydrophobic amino acids (I, V, and Y) are green, variable hydrophilic amino acids (S, N, D, and R) blue, and variable glycine pale orange. **c.** Superposition of the VR of TaqVP (red) and Mtd-P1 (orange) in Cα representation. The spheres represent the position of variable amino acids in each protein. **d.** Superposition of the VR of TaqVP (red) and TvpA (blue) in Cα representation. **e.** Superposition of the VR of TaqVP (red) and AvpA (magenta) in Cα representation. **f.** Stabilization of the main chain of VR (gray, Cα of variable amino acids indicated by spheres) by insert 1’ (magenta) in Cα representation. Dashed line indicates hydrogen bonds.

Remarkably, the TaqVP variable region and those of Mtd and TvpA are nearly identical in structure, despite their weak sequence relationship (rmsd 0.46 Å, 26 Cα, $p<0.001$, 28.9% sequence identity; and rmsd 1.10 Å, 30 Cα, $p<0.001$, 23.1% sequence identity, respectively) (Figs 3c and 3d). However, the TaqVP variable region does not share significant structural similarity with the variable region of the more distantly related AvpA (rmsd, 2.28 Å; 22 Cα; $p>0.1$) (Fig 3e). The nine variable amino acids in TaqVP VR support a potential diversity through adenine-mutagenesis of 3×10^9. Seven of the 9 variable amino acids are encoded by AAY codons in *TR* (D348, V354, Y361, N364, S368, and R370 by AAC and N351 by AAT), enabling substitution by 14 other amino acids at each of these positions. The remaining two variable amino acids are encoded by an ATC (I344) or AGT (to G372) codon, which enables substitution by three other amino acids. All nine of the TaqVP variable amino acids have structural equivalents in TvpA, while eight of the nine have structural equivalents in Mtd (Figs 3c and 3d). The structural similarity among these VRs suggests that a composite VR having maximal diversity may be designed from these.

The last seven amino acids (375-381) of TaqVP VR are invariant and located along the β5-strand, as in Mtd and TvpA. The corresponding DNA sequence is likely to function as the initiation of mutagenic homing (IMH) element, a critical component of DGRs that along with the IMH* element in *TR* defines the directionality of cDNA transfer [18] (Fig 1).

Inserts

TaqVP has three inserts within the core of the CLec-fold. Insert 1' (amino acids 129-217) is located between α1 and α2; an equivalent to insert 1' occurs in TvpA. Insert 1' in TaqVP appears to stabilize the VR through hydrogen bonding (Fig 3f), whereas insert 1' in TvpA interacts with helix α1 instead. Insert 2 of TaqVP is between α2 and β2 (amino acids 222-296). An equivalent occurs in all DGR variable proteins structurally characterized to date, but unlike these others, insert 2 in TaqVP does not interact with the VR. Insert 3 (amino acids 301-331) is between β2 and β3 of the CLec-domain and is composed of loops and a short α-helix; an equivalent for insert 3 occurs in AvpA.

Thermal Stability

The thermal stability of TaqVP was determined by monitoring its secondary structure as a function of temperature by circular dichroism. We observed that TaqVP started to unfold at 70 °C and was completely denatured by 90 °C, as monitored by the loss of ellipticity (Fig 4a). The thermal unfolding of TaqVP was irreversible. AvpA has previously been studied for its thermal stability and it was found that this protein also adopts a thermostable fold, starting to unfold at ~65 °C and becoming completely denatured by 80 °C [10]. We also determined the thermal stabilities of Mtd-P1 and TvpA. *Bordetella* bacteriophage Mtd-P1, which unlike the other DGR variable proteins studied here has multiple domains, was found to start irreversibly unfolding at 70 °C and was completely denatured by 80 °C (Fig 4b). TvpA began unfolding at 50 °C and was completely denatured by 70 °C (Fig 4c). For TvpA, the ellipticity became more negative upon unfolding. This shift to more negative ellipticity upon unfolding has been seen previously in the unfolding of dihydrofolate reductase due to the involvement of W47 and W74 forming an exciton pair [39]. TvpA has a structurally similar pairing of tryptophans (W138 and W263) that could form an exciton pair.

Fig 4. Thermal stability of DGR variable proteins. Circular dichroism signal (mean residue ellipticity, MRE) at 216 nm for the transition from 25 to 110 °C (green), and transition from 110 to 25 °C (blue) for TaqVP (**a**), Mtd-P1 (**b**), and TvpA (**c**). Means and standard deviations from three separate experiments are shown.

A recent structural proteomic analysis identified increased lysine and β-sheet content as promoting thermostability in proteins [40]. However, the lysine content for TaqVP, Mtd-P1, AvpA, and TvpA (in descending order of thermostability) is 2.7, 3.7, 7.7, and 5.8%, respectively, which is not very different from the ~5.3% average for 200-amino acid proteins [41]. The β-sheet content in TaqVP, Mtd-P1, AvpA, and TvpA is 23, 30, 24, and 19%, respectively, which is not atypical. Thus, the basis for thermostability in these proteins is not readily apparent and requires further study.

In summary, these results suggest that thermostability is a shared property of several structurally characterized CLec-fold DGR variable proteins (i.e., TaqVP, Mtd, and AvpA), and that these proteins may provide an advantageous scaffold for molecular display applications.

Acknowledgments

We thank Brent Hamaoka and Cosmo Buffalo for their help with the crystallization of TaqVP.

References

1 Liu M, Deora R, Doulatov SR, Gingery M, Eiserling FA, et al. (2002) Reverse Transcriptase-Mediated Tropism Switching in Bordetella Bacteriophage. Science 295: 2091-2094.

2 Wu L, Gingery M, Abebe M, Arambula D, Czornyj E, et al. (2018) Diversity-generating retroelements: natural variation, classification and evolution inferred from a large-scale genomic survey. Nucleic Acids Res 46: 11-24.

3 Le Coq J and Ghosh P (2011) Conservation of the C-type lectin fold for massive sequence variation in a Treponema diversity-generating retroelement. Proc Natl Acad Sci U S A 108: 14649-14653.

4 Boehm T, McCurley N, Sutoh Y, Schorpp M, Kasahara M, et al. (2012) VLR-based adaptive immunity. Annu Rev Immunol 30: 203-220.

5 Litman GW, Rast JP and Fugmann SD (2010) The origins of vertebrate adaptive immunity. Nat Rev Immunol 10: 543-553.

6 Minot S, Grunberg S, Wu GD, Lewis JD and Bushman FD (2012) Hypervariable loci in the human gut virome. Proc Natl Acad Sci U S A 109: 3962-3966.

7 Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, et al. (2013) Rapid evolution of the human gut virome. Proc Natl Acad Sci U S A 110: 12450-12455.

8 Ye Y (2014) Identification of diversity-generating retroelements in human microbiomes. Int J Mol Sci 15: 14234-14246.

9 Guo H, Arambula D, Ghosh P and Miller JF (2014) Diversity-generating Retroelements in Phage and Bacterial Genomes. Microbiol Spectr 2.

10 Paul BG, Bagby SC, Czornyj E, Arambula D, Handa S, et al. (2015) Targeted diversity generation by intraterrestrial archaea and archaeal viruses. Nat Commun 6: 6585.

11 Nimkulrat S, Lee H, Doak TG and Ye Y (2016) Genomic and Metagenomic Analysis of Diversity-Generating Retroelements Associated with Treponema denticola. Front Microbiol 7: 852.

12 Arambula D, Wong W, Medhekar BA, Guo H, Gingery M, et al. (2013) Surface display of a massively variable lipoprotein by a Legionella diversity-generating retroelement. Proc Natl Acad Sci U S A 110: 8212-8217.

13 Paul BG, Burstein D, Castelle CJ, Handa S, Arambula D, et al. (2017) Retroelement-guided protein diversification abounds in vast lineages of Bacteria and Archaea. Nat Microbiol 2: 17045.

14 Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, et al. (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523: 208-211.

15 Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, et al. (2016) A new view of the tree of life. Nat Microbiol 1: 16048.

16 Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, et al. (2016) Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat Commun 7: 13219.

17 Castelle CJ and Banfield JF (2018) Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. Cell 172: 1181-1197.

18 Doulatov S, Hodes A, Dai L, Mandhana N, Liu M, et al. (2004) Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. Nature 431: 476-481.

19 Handa S, Jiang Y, Tao S, Foreman R, Schinazi RF, et al. (2018) Template-assisted synthesis of adenine-mutagenized cDNA by a retroelement protein complex. Nucleic Acids Res 46: 9711-9725.

20 McMahon SA, Miller JL, Lawton JA, Kerkow DE, Hodes A, et al. (2005) The C-type lectin fold as an evolutionary solution for massive sequence variation. Nat Struct Mol Biol 12: 886-892.

21 Miller JL, Coq JL, Hodes A, Barbalat R, Miller JF, et al. (2008) Selective Ligand Recognition by a Diversity-Generating Retroelement Variable Protein. PLoS Biol 6: e131.

22 Handa S, Paul BG, Miller JF, Valentine DL and Ghosh P (2016) Conservation of the C-type lectin fold for accommodating massive sequence variation in archaeal diversity-generating retroelements. BMC Struct Biol 16: 13.

23 Zelensky AN and Gready JE (2005) The C-type lectin-like domain superfamily. FEBS J 272: 6179-6217.

24 Nuttall SD and Walsh RB (2008) Display scaffolds: protein engineering for novel therapeutics. Curr Opin Pharmacol 8: 609-615.

25 Hosse RJ, Rothe A and Power BE (2006) A new generation of protein display scaffolds for molecular recognition. Protein Sci 15: 14-27.

26 Brumm PJ, Monsma S, Keough B, Jasinovica S, Ferguson E, et al. (2015) Complete Genome Sequence of Thermus aquaticus Y51MC23. PLoS One 10: e0138674.

27 Doublie S (2007) Production of selenomethionyl proteins in prokaryotic and eukaryotic expression systems. Methods Mol Biol 363: 91-108.

28 Collaborative Computational Project N (1994) The CCP4 suite: programs for protein crystallography. Acta Crystallogr D Biol Crystallogr 50: 760-763.

29 Evans P (2006) Scaling and assessment of data quality. Acta Crystallogr D Biol Crystallogr 62: 72-82.

30 Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, et al. (2011) Overview of the CCP4 suite and current developments. Acta Crystallogr D Biol Crystallogr 67: 235-242.

31 Terwilliger TC and Berendzen J (1999) Automated MAD and MIR structure solution. Acta Crystallogr D Biol Crystallogr 55: 849-861.

32 Murshudov GN, Vagin AA and Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr 53: 240-255.

33 Adams PD, Grosse-Kunstleve RW, Hung LW, Ioerger TR, McCoy AJ, et al. (2002) PHENIX: building new software for automated crystallographic structure determination. Acta Crystallogr D Biol Crystallogr 58: 1948-1954.

34 Emsley P and Cowtan K (2004) Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr 60: 2126-2132.

35 Davis IW, Murray LW, Richardson JS and Richardson DC (2004) MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. Nucleic Acids Res 32: W615-619.

36 Ye Y and Godzik A (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics 19: ii246-ii255.

37 Dierks T, Dickmanns A, Preusser-Kunze A, Schmidt B, Mariappan M, et al. (2005) Molecular Basis for Multiple Sulfatase Deficiency and Mechanism for Formylglycine Generation of the Human Formylglycine-Generating Enzyme. Cell 121: 541-552.

38 Tramontano A, Janda KD and Lerner RA (1986) Catalytic antibodies. Science 234: 1566-1570.

39 Kuwajima K, Garvey EP, Finn BE, Matthews CR and Sugai S (1991) Transient intermediates in the folding of dihydrofolate reductase as detected by far-ultraviolet circular dichroism spectroscopy. Biochemistry 30: 7693-7703.

40 Leuenberger P, Ganscha S, Kahraman A, Cappelletti V, Boersema PJ, et al. (2017) Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. Science 355.

41 Carugo O (2008) Amino acid composition and protein dimension. Protein Sci 17: 2187-2191.

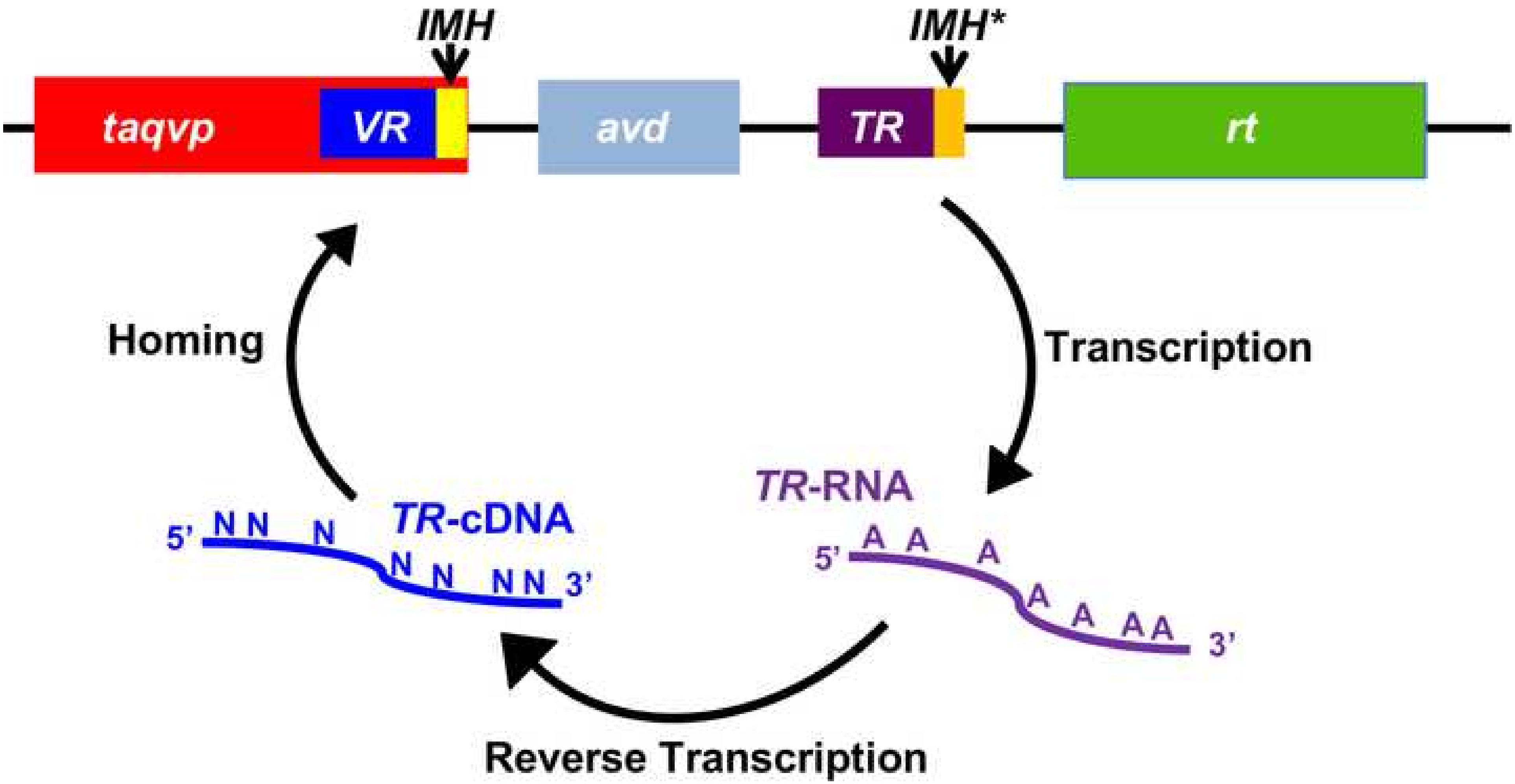
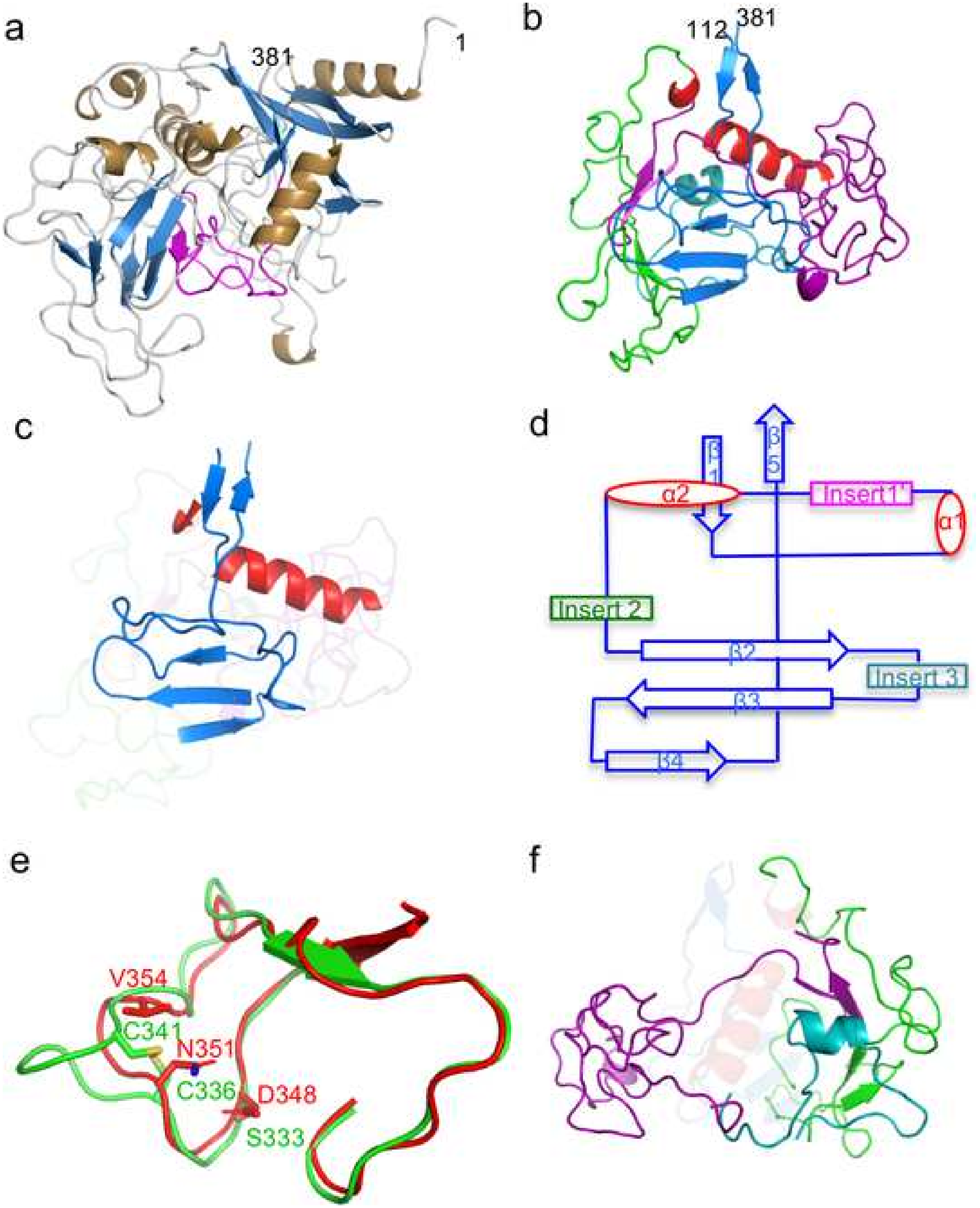

IMH
IMH*
taqvp
VR
avd
TR
rt
Homing
Transcription
TR-RNA
5' A A A A A AA 3'
TR-cDNA
5' N N N N N N N 3'
Reverse Transcription

Figure 2


a
381
1
b
112
381
c
d
β1
β5
α2
Insert1'
α1
Insert 2
β2
Insert 3
β3
β4
e
V354
C341
N351
C336
D348
S333
f

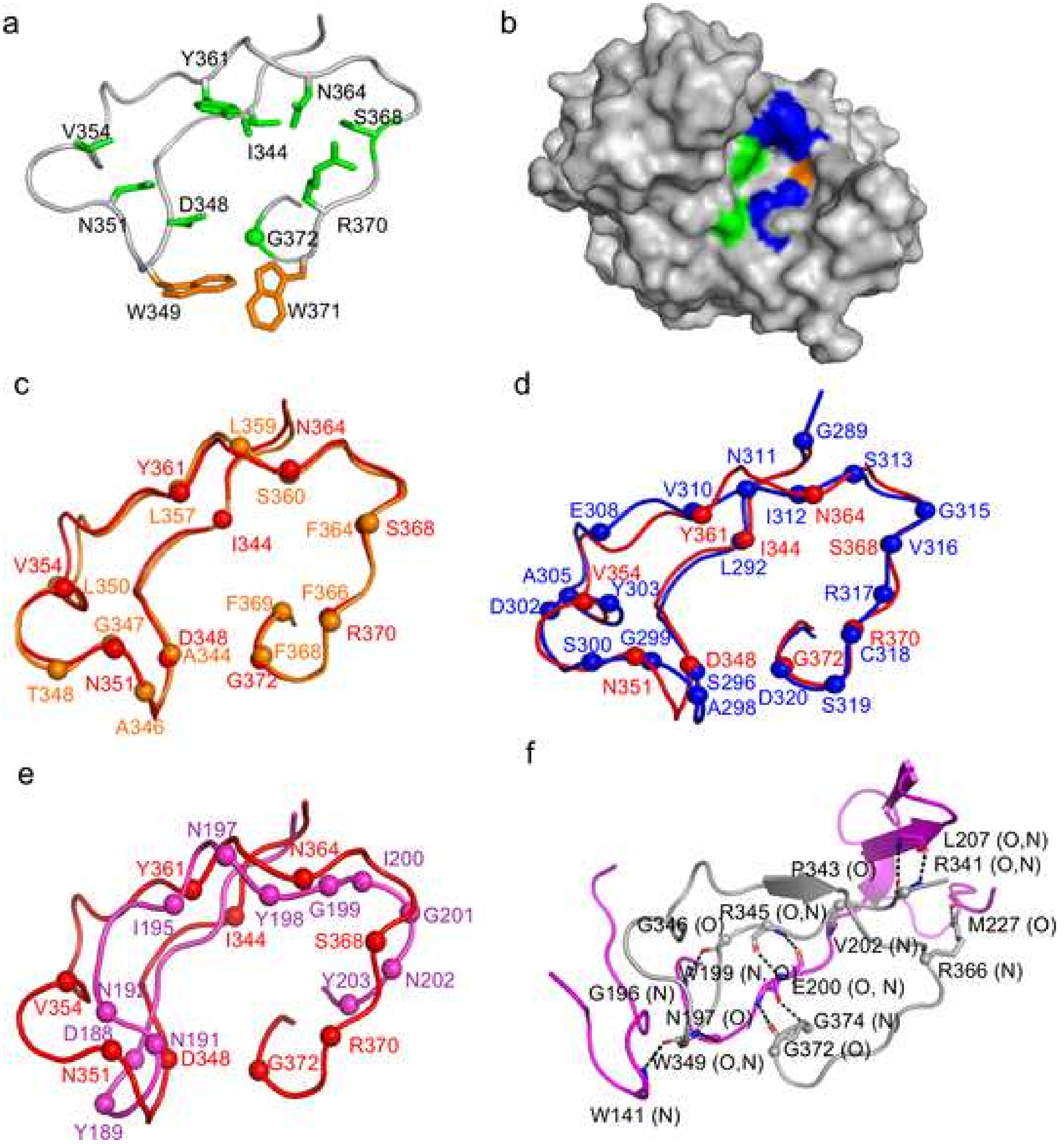
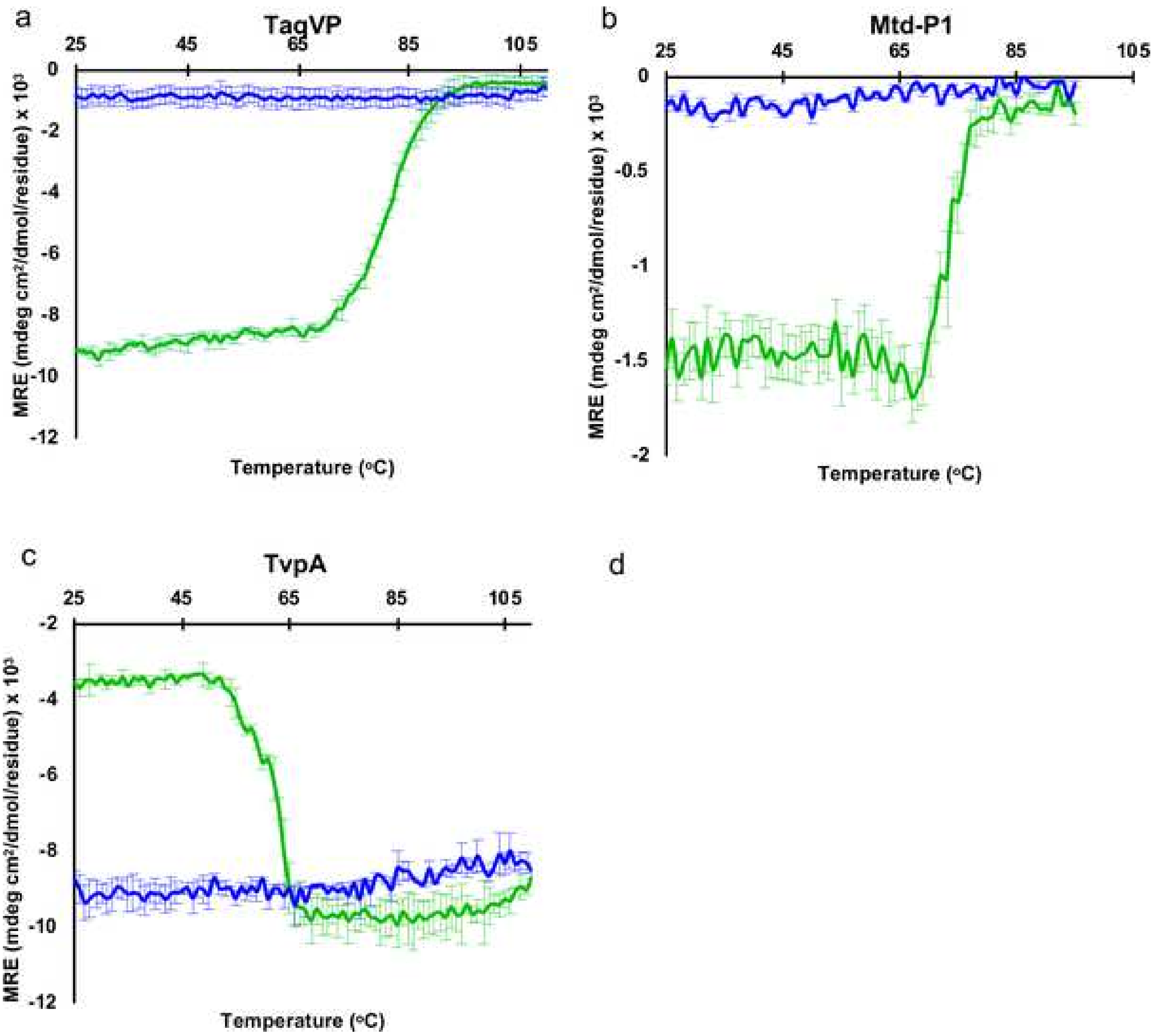

a
Y361
N364
V354
S368
I344
D348
N351
R370
G372
W349
W371
b
c
L359
N364
Y361
S360
L357
I344
F364
S368
V354
L350
F369
F366
G347
D348
A344
R370
F368
T348
N351
G372
A346
d
G289
N311
S313
V310
E308
I312
N364
G315
Y361
I344
S368
V316
L292
A305
V354
Y303
R317
D302
R370
S300
G299
C318
D348
G372
S296
N351
D320
S319
A298
e
N197
N364
I200
Y361
G199
G201
I195
Y198
I344
S368
Y203
N202
N192
V354
D188
N191
D348
N351
G372
R370
Y189
f
L207 (O,N)
P343 (O)
R341 (O,N)
R345 (O,N)
G346 (O)
M227 (O)
V202 (N)
W199 (N, O)
R366 (N)
E200 (O, N)
G196 (N)
N197 (O)
G374 (N)
G372 (O)
W349 (O,N)
W141 (N)

Figure 4


a
TaqVP
25
45
65
85
105
0
-2
-4
-6
-8
-10
-12
MRE (mdeg cm²/dmol/residue) x 10³
Temperature (ᵒC)
b
Mtd-P1
25
45
65
85
105
0
-0.5
-1
-1.5
-2
MRE (mdeg cm²/dmol/residue) x 10³
Temperature (ᵒC)
c
TvpA
25
45
65
85
105
-2
-4
-6
-8
-10
-12
MRE (mdeg cm²/dmol/residue) x 10³
Temperature (ᵒC)
d