Discrete analysis of camelid variable domains: sequences, structures, and in-silico structure prediction

Akhila Melarkode Vattekatte, Nicolas Ken Shinada, Tarun J Narwani, Floriane Noël, Olivier Bertrand, Jean-Philippe Meyniel, Alain Malpertuy, Jean-Christophe Gelly, Frédéric Cadet, Alexandre de Brevern

Discrete analysis of camelid variable domains: sequences, structures, and in-silico structure prediction

Akhila Melarkode Vattekatte[1,2,3,4], Nicolas Ken Shinada[1,2,4,5], Tarun J. Narwani[1,2,4], Floriane Noël[1,2,4,6,7], Olivier Bertrand[1,2,4], Jean-Philippe Meyniel[8], Alain Malpertuy[9], Jean-Christophe Gelly[1,2,4,10], Frédéric Cadet[1,2,3,11] and Alexandre G. de Brevern[1,2,3,4,10]

[1] Biologie Intégrée du Globule Rouge UMR_S1134, Inserm, Univ. Paris, Univ. de la Réunion, Univ. des Antilles, Paris, France
[2] Laboratoire d'Excellence GR-Ex, Paris, France
[3] Faculté des Sciences et Technologies, Saint Denis, La Réunion, France
[4] Institut National de la Transfusion Sanguine (INTS), Paris, France
[5] Discngine SAS, Paris, France
[6] PSL Research University, INSERM, UMR 932, Institut Curie, Paris, France
[7] Université Paris Sud, Université Paris-Saclay, Orsay, France
[8] ISoft, Saint-Aubin, France
[9] Atragene, Ivry-sur-Seine, France
[10] IBL, Paris, France
[11] Peaccel, Protein Engineering Accelerator, Paris, France

ABSTRACT

Antigen binding by antibodies requires precise orientation of the complementarity-determining region (CDR) loops in the variable domain to establish the correct contact surface. Members of the family Camelidae have a modified form of immunoglobulin gamma (IgG) with only heavy chains, called Heavy Chain only Antibodies (HCAb). Antigen binding in HCAbs is mediated by only three CDR loops from the single variable domain (V_HH) at the N-terminus of each heavy chain. This feature of the V_HH, along with their other important features, e.g., easy expression, small size, thermo-stability and hydrophilicity, made them promising candidates for therapeutics and diagnostics. Thus, to design better V_HH domains, it is important to thoroughly understand their sequence and structure characteristics and relationship. In this study, sequence characteristics of V_HH domains have been analysed in depth, along with their structural features using innovative approaches, namely a structural alphabet. An elaborate summary of various studies proposing structural models of V_HH domains showed diversity in the algorithms used. Finally, a case study to elucidate the differences in structural models from single and multiple templates is presented. In this case study, along with the above-mentioned aspects of V_HH, an exciting view of various factors in structure prediction of V_HH, like template framework selection, is also discussed.

Subjects Biochemistry, Bioinformatics

Keywords Secondary structure, Nanobodies, Complementarity determining regions, Structural alphabet, Frameworks, Sequence structure relationship, Antibodies

Submitted 6 June 2019
Accepted 16 December 2019
Published 6 March 2020

Corresponding author
Alexandre G. de Brevern, alexandre.debrevern@univ-paris-diderot.fr

Academic editor
Mohammed Gagaoua

Additional Information and Declarations can be found on page 21

DOI **10.7717/peerj.8408**

How to cite this article Melarkode Vattekatte A, Shinada NK, Narwani TJ, Noël F, Bertrand O, Meyniel J-P, Malpertuy A, Gelly J-C, Cadet F, de Brevern AG. 2020. Discrete analysis of camelid variable domains: sequences, structures, and in-silico structure prediction. *PeerJ* **8**:e8408 http://doi.org/10.7717/peerj.8408

INTRODUCTION

Immunoglobulin Gamma (IgG) (see Fig. 1A) is a major component of the immune system in vertebrates. Members of the family Camelidae have a modified form of IgG, called Heavy Chain only Antibodies (HCAbs). HCAbs, as their name suggests, are completely devoid of (i) light chains and (ii) C_H1 domain in the heavy chain (see Fig. 1B) (*Hamers-Casterman et al., 1993*). Interestingly, the N-terminal domain of each chain of HCAb, named V_HH, is functional when expressed independently. A V_HH domain is 20 times smaller than complete IgG, ranging from 120 to 150 amino acids in length. Each V_HH domain (see Figs. 1C–1E) has only 3 Complementary Determining Regions (CDR) to bind to their antigens. These loops connect the more structured regions called the Framework Regions (FR), considered sequentially and structurally well conserved.

Due to the lack of a light chain variable domain (V_L), V_HH sequences have evolved to adapt to the hydrophilic environment (*Muyldermans et al., 1994*), also leading to higher thermal tolerance (*Van der Linden et al., 1999*; *Stefan et al., 2002*). Additional topological advantages of V_HH include convex shaped CDR3 found in V_HH that bind to enzymes (*Muyldermans, 2013*), to access epitopes which are otherwise inaccessible to classical antibodies. The small size, ease of expression, unique biochemical and biophysical properties of these domains made them (*Nabuurs et al., 2012*) appealing tool for applied biotechnology, healthcare therapeutics and diagnostics. For instance, V_HH domains can cross the blood–brain barrier (*Rutgers et al., 2011*; *Nabuurs et al., 2012*) or penetrate tumour core, helping in non-invasive screening of tumours (*Massa et al., 2014*; *Rashidian et al., 2015*). AbLynx® (*Ablynx, 2016*) had developed a V_HH against acquired Thrombotic Thrombocytopenic Purpura which was successful in clinical trials (*Peyvandi et al., 2016*), and is patented under the name Caplacizumab in Europe.

In their 2017 study, Zuo and co-workers (*Zuo et al., 2017*) found more than 2,300 sequences of V_HH domains, but 90% in patents and only 74% in PDB structures. With the availability of huge numbers of V_HH sequences in the large majority of the proteins, it is interesting and essential to predict models of V_HH domains, to understand their binding and specific properties. In 2010, our lab proposed the first structural model using comparative modelling of a V_HH have been designed against ACKR1 (*Smolarek et al., 2010b*). Since our study, 21 investigations have used *in silico* structure prediction of V_HH, to either understand structural features or to understand the interactions between V_HH and its ligand using the molecular docking summarised in Table S1.

The most common approach to protein 3D structure prediction using template-based modelling is comparative modelling. The principle is to build the structural model of a query protein sequence using the structure of a homologous protein as a template. In case of comparative modelling of variable domains of IgGs, peculiar challenges arise, which are: (i) the presence of interspersed amino acid regions with varying sequence conservation, (ii) hypervariable regions showing high diversity in length and in conformation and (iii) prediction of the inclination angle between the V_H and the V_L domains which determines the antigen-binding interface. For V_HH, the third criterion does not apply, but the CDR3 loop is longer and is more conformationally diverse compared to its V_H counterparts.

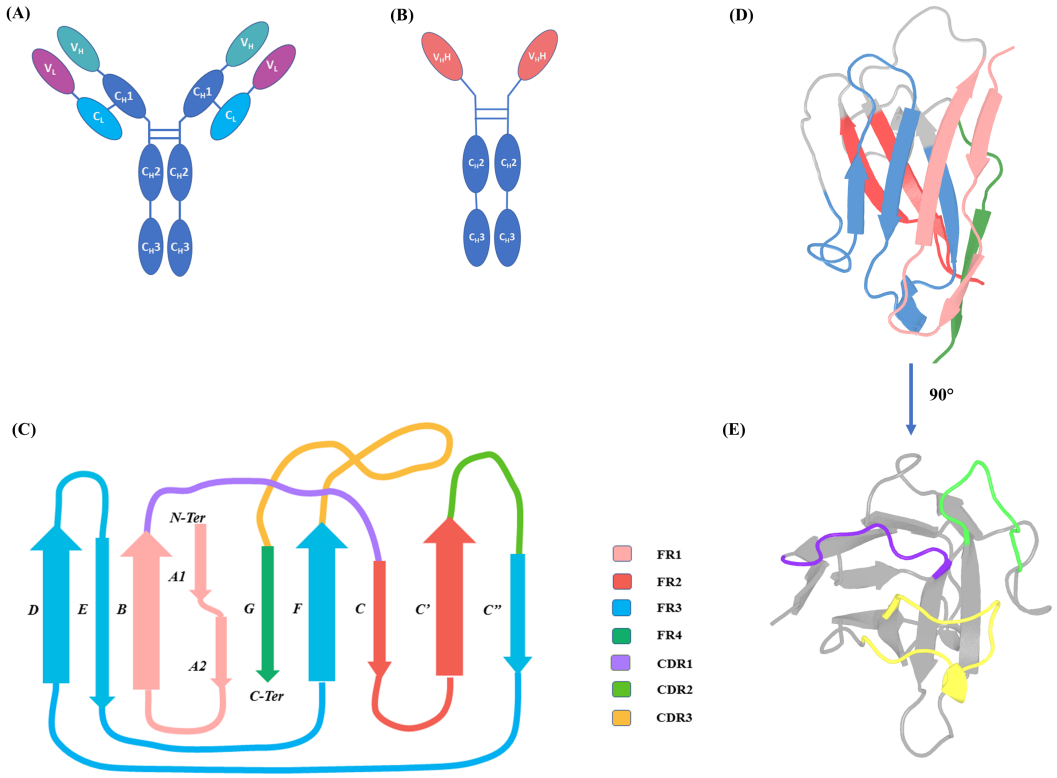

Figure 1 **IgG and HCAb.** (A) IgG schematic representation with heavy chain (domains V_H, C_H1, C_H2 and C_H3) and light chain (domains V_L and C_L) and (B) Heavy Chain Only Antibodies (HCAbs), schematic representation with domains V_HH, C_H2 and C_H3; (C) 2D representation of Immunoglobulin fold of V_HH domain with demarcated Framework Regions (FRs) and Complementarity Determining Regions (CDRs), loop lengths are approximated for representation purposes. FR1 is composed of β-strands A1, A2 and B, FR2 is composed of β-strands C and C′, FR3 of 4 β-strands C″, D, E and F, FR4 end the V_HH sequence by last β-strand, G. (D) 3D cartoon representation of V_HH (PDB ID: 1BZQ chain K (*Decanniere et al., 1999*) and (E) 90° rotation of the same with the CDRs coloured.

Full-size DOI: 10.7717/peerj.8408/fig-1

Antibody Modelling Assessment group (AMA) has held two assessment meetings in 2011 and 2014 to rank different methodologies such as the Dassault Systèmes BIOVIA (*Fasnacht et al., 2014*), RosettaAntibody (*Sircar, Kim & Gray, 2009*), Schrödinger® (*Beard et al., 2013*; *Zhu et al., 2014*; *Salam et al., 2014*), PIGS (*Marcatili, Rosi & Tramontano, 2008*) and KotaiAntibody (*Yamashita et al., 2014*). Most of them use hybrid modelling, which is to model CDR loops separately (by comparative modelling or ab-intio) and using comparative modelling for FR. Since only V_H/V_L from IgGs were used, the protocols tested by AMA may not be extrapolated to V_HH structure prediction. Although some of these methods are capable of modelling V_HH, most of the V_HH structures in our survey used classical comparative modelling protocols.

Another important point is that V_HH topology may seem simple with a conserved immunoglobulin fold composed of "conserved" FRs and a reduced number of CDRs; the complexity arising due to longer CDR1 and CDR3 contribute to non-trivial task of structure prediction. We have shown the crucial choice of the structural template (*Smolarek et al.,*

2010a) that can lead to a totally different orientation of the CDRs. No study has summarised the different approaches since our study, or possible limitations and biases. Our work is intended to make a contribution in this area. The present paper provides an update on (i) different sources of information on V_HH sequence and structures, (ii) global and local properties of V_HH sequences and structures, (iii) different methods used in 3D structure prediction, and (iv) drawbacks of the most frequently used prediction protocol. The main goal of this paper is to provide useful findings to researchers interested in analysing V_HH domains and its structure prediction.

MATERIALS AND METHODS

Datasets

V_HH (nanobody)sequences were obtained from NCBI Genbank (*Benson et al., 2005*) (https://www.ncbi.nlm.nih.gov/genbank/) and Uniprot (*Boutet et al., 2016*) (http://www.uniprot.org/). V_HH structures were taken from the Protein Databank (PDB, https://www.rcsb.org/pdb/home/home.do) (*Berman et al., 2000*). This dataset was put together by March, 2018. Sequences were aligned using Clustal Omega v1.2.1 (http://www.clustal.org) (*Sievers et al., 2011*) and using Jalview v.2.10 (http://www.jalview.org) (*Waterhouse et al., 2009*), and analysed using WebLogo (*Schneider & Stephens, 1990*; *Crooks et al., 2004*). CDRs were analysed using PyIgClassify classification (http://dunbrack2.fccc.edu/pyigclassify/) (*Adolf-Bryfogle et al., 2015*) (described in greater detail below).

Comparative modelling of V_HH

Modeller.9v.16 was used in the study to model V_HH. Briefly, an alignment file in the prescribed format of the query with the desired template(s) is generated. From this alignment, the algorithm derives spatial restraints or probability density functions for each residue to be modelled. 3D model(s) of the query protein are obtained with the optimisation of the input restraints. The best structural model is selected with best DOPE score (*Shen & Sali, 2006*; *Melo & Sali, 2007*).

In the study case, two different scenarios were used: (i) 4 templates were used to propose a multi-template approach and (ii) individual templates were used, leading to five distinct cases. In each case a total of 100 models were generated and were ranked using DOPE score.

Assessment of structural similarity and disulphide bridge conformations of V_HH domains

The most popular method to evaluate similarity is the distanc-based similarity measure Root Mean Squared Deviation (RMSD); it is calculated by the averaging of distances between *n*-pairs of equivalent atoms. It can be applied to a whole protein or a subset of specified equivalent atoms from two proteins. It must be noted that RMSD value takes only into account equivalent residue, i.e., it is not appropriate when the segments are of different lengths. RMSD was computed with ProFit (http://www.bioinf.org.uk/programs/profit/) that is based on the McLachlan algorithm (*McLachlan & IUCr, 1982*) and iPBA (http://www.dsimb.inserm.fr/dsimb_tools/ipba/) (*Gelly et al., 2011*); this last also provides

Global Distance Test Total Score (GDT-TS) values (*McLachlan & IUCr, 1982*; *Zemla, 2003*).

A recent classification of disulphide bridges (*Schmidt, Ho & Hogg, 2006*) includes 20 kinds of disulphide bonds based on the signs of the five dihedral angles between the Cα atoms of any two cysteines involved and the Dihedral strain energy. For example, the sign pattern '- - - - - -' indicates that all the dihedral angles $\chi 1$, $\chi 2$, $\chi 3$, χ '1, χ '2 between the two C α atoms Cys and Cys' have negative values.

Local conformational analysis

Secondary structures were assigned with the most widely used algorithm, namely DSSP (CMBI version 2000) with default parameters (Kabsch et Sander 83). Protein Blocks (PBs) were also used. PBs are a structural alphabet composed of 16 local prototypes (*Joseph et al., 2010*) five residues in length. PBs give a reasonable approximation of all local protein 3D structures (*de Brevern, Etchebest & Hazout, 2000*) and are very efficient in protein superimpositions (*Joseph, Srinivasan & de Brevern, 2012*) and MD (Molecular Dynamics) analyses (*de Brevern et al., 2005*). They are labelled from *a* to *p*. PBs *m* and *d* can be roughly described as prototypes for α-helix and central β-strand, respectively. PBs *a* to *c* primarily represent β-strand N-caps and PBs *e* and *f* represent β-strand C-caps; PBs *g* to *j* are specific to coils; PBs *k* and *l* to α-helix N-caps, while PBs *n* to *p* to α-helix C-caps. PB (*de Brevern, 2005*) assignment was carried out using our PBxplore tool (available at GitHub) (*Barnoud et al., 2017*).

The equivalent number of PBs (N_{eq}) is a statistical measurement similar to an entropy index. It represents the average number of PBs for a residue at a given position which is calculated as follows (*de Brevern, Etchebest & Hazout, 2000*):

$$N_{eq} = \exp\left(-\sum_{x=1}^{16} f_x Inf_x\right) \tag{1}$$

Where, f_x is the probability of PB *x*. A N_{eq} value of 1 indicates that only one type of PB is observed, while a value of 16 is equivalent to a random distribution.

To detect a change in PBs profile, a ΔPB value was calculated. It corresponds to the absolute sum of the differences for each PB between the probabilities of a PB *x* being present in the first and the second structures (*x* goes from PB *a* to PB *p*). ΔPB is calculated as follows:

$$\Delta PB = \sum_{x=1}^{16} |(f_x^1 - f_x^4)|. \tag{2}$$

Where, $f^1{}_x$ and $f^2{}_x$ are the percentages of occurrence of a PB *x* in the analysed structures. A value of 0 indicates perfect PB identity, while a score of 2 indicates a total difference.

The change in PB entropy at given position between two different sets of structures analysed is denoted using ΔN_{eq} which is calculated as the following:

$$\Delta N_{eq} = |N_{eq1} - N_{eq2}|. \tag{3}$$

PyIgClassify database

CDRs support most of the interactions with the epitope. Different classifications have been proposed to analyse them (*Chothia & Lesk, 1987*; *Martin & Thornton, 1996*; *Al-Lazikani, Lesk & Chothia, 1997*; *Shirai, Kidera & Nakamura, 1999*). The most recent classification of CDRs was proposed by North and co-workers (*North, Lehmann & Dunbrack, 2011*). The AHo numbering scheme (*Honegger & Plückthun, 2001*) was used by the authors to enumerate the residues in variable domains. The clusters are regularly updated and are made available through the PyIgClassify database (http://dunbrack2.fccc.edu/PyIgClassify/) (*Adolf-Bryfogle et al., 2015*). All the CDRs except CDRH3 are included in the classification. The classified CDR loop conformations are based on the (a) type of CDR, (b) length of the CDR loop, and (c) affinity-propagation clustering method (*Frey & Dueck, 2007*), using a dihedral angle distance function was used to group CDRs. The clusters are grouped into three categories, (1) Type I, i.e., one cluster CDR-lengths, (2) Type II, i.e., predictable CDR-lengths, and (3) Type III, i.e., unpredictable CDR-lengths. According to the authors, CDR H1 (i.e., CDR1 for V_HH) falls into the Type II category, where the CDR-length combination has multiple possible structures. CDR H2 (i.e., CDR2 for V_HH) falls under Type I, the one-cluster CDR-length. Since Protein Blocks are descriptors of local backbone conformations, they have also been used to analyse the conformations of dense clusters, which include V_HH.

RESULTS

Sequence datasets

Uniprot contains a limited number of V_HH annotated sequences (only 9), while 245 are retrieved from Genbank. In the latter, the majority of V_HH domains (233) are from Camelids. In the PDB, 140 V_HH structures were available at the time of dataset generation. Sequences from PDB and Genbank were combined to constitute a master dataset of 373 sequences. After removing the redundant sequences at >95% sequence identity, 325 representative sequences were selected (see Dataset S1). It is the largest sequence dataset ever used to analyse V_HH domains. Our study focuses on V_HH domains, since others have compared a more limited number of sequences (90) and always compare with V_H as a priority (*Mitchell & Colwell, 2018b*).

Multiple sequence alignment

A first step in the analysis of a specific protein family is to look at amino acid sequence characteristics using a multiple sequence alignment (MSA). Such an alignment was generated with sequence dataset using Clustal Omega (see Fig. S1). Figure 2 shows the analysis of this MSA represented as a sequence logo, where residue conservation at each position was calculated as information content (bits). As expected, FR positions strikingly appear as conserved sequence blocks evidenced by high bit scores. The interspersed CDRs, on the other hand, have less information content in terms of bits. This is mainly due to inherent sequence variability. Results are in good correspondence with the recent analyses of *Mitchell & Colwell (2018a)*, but show more information in terms of bits for the CDRs than this study. This can be due to the use of an alignment tool that can provide less efficient

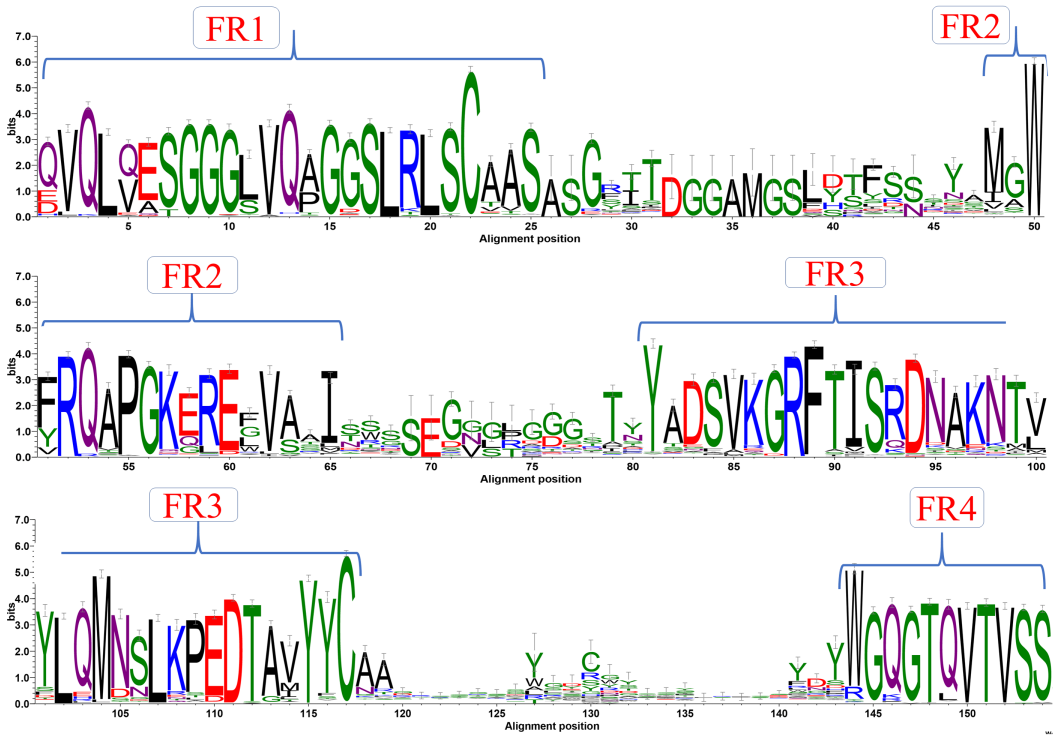

Figure 2 **Sequence conservation.** Sequence logo representation of multiple sequence alignment of complete dataset of V_HH sequences. The relative frequency of amino acids at each position is shown here as sequence logo. The residues are colour coded according to their chemical properties. The residue positions are not in accordance with the numbering systems, sequence alignment creates a longer length than canonical V_H.

Full-size DOI: 10.7717/peerj.8408/fig-2

alignment in variable regions, or simply to a larger number of analysed V_HH domains, since both free and complexed forms are studied. For the first time, this analysis was also performed for each genus (see Fig. S2 for a summary and Figs. S4–S5 for complete sequence profiles). Interestingly, even with the divergence of species, sequence conservation also shows similar conservation for all the different regions with no significant differences.

V_HH sequences have a median length value of 123 amino acids (aa) with minimal and maximal length of 109 aa and 137 aa, respectively (see Fig. S6). The amino acid length distribution in different regions of V_HH (see Fig. S7) shows diversity in CDR lengths, especially in CDR3. The median values for CDR lengths are 8 aa, 8 aa and 16 aa for CDR1, CDR2 and CDR3, respectively. The average length of CDR3 in V_HH is higher than conventional human or mouse V_H sequences (*Muyldermans et al., 2009*). However, an important point is that FRs are not of absolute invariant length: FR1 is 25 aa in length, FR2 is 18 aa, and FR3 is 37 aa, but with some differences of 2–3 residues. These differences are often forgotten, i.e., in the recent (*Mitchell & Colwell, 2018a*; *Mitchell & Colwell, 2018b*) studies, but must be taken into account, or else the analyses could be biased.

The pairwise sequence identity between sequences in the dataset has a median value of 62% and is always above 35% (see Fig. S8A). Thus, it can be considered a relevant case for

homology modelling. The variability of amino acids is not constant in these domains. FRs are more conserved, with sequence identity 84, 72, 81 and 90% (median values) for FRs 1, 2, 3 and 4, respectively (see Figs. S8E–S8H). In the case of CDRs, low sequence identities are observed (below 30%); CDR3 is the lowest with 18%, followed by CDR2 with 25% and CDR1 with 28% (see Figs. S8B–S8D). These results also underline why it is so important to keep a high sequence threshold for building of the dataset, since FRs are highly redundant. Hence, a threshold of 30% would have selected only one V_HH.

Sequence characteristics

At the time of their discovery, V_HH sequences were known to have unique amino acid substitutions in FR2 (*Vu et al., 1997*) compared to V_H found in camelids at positions V42F/Y, G49E/K, L50R/C, W52G/L (IMGT® numbering scheme, see Fig. S1). It was thought that there is a single type of V_HH germline sequence, which has sequence similarity to Clan III of $V_H$germline sequences from humans. However, the discovery of V_HH without this tetrad added exceptions to the former theory. These V_HH domains are assumed to be derived from a different clan similar to Clan II of V_H of human germline sequences (*Deschacht et al., 2010*), and represent 23 sequences (8 from PDB and 15 from NCBI Genbank). Hence this new analysis underlines that the amino acid substitutions of V_HH of FR1 and FR2 are entirely non-specific: (a) V(F/Y) also has V at the alignment position 51 in Fig 2.3, in 20% of cases, (b) in G(E/K), E is majority, but K is not found in our dataset and D can be also found (no G at all) at position 58, (c) in L(R), R is majority at position 59, but in a few cases L is found, like for classical V_H, and finally (d) W(G/L) alignment position 61 is explicit, since G and L represent 50% of the residues while F represents the rest, with V_H typically another aromatic W. It is the same thing with L11S, where S is found in 1/3 of the cases and the rest is L –supposedly not typical.

A second kind of unique characteristics are those of the Cystine residues. The conserved Cys23 (FR1)/Cys104 (FR3) disulphide bond is observed in all the variable regions irrespective of the type of chain or species (highlighted in green in Figs. S1). Some of the camelid V_HH sequences are known to possess an additional disulphide bond between CDR3 or CDR1, CDR3 or CDR2/FR2. In sequences from llamas and camels (see Figs. S3, S4) the presence of an extra cysteine in CDR3 is compensated by another cysteine in CDR1 or CDR2 or CDR2/FR2 boundary. In case of V_HH from llama, 8 out of 68 structures from PDB show this signature, whereas in camels this number is slightly higher, with 15 out of 34 structures from PDB. In the case of V_HH sequences from alpacas, this extra cysteine bond is formed between cysteines of CDR3 and FR2/CDR2 boundary with 3 out of 11 structures from PDB (see Figs. S5).

Structural analysis of V_HH domain

As mentioned previously, V_HH adopts the immunoglobulin fold formed by the anti-parallel arrangement of 9 β-stands connected by loops. Each V_HH domain has three CDR loops. The fold is held together by the inter strand hydrogen bonds and disulphide bond(s). The conserved disulphide bond is between Cys23 of FR1 and Cys104 of FR3 (IMGT numbering). In few cases the presence of a second disulphide bond is also observed, which

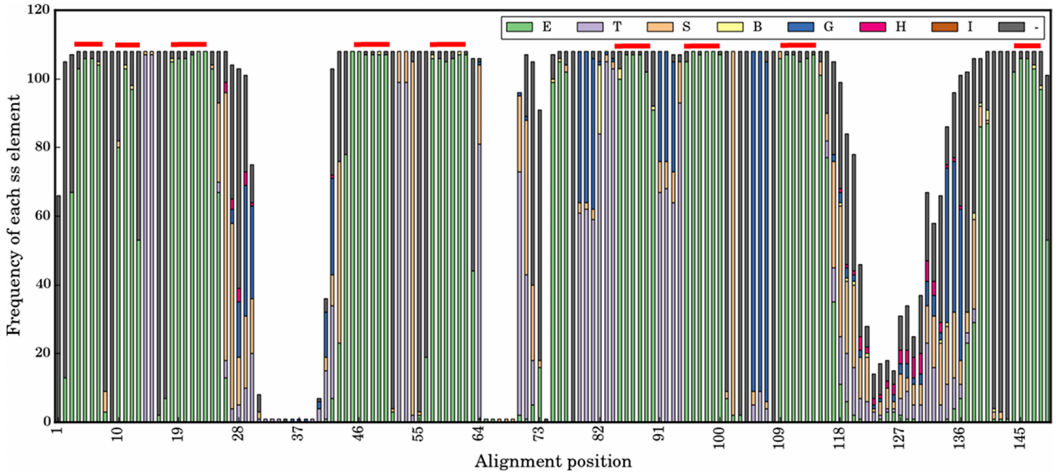

Figure 3 Secondary structural profile of the 105 V_HH domains. The eight classes of DSSP elements are colour coded in the Figure legend. The symbols are E for extended conformation, (b-strand from b-sheet, T for hydrogen bond turn, S for bend, B for isolated b-bridge, G for 3_{10}-helix, H for a-helix, I for p-helix, and '-.' for coils. The different β-strands that form the FRs are indicated above in red: from left to right are the β-strands A, A′, B, C, C′, C″, D, E, and F. The *X*-axis represents the numbering in the alignment.
Full-size DOI: 10.7717/peerj.8408/fig-3

is known to increase the overall stability in the respective V_HH structures (*Zabetakis et al., 2014*). A non-redundant set of 105 V_HH domains was assessed for structural domain similarity, using RMSD. The median RMSD value for the whole domain was at 2.63 Å (Figs. S9A–S9H), all the FRs showed a median value <0.8 Å (Figs. S9E to S9H), while they rose to 2.01, 1.56 and 3.51 Å for CDR1, CDR2 and CDR3 respectively (see Figs. S9B to S9D).

Secondary structure analysis of V_HH domain

At a more local conformational level, Fig. 3 is a representation of the secondary structures when aligned as per sequence-aligned positions. The nine conserved β-strands (in alignment positions 3–7, 10–13, 18–26, 44–50, 58–62, 85–90, 95–99, 109–119, 144–148) can easily be observed. The CDR regions represented in the alignment at positions 29-41, 64-75 and 116-139, are mainly associated with turns, bends and coil secondary structure elements (SSEs). Interestingly, the connecting loops between some β-strands, encompassed in FRs, are composed of turns, bends and helical SSEs, i.e., positions 13–17 in FR1, 51–59 in FR2, 78–84, 91–94 and 100–108 in the FR3 region. This suggests some of them have different backbone constraints compared to the rest. The middle region of FR4 is an interesting region with almost all coils, which are the irregular SSEs for alignment positions 140–143.

The noteworthy point in the analysis of the different canonical β-strands that are in FRs is that some are close to pure β-strands such as β-strand F (for which it is 99%), but for most of them some positions supposed to be only β-strands are not. They are mainly N-terminus first residue(s). For instance, the first position of β-strand A1 is only 70% for β-strand; it is 80% for β-strand C or A2. The most striking case is the C-terminus of β-strand F for which many residues are either β-strand or β-turns, i.e., clearly different types of local

conformations. These results (i) are in concordance with previous analyses using PBs (*Noël, Malpertuy & de Brevern, 2016*), and (ii) show that FRs are not as conserved in terms of secondary structure, i.e., this finding can so have a direct impact on the proposition of structural models.

Analysis of disulphide bond geometry in V_HH domains

An additional disulphide bridge is present in some V_HH domains along with the conserved-canonical disulphide bond between Cys 23 and Cys 104 (IMGT numbering system). These additional disulphide bridges were observed between Cysteine residues found (i) CDR1 and CDR3, namely type 1, (ii) FR2 and CDR3, namely type 2 and (iii) both in CDR3 (see Figs. S10 to S10C for schematic representation). In our initial dataset, 25 V_HH domains were observed to have the type 1 (12 V_HH domains) and type 2 (13 V_HH domains) additional disulphide bridges. Figure 4 illustrates their position in V_HH structure, with both types involved in bending CDR3 onto the β-sheet surface.

Using a recent classification of disulphide bridges (*Schmidt, Ho & Hogg, 2006*) (see Material and Methods section and Figs. S11 for schematic representation) we have analysed the geometry of disulphide bonds in subset of V_HH structures which contain two disulphide bonds. The sign pattern of the canonical and the additional disulphide bonds of both types have been characterized (see Table S3). The greater the number of negative values a dihedral angle pattern has, the lower is the strain energy, suggestive of a stable disulphide bond. Although, the canonical disulphide bond is conserved in terms of position of cysteines, the sign pattern of the dihedral angles varies much more compared to the additional disulphide bond which is surprising. Interestingly, most of the additional disulphide bonds all have negative sign patterns, suggesting a lower energy bond which might increase stability. Two V_HH domains (PDB id 1YC8 from type 1 and 4Y7M from type 2 class in Table S3) have the dihedral angle pattern characteristic of allosteric disulphide bonds according to the classification. This analysis shows that both canonical and the additional disulphide bridges do not play a simple structural role, as many are not favourable, and can be or are associated to functional roles of V_HH domains.

Analysis of local conformations of CDR1 and CDR2

The CDR1 region is defined between the first conserved cysteine (Cys 22) and the conserved tryptophan (Trp 41). The recent PyIgClassify database has 30 clusters of CDRH1 and 19 of these clusters have V_HH structures. Out of the 19 clusters, 13 are very sparsely populated in V_HH structures (<10 structures). Clusters H1-13-1, H1-13-3, and H1-13-5 were analysed, as they were associated to a correct number of occurrences (>20 structures). PB analyses of these clusters show variations in PB assignment in the CDR1 region (see PB frequency maps in Figs. 5A to 4C). The residue region from 21 to 33 is the CDR1 according to PyIgClassify. The PB motif includes 3 residues flanking the CDR1, representing the transition from the β-strand region represented by PB *d* to loop and back to β-strand. The first cluster (H1-13-1, see Fig. 5A) has 21 out of 37 (57%) structures sharing a strict common PB signature *ddddehiafklpccdddddd*; the remaining structures of this cluster are close to this PB series as shown with low N_{eq} (see Fig. 5G). The highest N_{eq} value for H1-13-1 cluster is

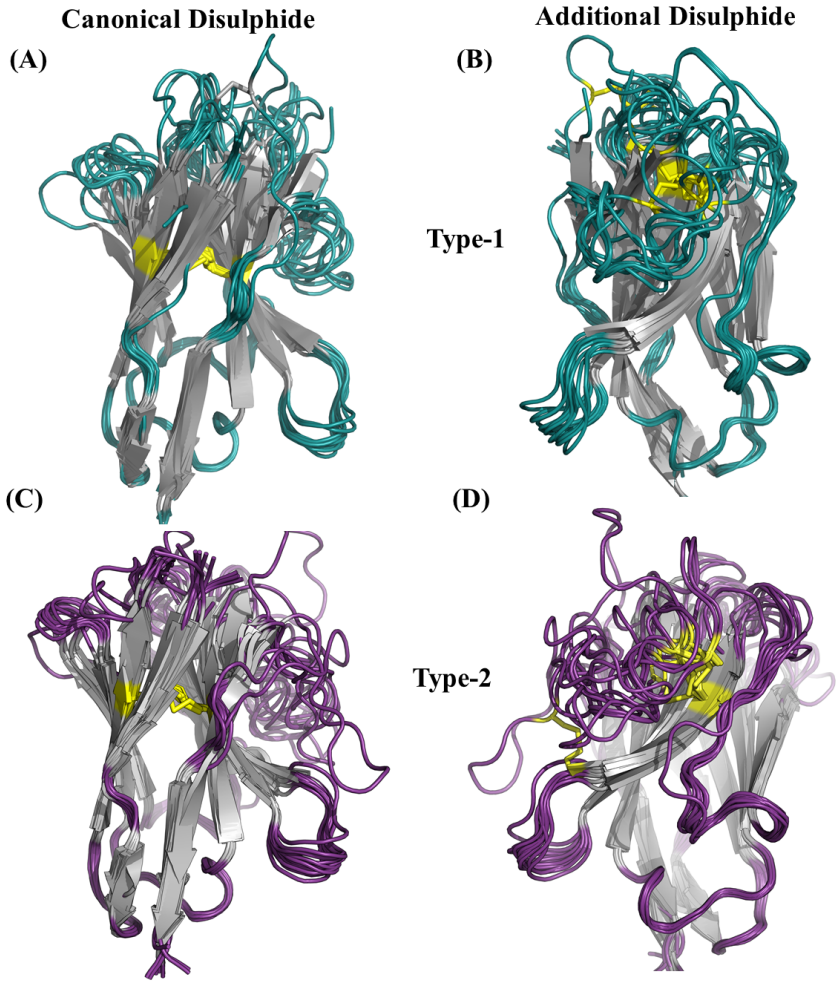

Figure 4 Disulphide bridges of V_HH domains with additional disulphide bonds. (A) Conserved disulphide bond in type 1 V_HH domains (B) Non conserved disulphide bonds in type 1 V_HH domains, (C) Conserved disulphide bond in type 2 V_HH domains and (D) non-conserved disulphide bonds in type 2 V_HH domains. The disulphide bridges are indicated in yellow connecting any two cysteines.
Full-size DOI: 10.7717/peerj.8408/fig-4

not more than 2 for residue position 32, with PB *p* either changing to *c* or *g*. In the other two clusters no strict common PB signature was observed. The third cluster (H1-13-5, see Fig. 5C) has the highest N_{eq} value. The PBs series of the first cluster *dehiafkl* is replaced by *dehjfklp* — a striking difference. The observed structural words in the first cluster PBs *dehia*, *fklpc* and *cdddd* are highly recurrent and reported to have an RMSD below 1 Å, while in the third cluster, the transitions are from the most to the less frequent ones (see the most frequent PB series in *de Brevern et al., 2002*). As seen in Figs. 5B and 5G, the cluster H1-13-3 has the highest N_{eq} values and does not show specific PB series like the two other clusters. We can notice small tendencies towards α helical PBs like PB *m*. A similar in-depth analysis was performed for three CDR2 clusters that also show heterogeneity in the three clusters (see Figs. 5D to 5E and Data S1). In summary, the analyses of CDR1 and CDR2 V_HH in light of PyIgClassify clusters using PBs show a large diversity.

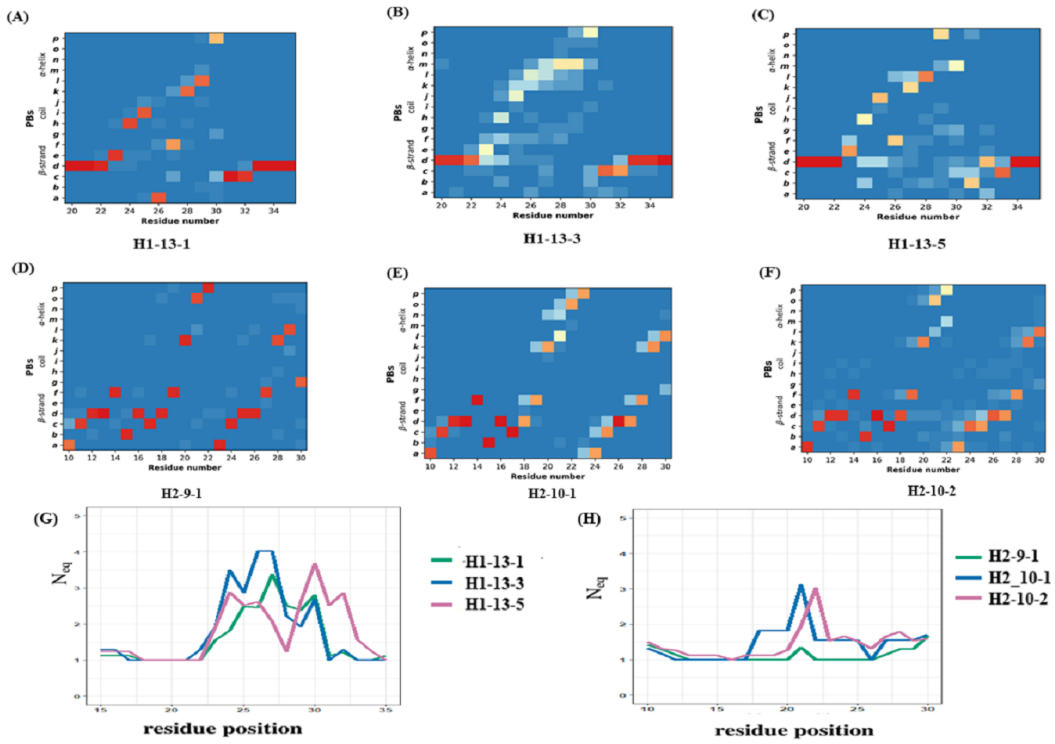

Figure 5 **Local conformational analysis of CDR clusters from PyIgClassify.** PB maps of CDR H1 region from V_HH sequences from CDR H1 clusters (A) H1-13-1, (B) H1-13-3, and (C) H1-13-5, CDR H2 region from V_HH sequences from (D) H2-9-1, (E) H2-10-1, and (F) H2-10-2 cluster, and N_{eq} of (G) three CDR H1 clusters and (H) three CDR H2 clusters. The numbering of residues in each plot is according to the IMGT numbering system.

Full-size DOI: 10.7717/peerj.8408/fig-5

Inter-cluster comparison

The previous analyses showed the variations within a cluster. Using Δ PB and ΔN_{eq}, it is possible to compare the clusters directly (see 'Materials and Methods' section and Fig. S12 to S12D). ΔPB profile allows a detailed description of the conformational diversity between any two sets of protein structural regions. For H1-13-1 and H1-13-5, Δ PB is always higher than 1.5, showing a completely different PB series occurrence. Hence, these two clusters sample two different conformational spaces.

For H2-9-1, H2-10-1 and H2-10-2, ΔPB shows that PB series are closely related in the N-terminal regions between H2-9-1 and H2-10-2, and in C-ter for H2-10-1 and H2-10-2, leading to the idea of composite series with a common PB at position 26 (all ΔPB values of 0.1), information that cannot be provided using RMSD quantification. This analysis shows the relevancy of PBs to compare V_HH structures and V_HH structural models in the following sections. It allows a precise comparison of V_HH structures, which can be considered as highly similar but are not identical, and which quantify this local distance.

V_HH structure prediction survey

V_HH structures are essential to understand binding with partners. Molecular modelling is a simple idea to propose structural models. We performed a complete survey of the V_HH

structure prediction studies published so far (see Table 1). Two main methodologies have been used. (i) generic homology modelling like Modeller and threading approaches like i-TASSER, and (ii) hybrid modelling where FRs and CDRs are modelled separately and then assembled together, like RosettaAntibody (*Sircar, Kim & Gray, 2009*), ABodyBuilder (*Dunbar et al., 2014*; *Leem et al., 2016*), or BioLuminate from Schrödinger® (*Beard et al., 2013*; *Zhu et al., 2014*; *Salam et al., 2014*). Of the above-mentioned approaches, Modeller was used in 50% of the cases in the literature.

All the different structure prediction studies are presented in brief in Table 1 and in detail in Supplemental Information 2). We have presented and classified all the 22 studies into three groups based on the method followed; the reader can, if he wishes, go directly to the following section where we test critically an example of V_HH structure prediction.

(a) V_HH structure prediction using Modeller

Modeller is the most popular comparative modelling tool known to the community. It is mainly used as a standalone algorithm, but also through specialised servers like Protein Structure Prediction server, PS^2V^2 (*Chen, Hwang & Yang, 2009*) and EsyPred3D (*Lambert et al., 2002*) or in commercial software like BIOVIA™ (earlier Discovery studio). The standalone Modeller version was used to propose structural models of V_HH in studies against DARC (Duffy antigen Receptor Chemokine) (*Smolarek et al., 2010b*), VEGFR2 (Vascular Endothelial Growth Factor Receptor 2) (*Shahangian et al., 2015*), TNFR1 α (Tumour Necrosis Factor Receptor 1 α) (*Steeland et al., 2015*), PLA_2 (Phospholipase A2) (*Chavanayarn et al., 2012*), BMP4 (Bone morphogenic protein 4), MMP8 (Matrix metalloproteinase 8) (*Demeestere et al., 2016*), PLA_2 toxins *B. jararacussu* Bothropstoxin-I (BthTX-I) and Bothropstoxin-II (BthTX-II) (*Prado et al., 2016*). A histone binding V_HH (*Jullien et al., 2016*) was proposed through BIOVIA™ suite.

(b) Other generic 3D prediction approaches

The fold prediction software Phyre2 was used to model V_HH against Urease (*Hoseinpoor et al., 2014*). The popular i-TASSER web server was used for V_HH designed against PrA (ProteinA) (*Fridy et al., 2015*) and HCV Non-structural protein NS3/4A (*Jittavisutthikul et al., 2015*). Raptor-X, another threading-based server, was used to model V_HH against adenylate cyclase-hemolysin toxin and the repeats in toxin (CyaA-RTX protein) (*Malik et al., 2016*) subdomains (see Supplemental Information 1).

(c) Antibody specific modelling protocols

Rosetta (*Rohl et al., 2004*) is one of the two most successful *de novo* approaches. A specialised suite called RosettaAntibody (*Sircar, Kim & Gray, 2009*) by Gray's group, was the first to put forward a modelling protocol for V_HH. The Rosetta Antibody suite was modified to model single chain V_HH.

Briefly, the templates for FR regions are selected using BLAST against antibody databank containing structures from PDB, and for CDRs the templates are chosen from BLAST bit scores. Once the FRs are modelled, the CDRs are grafted onto the FRs by optimal superposition of the backbone atoms of two overlapping residues at each end of the loop. This study highlights the difficulty in V_HH modelling, and specifically the fact that an

Table 1 Summary of structural modelling studies in chronological order. From left to right, the columns are the names of the authors, year of publication, algorithm of choice for template selection, number of templates used per query, algorithm used for modelling, algorithm used for model validation, algorithm used for model refinement, target/ antigen against which the V_HH in the study is generated, name of the V_HH used in the study, organism from which the respective V_HH is synthesized and PDB ID of the template(s) used in the study.

Authors	Year	Template selection	Templates/query	Modeling methods	Antigen	VHH (main study)	Template(s)
Smolarek et al.	2010	PSI-BLAST against PDB	one	Modeller	DARC- C terminus	CA52	1OP9 and 1 JTO
Govaert et al.	2011	mutation studies	not applicable	Esypred3D,Robetta for mutants	not applicable	cAbAn33, cAbLys3, and cAbPSA-N7	not mentioned
Sircar et al.	2011	BLAST	one or many	Rosetta Antibody VHH suite	not applicable	not applicable	not applicable
Chavanayarn et al.	2012	BLAST	one	not mentioned	Phospholipase 2 of Naga koultia	P3-1, P3_3	1VHP and 1MVF
Thueng-in et al.	2012	BLAST	one	not mentioned	NS5B(RNA dependent RNA polymerase) of hcv	VHH6, VHH24 (clone names)	1VHP, 1F2X
Phalaphola et al.	2013	BLAST	one	not mentioned	NS3-C (HCV helicase)	VH6,VHH9,VH59	1OHQ,1XFP,3BN9
Inoue et al.	2013	not mentioned	one	MOE (CCG)	Hen Egg White Lysozyme	cAb-CA05-(C-C-L), cAb- CA05-(#16-#09-L), cAb-CA05-(#16-#19-L)	1RI8
Hoseinpoor et al.	2014	PSI-BLAST (Phyre2)	one	Phyre2	H.pylori Urease	HMR23	not available
Steeland et al.	2015	not mentioned	multiple	swiss model server	TNF receptor 1	Nb70 and Nb96	4FZE, 4JVP, 3P0G, and 2KH2
Shirin Shahangiana et al.	2015	BLAST against PDB	one	Modelller 9.13	VEGFR	VEvhh1, VEvhh2, VEvhh3	1OP9-A , 1MVF-A and 2X6M-A respectively
Jittavisutthikul et al.	2015	i-TASSER	not mentioned	I-TASSER and ModRefiner	NS3/4A	VHH24, VHH28, VHH41	not mentioned
Unger et al.	2015	BioLuminate suite	multi (different for frs and cdrs of each vhh)	BioLuminate	CDTa/b(clostridium difficile toxin a/b)	1+8, 1-14, 1+18 (3 VHH clones)	numerous
Fridy et al.	2015	based of target-IgG complex PrA1-Fab	one	I-TASSER	not mentioned	LaP-1 to 4	not mentioned
Calpe et al.	2015	BLAST	one	Modeller	Bone Morphogenic factor 4 (BMP)	C4, E7, C8	4BSE, 1SJX
Prado et al.	2016	BLAST against PDB	one	Modeller 9.10	BthTX-I and BthTX-II	KF498607, KF498608, KC329715 and KC329718	4KRP-B, 4DKA-A, 3EZJ-B, 4KRP-B
Jullien et al.	2016	Discovery studio	one	Modeller 9.10 from Accelrys Discovery Studio v. 3.1 (DS 3.1)	Histones 2A and 2B	nabobody against chromatin (Chromatibody)	4IDL
Demeestere et al.	2016	Swiss model web server	multiple	Modeller	Matrix metallo proteinases 8	Nb14	4LAJ, 3EZJ, 3TPK and 4M3J
Se et al.	2016	PS2V2 server	one	Modeller through PS2V2 server	Bap protein Acinetobacter baumanii	VHH1	1MVF
Leem et al.	2016	BLAST	many	Modeller and FREAD (for CDR loop prediction)	not applicable	not applicable	not applicable
Soler et al.	2016	Pre existing knowledge	many	Swiss Model server	lysozyme	NbHuL6, cAbLys, cAbCII10	NbHuL6-3EBA and 3DWT, cAbLys and cAbCII10 1ZMY, 1JTP
Malik et al.	2016	”best fit structures”	one	Raptorx	*CyaA-RTX Segment*	VHH2,VH5,VH18,VHH37	1F2K,4O9H,2KH2,4HEP

approach for IgG is not optimal for V_HH. The scripts specific for V_HH are not available anymore in the newer version of RosettaAntibody, but as suggested by the group, it is possible to model V_HH by submitting a dummy light chain in the protocol and then deleting it from the models once they are modelled (*Sircar et al., 2011*).

The most recent class of antibody-specific modelling approaches is ABodyBuilder. Templates are selected using the sequence identity of the FR regions as a criterion against SAbDab (Structural Antibody Database) (*Dunbar et al., 2014*). Modeller then models FR regions. Next, CDRs are modelled using FREAD, a loop prediction developed by the same group using database search. FREAD basically searches against the CDR database created for each CDR (six in total). The SAbDab is the first database to have text search for querying V_HH; however, the text search also lists modified V_H/V_L which exist as single domain antibodies (*Dunbar et al., 2014*; *Leem et al., 2016*).

Case study

In case of antibodies, the acceptable range of RMSD is different for CDRs and FRs, often leading to global RMSD <3 Å, but it may not be the only near-native structure/conformation possible (*Kufareva & Abagyan, 2012*). Improvement of model quality was detailed by using multiple templates (*Chakravarty et al., 2004*; *Larsson et al., 2008*). Here, we decided to reproduce the study of Steeland and co-workers (*Steeland et al., 2015*) to assess the interest and impact of such an approach. They predicted structural models of V_HH domains (named Nb70 and Nb96) against Tumour Necrosis Factor receptor 1α using multiple templates: (i) Single chain variable fragment (ScFv) from mouse against Interleukin 1 β (PDBID 2KH2 (*Wilkinson et al., 2009*)), named temp-m), (ii) a llama V_HH used to stabilize β2 adrenoceptor (PDBID: 3P0G (*Rasmussen et al., 2011*), referred to as temp-l), (iii) an alpaca V_HH against Hepatitis C virus (HCV) glycoprotein E2 (PDBID: 4JVP (*Tarr et al., 2013*), temp-a), and (iv) a V_H from human F_{ab} which is Antibody-dependent cell-mediated cytotoxicity anti-HIV 1 antibody (PDBID: 4FZE (*Tolbert, Wu & Pazgier, 0000*), henceforth referred to as temp-h). In the following four sections we will present (a) the analysis of sequence relationship, (b) analysis of the structures, (c) the modelling results and finally (d) how the different structural templates have different impacts on the proposition of structural models.

Sequence and structure analysis of templates and query

The query sequence (Nb70) shared a high percent sequence identity (76–70%) with temp-l, temp-m and temp-a (70.6%) and a weak sequence identity (45%) with temp-h (45.3%) (See Table S1A). Hence, three of the structural templates can be considered as good 'classical' templates. Amongst them, the first three templates have 67–61% sequence identity, and only 46–37% with temp-h (see Table S1A).

Analysis of FRs and CDRs provided a slightly different view (see Table S1B). Indeed, temp-h still has the worst sequence identity with Nb70 for all FRs and CDRs, except CDR1 (50% sequence identity) and CDR3 (with a sequence identity of only 16%). For the FRs of the others templates, sequence identity reached 96% for FR1, 77% for FR2, 94% for FR3 and even 100% for FR4. For the CDRs, it is lower but remains good, with 50%, 75%

and 58%, respectively. Another important point is that CDRs 1 and CDR 2 are of same length in all the sequences (see Fig. S13). The CDR3 region of query is 12 residues long, while temp-a and temp-h are longer, with 20 and 18 aa respectively. These results lead us to conclude that comparative modelling is probably the right choice.

Structure analysis of templates

Sequence similarity from the above analysis suggests good conservation, but structural similarity between the templates provides a different view (see Table S1C). Temp-m is the closest to all three templates, with RMSD values mainly from 1.9 to 2.4 Å, while temp-a is further away with a RMSD value of 3.7 Å with temp-l and of 4.5 Å with temp-h. A large difference between structural templates is observed and does not correspond directly to the information seen in the previous section (see Figs. 6A & 6B). The most conspicuous region in the superposed structures is CDR3 for temp-a and temp-h; in the structure of temp-a, the torso of the CDR3 is bent towards the FR2 region due to the presence of additional disulphide bond. Whereas in the case of temp-h, the CDR3 loop, though long, has a protruding tip (upwards). The RMSD values range from 0.7 to 1.5 Å for FR1, 1.7 to 3.1 Å for CDR1, 0.4 to 0.9 Å for FR2, 0.3 to 2.2 Å for CDR2, 0.8 to 1.3 Å for FR3, and 0.4 to 1.0 Å for FR4, while CDR3 had segmented values due to difficulty in superimposition. As expected, the general trend of high RMSD in CDRs compared to FRs is observed in all the pairwise comparisons (see Table S2). It is interesting to note the deviations in FR regions between temp-l, temp-a and temp-h, that might be unexpected.

(c) Analysis of the structural models

The structural similarity between the best-selected model from multiple template modelling (the reference) and the models generated from single template is shown in Figs. 6C to 6F. The best model constructed using temp-m adopts the conformations similar to that of the best model using multi-templates, especially in the CDR3 region. This is surprising, since it shares a higher sequence identity with temp-l in this region, but CDR3 lengths are the same in both the query and temp-m.

Using residue-wise RMSD comparison as quantification (see Fig. S11), the model generated from temp-m is close to the best model generated from multiple templates, even in the CDR3 region (only the CDR2 region has a significantly higher value). The other best models display much higher residue-wise RMSD in all the CDR regions, even ranging 15 Å in CDR3 for the model generated with temp-a.

Interestingly, one of the structural templates provided a major contribution to the final models, although (i) it does not have a stronger sequence identity than the rest, and even, (ii) the only region for which it had the best sequence identity, CDR2, is the only region that is far from the selected model. Thus, the multi-template approach for V_HH appears somewhat complex.

(d) Analysis of local conformational sampling by Modeller

Analyses of best models only provided information for best DOPE score selected models; it does not provide information about potential sampling proposed by the comparative modelling approach. Using PBxplore, Protein Blocks were assigned to each of the models

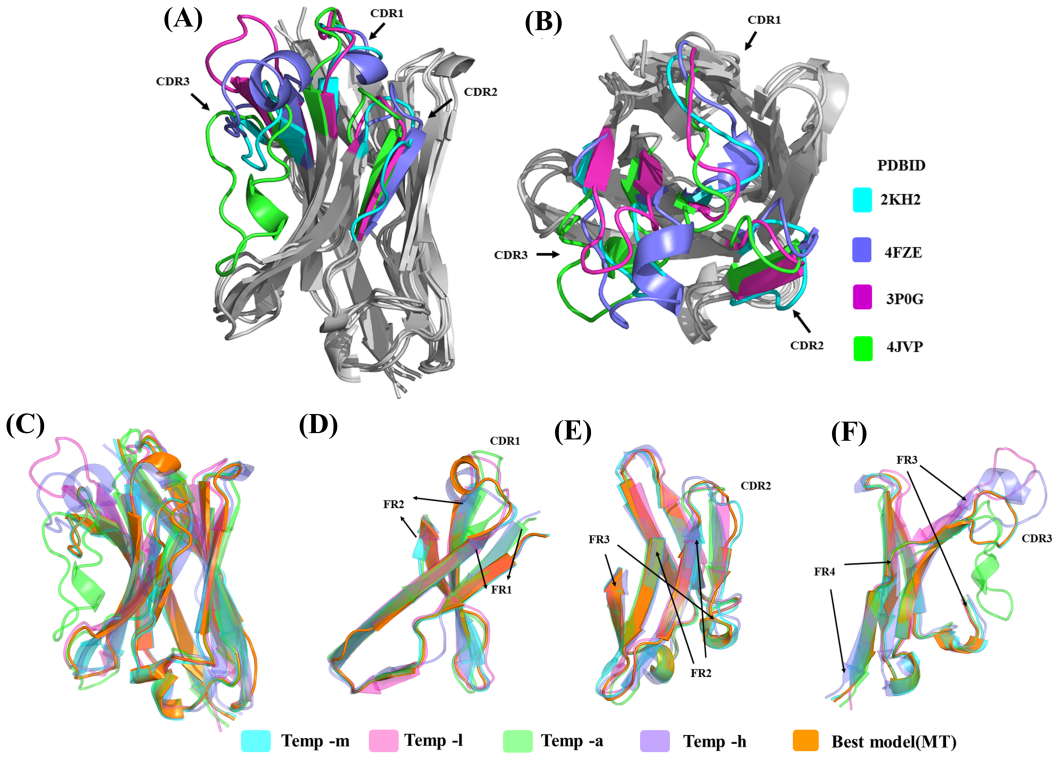

Figure 6 Analyses of structural template and best structural models. First, the four templates Temp -m, Temp -l, Temp -a and Temp -h are superimposed. The colour coded regions in teal, pink, green and violet respectively are the CDRs. (A) Lateral view and (B) top view. Then, best models selected using DOPE score are superimposed. Four templates Temp-m, Temp-l, Temp-a, Temp-h and best structural model colour coded regions in teal, pink, green, violet and orange respectively are shown with different orientations. (A) Global view, (B) zoom on CDR1, FR1 and FR2, (C) on CDR2, FR2, and FR3, and (D) CDR3 with FR3 and FR4.

Full-size DOI: 10.7717/peerj.8408/fig-6

and N_{eq} entropy computed at each position. A summary of N_{eq} values >1 is provided in Table 2; in all the cases, modelling with multiple templates is the one with the highest number of residue positions with N_{eq}>1 (49 residue positions), suggesting the largest conformational diversity. The least number of residue positions with N_{eq}>1 is for models generated using temp-m (23 residue positions). Notably, models generated from temp-h had two residue positions with high N_{eq} values in the CDR3 region. Figure 7A is a position-wise distribution of N_{eq} values in each case. The interesting results are that (i) the N_{eq} of template temp-m is most often low compared to the multi-template case, and, (ii) the models from multiple templates case show maximum N_{eq} in the CDR2 region (see also Supplemental Information 2).

Analysis of PB distribution (see Fig. S15) shows preferred PB '*d* 'for most residues of FRs in β-sheets in all the cases. The three preceding and succeeding residues of CDRs were considered anchor residues in each loop. Surprisingly, no PB variability was observed for CDR1 anchors. In the case of CDR2, only the position 51 from models generated from temp-l showed a slight variation of N_{eq} to 1.22. For CDR3, the preceding anchors showed

Table 2 Distribution of N_{eq} in each modelling scenario.

N_{eq}	Temp-m	Temp-l	Temp-a	Temp-h	All templates
1–2	19	34	35	23	39
2–3	4	1	3	4	6
3–4			3	2	3
4–5			1		
5–6				1	1
6–7				1	

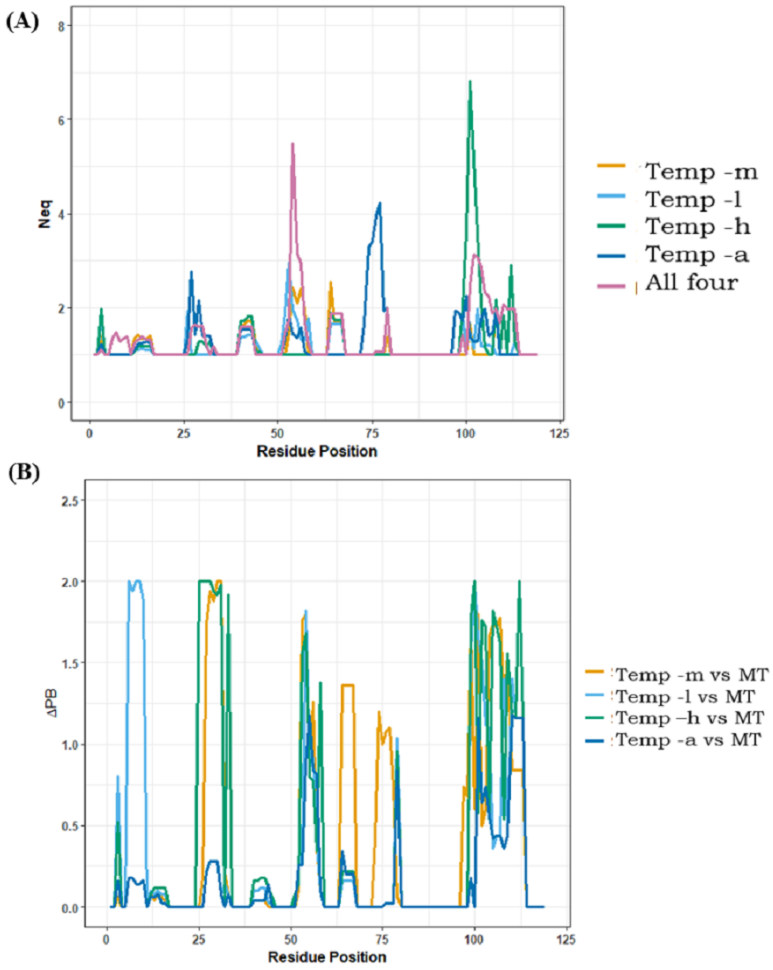

Figure 7 **Positional PB entropy N_{eq} and ΔPB.** (A) All five scenarios of modelling are represented in separate colours. *X*-axis represents residue positions and *Y*-axis represents N_{eq}. (B) ΔPB between multi-template scenario and each mono template scenario.

Full-size DOI: 10.7717/peerj.8408/fig-7

no PB variation; however, the succeeding anchor positions in the modelling scenario of temp-h and temp-a showed variations.

CDRs (aa regions 26–34, 52–58, 97–108), as expected, have more diverse conformations than FRs. Amongst the five scenarios, the most PB diversity in CDR3 is seen in 11 out of 12 residue positions from models generated from temp-a, followed by models generated from temp-l (9/12 residues positions), multiple templates (8/12 residue positions), temp-h (7/12 residue positions) and temp-m (1/12 residue position). This observation suggests that adding more information in the case of multiple templates, and poor template target alignments can cause unexpected conformational diversity.

The local conformational diversity in terms of PBs between mono template(s) and multi-template structural models can be understood by analysing the differences in PBs, quantified by ΔPB. Figure 7B shows ΔPB calculated between multi-template models and each individual template model. Among all four cases of comparison, the case of modelling with temp-m and multi-template shows the least change in PBs at each residue. The differences in local conformations are expected in the CDRs, while the changes in FRs are the ones least expected. These comparisons may also suggest that in case of multi-template modelling, temp-m mostly influenced model conformations due to better alignment quality in CDR regions, and a shared second-best sequence identity with the query. Please note that Nb96 produced roughly the same results, while CDR3 is longer and so more complex to analyse.

DISCUSSION

Analysis and prediction of simple V_HH fold seems a trivial task at first sight. Analysis of the sequence content of FRs and CDRs correlates partially with the recent study of Mitchell and Colwell (*Mitchell & Colwell, 2018a*). They have focused their study on a dataset of V_HH domain in complex with respective antigens and compared to IgGs. At present, it appears obvious that the expected specific amino acid signature of V_HH is not a universal feature in these domains and only depends on the germline; few unmodified V_HH domains lack them. Analysis of V_HH structures also showed that the fold is far from conserved in terms of local protein conformation. Assignment of secondary structures showed some important deviations, irrespective of the bound or unbound state of the V_HH. Similarly, analysis of CDRs in light of PBs showed (i) that actual CDR classification is difficult to apply to V_HH, (ii) use of a global measure in CDR classification tends to associate different types of local protein conformations, i.e., PB series and (iii) in fact, some common PB series can be found in different CDR clusters. Interestingly, the additional disulphide bridges of whatever type are often not strongly favourable, which is slightly counterintuitive at first sight.

A survey of structure prediction studies of V_HH showed that most studies resorted to generic template-based modelling approaches. Analysis of the impact of template conformations on the model(s) generated, with the example of V_HH modelled by Steeland and co-workers (*Steeland et al., 2015*), underlined the difficulty of choosing the template(s). Indeed, three of the four templates have roughly common sequence identity measured

at around 74%, each being slightly more identical depending on the CDRs and FRs, i.e., none can be selected as the best of the best. Nonetheless, each of them had a unique 3D conformation and proposed (a) a different best model, but also (b) sampled different conformations. It is, in fact, surprising that the more dissimilar template (45% sequence identity) does not produce V_HH models distinct from models predicted using other templates. These results clearly indicate the need for more detailed study and approaches to propose optimized methodology for V_HH structure prediction (see also Supplemental Information 3).

Additionally, we have performed a similar analysis with solved V_HH structure to confirm our hypothesis. A V_HH domain (PDBID: 1QD0 (*Spinelli et al., 2000*) was selected as the query and four other V_HH sharing sequence identities between 73.9% and 67.5%, representing a real world case. The four templates' RMSD to the query ranged from 2.01 Å to 4.41 Å. The best structural model obtained using the structural template with highest sequence identity had a 2.1 Å RMSD value. The addition of other templates provided a clear increase to 2.2 Å with two templates, a further increase to 2.6 Å for 4 templates (see Supplementary Analysis 1 and Fig. S16 for more details). The findings corroborate results previously shown in the case study of Nb70 V_HH domain modelling. These examples taken from the literature are case studies and show how the authors proceeded. They could be improved by a better control of structural templates, or additional constraints such as secondary structures, distances or even Protein Blocks. We underlined that one must pay attention to details of templates such as structural similarity between templates when using multiple templates, and the length of CDRs in addition to sequence identity

CONCLUSION

Our paper in addition to reaffirming the sequences-structure characteristics of V_HH domains reported in the recent literature, also makes unique observations regarding (i) the variation in the amino acid signature in the FR2 region, (ii) conservation of β-strands and presence of other kinds of secondary structures, (iii) sheds light on conformations of di sulphide bridges and (iv) inter and intra cluster variations from PyIgclassify CDR clusters in terms of local conformations using Protein Blocks. All the above analyses might help the community to appreciate the topological enigma of these domains. As Protein Blocks were able to identify many fine variations in the well accepted CDR classification and FRs in our previous paper, we used it to analyse local conformations of modelled structures of V_HH domain from a study. The variations in local conformations in models are influenced template quality and conformations. In most cases of variable domain comparative modelling, the templates for modelling FRs are selected based only on sequence identity. In the specific case study that we chose, more complexity arose due to the usage of multiple templates which were chosen based on BLAST searches. We intend to draw the attention of the research community by a precise analysis of the models from this exercise to the influences of template in terms of CDR3 length and sequence identity and Protein Blocks sequences. The latter is an innovative approach developed in our lab to shed light on local conformational changes. This exercise suggests that the selected templates might not be

the best possible templates when chosen using FR sequence identity. Finally, the use of multiple templates when the templates are overlapping more restraints that may not be desirable.

ACKNOWLEDGEMENTS

We would like to thank Catherine Etchebest for fruitful discussions.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by grants from the Ministry of Research (France), University Paris Diderot, Sorbonne, Paris Cité (France), University of La Réunion, Réunion Island, National Institute for Blood Transfusion (INTS, France), National Institute for Health and Medical Research (INSERM, France) and labex GR-Ex. The labex GR-Ex, reference ANR-11-LABX-0051 is funded by the program ''Investissements d'avenir'' of the French National Research Agency, reference ANR-11-IDEX-0005-02. Tarun J. Narwani and Alexandre G de Brevern received a collaborative grant (number 5302-2) from Indo-French Centre for the Promotion of Advanced Research/CEFIPRA. Nicolas K. Shinada is supported by Discngine, Paris, France and ANRT, France. Akhila Melarkode Vattekatte is supported by Allocation de Recherche Réunion granted by the Conseil Régional de la Réunion and the European Social Fund EU (ESF). The authors were granted access to high performance computing (HPC) resources at the French National Computing Centre CINES under grant no. A0010707621 and A0040710426 funded by the GENCI (Grand Equipement National de Calcul Intensif). Calculations were also performed on an SGI cluster granted by Conseil Régional Ile de France and INTS (SESAME Grant). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Ministry of Research (France).
University Paris Diderot, Sorbonne, Paris Cité (France).
University of La Réunion, Réunion Island.
National Institute for Blood Transfusion (INTS, France).
National Institute for Health and Medical Research (INSERM, France).
labex GR-Ex: ANR-11-LABX-0051.
French National Research Agency: ANR-11-IDEX-0005-02.
Indo-French Centre for the Promotion of Advanced Research/CEFIPRA: 5302-2.
Discngine, Paris, France and ANRT, France.
Conseil Régional de la Réunion.
The European Social Fund EU (ESF).
French National Computing Centre CINES: A0010707621.
GENCI: A0040710426.

Conseil Régional Ile de France.
INTS (SESAME Grant).

Competing Interests

Frederic Cadet is associated with PEACCEL, Paris, France. Jean-Christophe Gelly and Alexandre G. de Brevern are associated with IBL, Paris, France. Jean-Philippe Meyneil is employed by ISoft, Paris, France. Alain Malpertuy is employed by Atragene, Paris, France. Nicolas K. Shinada is sponsored by Discngine, Paris, France and ANRT, France. All other authors declare no competing interests.

Author Contributions

- Akhila Melarkode Vattekatte conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Nicolas Ken Shinada, Tarun J. Narwani and Floriane Noël performed the experiments, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Olivier Bertrand analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Jean-Philippe Meyniel, Jean-Christophe Gelly and Frédéric Cadet conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Alain Malpertuy conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Alexandre G. de Brevern conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The sequences of all VHH used in the study are available as Dataset S1.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.8408#supplemental-information.

REFERENCES

Ablynx. 2016. Ablynx 2016 Annual report.

Adolf-Bryfogle J, Xu Q, North B, Lehmann A, Dunbrack Jr RL. 2015. PyIgClassify: a database of antibody CDR structural classifications. *Nucleic Acids Research* **43**:D432–D438 DOI 10.1093/nar/gku1106.

Al-Lazikani B, Lesk AM, Chothia C. 1997. Standard conformations for the canonical structures of immunoglobulins 1 1Edited by I. A. Wilson. *Journal of Molecular Biology* **273**:927–948 DOI 10.1006/jmbi.1997.1354.

Barnoud J, Santuz H, Craveur P, Joseph AP, Jallu V, de Brevern AG, Poulain P. 2017. PBxplore: a tool to analyze local protein structure and deformability with protein blocks. *PeerJ* **5**:e4013 DOI 10.7717/peerj.4013.

Beard H, Cholleti A, Pearlman D, Sherman W, Loving KA. 2013. Applying physics-based scoring to calculate free energies of binding for single amino acid mutations in protein-protein complexes. *PLOS ONE* **8**:e82849 DOI 10.1371/journal.pone.0082849.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2005. GenBank. *Nucleic Acids Research* **33**:D34–D48 DOI 10.1093/nar/gki063.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. *Nucleic Acids Research* **28**:235–242 DOI 10.1093/NAR/28.1.235.

Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios I. 2016. UniProtKB/swiss-prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. In: *Methods in Molecular Biology*. 23–54 DOI 10.1007/978-1-4939-3167-5_2.

Chakravarty S, Sanchez R, Sali A, Heintz N, Orengo CA, Madhusudhan MS, Mirkovic N, Sali A, Rost B, Yerkovich B. 2004. Systematic analysis of added-value in simple comparative models of protein structure. *Structure* **12**:1461–1470 DOI 10.1016/j.str.2004.05.018.

Chavanayarn C, Thanongsaksrikul J, Thueng-in K, Bangphoomi K, Sookrung N, Chaicumpa W. 2012. Humanized-single domain antibodies (VH/VHH) that bound specifically to naja kaouthia phospholipase A2 and neutralized the enzymatic activity. *Toxins* **4**:554–567 DOI 10.3390/toxins4070554.

Chen CC, Hwang JK, Yang JM. 2009. (PS)2-v2: template-based protein structure prediction server. *BMC Bioinformatics* **10**:366 DOI 10.1186/1471-2105-10-366.

Chothia C, Lesk AM. 1987. Canonical structures for the hypervariable regions of immunoglobulins. *Journal of Molecular Biology* **196**:901–917 DOI 10.1016/0022-2836(87)90412-8.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Research* **14**:1188–1190 DOI 10.1101/gr.849004.

de Brevern AG. 2005. New assessment of a structural alphabet. *In Silico Biology* **5**:283–289.

de Brevern AG, Etchebest C, Hazout S. 2000. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins: Structure, Function and Genetics* **41**:271–287 DOI 10.1002/1097-0134(20001115)41:3<271::AID-PROT10>3.0.CO;2-Z.

de Brevern AG, Valadié H, Hazout S, Etchebest C. 2002. Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship. *Protein Science* **11(12)**:2871–2886 DOI 10.1110/ps.0220502.

de Brevern AG, Wong H, Tournamille C, Colin Y, Le Van Kim C, Etchebest C. 2005. A structural model of a seven-transmembrane helix receptor: the Duffy antigen/receptor for chemokine (DARC). *Biochimica et Biophysica Acta* **1724**:288–306 DOI 10.1016/j.bbagen.2005.05.016.

Decanniere K, Desmyter A, Lauwereys M, Ghahroudi MA, Muyldermans S, Wyns L. 1999. A single-domain antibody fragment in complex with RNase a: non-canonical loop structures and nanomolar affinity using two CDR loops. *Structure* **7**:361–370 DOI 10.1016/S0969-2126(99)80049-5.

Demeestere D, Dejonckheere E, Steeland S, Hulpiau P, Haustraete J, Devoogdt N, Wichert R, Becker-Pauly C, Van Wonterghem E, Dewaele S, Van Imschoot G, Aerts J, Arckens L, Saeys Y, Libert C, Vandenbroucke RE. 2016. Development and validation of a small single-domain antibody that effectively inhibits matrix metalloproteinase 8. *Molecular Therapy* **24**:890–902 DOI 10.1038/mt.2016.2.

Deschacht N, De Groeve K, Vincke C, Raes G, De Baetselier P, Muyldermans S. 2010. A novel promiscuous class of camelid single-domain antibody contributes to the antigen-binding repertoire. *The Journal of Immunology* **184(10)**:5696–5704 DOI 10.4049/jimmunol.0903722.

Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, Shi J, Deane CM. 2014. SAbDab: the structural antibody database. *Nucleic Acids Research* **42**:D1140–D1146 DOI 10.1093/nar/gkt1043.

Fasnacht M, Butenhof K, Goupil-Lamy A, Hernandez-Guzman F, Huang H, Yan L. 2014. Automated antibody structure prediction using Accelrys tools: results and best practices. *Proteins: Structure, Function, and Bioinformatics* **82**:1583–1598 DOI 10.1002/prot.24604.

Frey BJ, Dueck D. 2007. Clustering by passing messages between data points. *Science* **315**:972–976 DOI 10.1126/science.1136800.

Fridy PC, Thompson MK, Ketaren NE, Rout MP. 2015. Engineered high-affinity nanobodies recognizing staphylococcal Protein A and suitable for native isolation of protein complexes. *Analytical Biochemistry* **477**:92–94 DOI 10.1016/j.ab.2015.02.013.

Gelly J-C, Joseph AP, Srinivasan N, de Brevern AG. 2011. iPBA: a tool for protein structure comparison using sequence alignment strategies. *Nucleic Acids Research* **39**:W18–W23 DOI 10.1093/nar/gkr333.

Hamers-Casterman C, Atarhouch T, Muyldermans S, Robinson G, Hammers C, Songa EB, Bendahman N, Hammers R. 1993. Naturally occurring antibodies devoid of light chains. *Nature* **363**:446–448 DOI 10.1038/363446a0.

Honegger A, Plückthun A. 2001. Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *Journal of Molecular Biology* **309**:657–670 DOI 10.1006/JMBI.2001.4662.

Hoseinpoor R, Mousavi Gargari SL, Rasooli I, Rajabibazl M, Shahi B. 2014. Functional mutations in and characterization of VHH against Helicobacter pylori urease. *Applied Biochemistry and Biotechnology* **172**:3079–3091 DOI 10.1007/s12010-014-0750-4.

Jittavisutthikul S, Thanongsaksrikul J, Thueng-in K, Chulanetra M, Srimanote P, Seesuay W, Malik A, Chaicumpa W. 2015. Humanized-VHH transbodies that inhibit HCV protease and replication. *Viruses* **7**:2030–2056 DOI 10.3390/v7042030.

Joseph AP, Agarwal G, Mahajan S, Gelly J-C, Swapna LS, Offmann B, Cadet F, Bornot A, Tyagi M, Valadié H, Schneider B, Etchebest C, Srinivasan N, de Brevern

AG. 2010. A short survey on protein blocks. *Biophysical Reviews* **2**:137–145 DOI 10.1007/s12551-010-0036-1.

Joseph AP, Srinivasan N, de Brevern AG. 2012. Progressive structure-based alignment of homologous proteins: adopting sequence comparison strategies. *Biochimie* **94**:2025–2034 DOI 10.1016/j.biochi.2012.05.028.

Jullien D, Vignard J, Fedor Y, Béry N, Olichon A, Crozatier M, Erard M, Cassard H, Ducommun B, Salles B, Mirey G. 2016. Chromatibody, a novel non-invasive molecular tool to explore and manipulate chromatin in living cells. *Journal of Cell Science* **129**:2673–2683 DOI 10.1242/jcs.183103.

Kufareva I, Abagyan R. 2012. Methods of protein structure comparison. *Methods in Molecular Biology* **857**:231–257 DOI 10.1007/978-1-61779-588-6_10.

Lambert C, Leonard N, De Bolle X, Depiereux E. 2002. ESyPred3D: prediction of proteins 3D structures. *Bioinformatics* **18**:1250–1256 DOI 10.1093/bioinformatics/18.9.1250.

Larsson P, Wallner B, Lindahl E, Elofsson A. 2008. Using multiple templates to improve quality of homology models in automated homology modeling. *Protein Science* **17**:990–1002 DOI 10.1110/ps.073344908.

Leem J, Dunbar J, Georges G, Shi J, Deane CM. 2016. ABodyBuilder: automated antibody structure prediction with data–driven accuracy estimation. *mAbs* **8**:1259–1268 DOI 10.1080/19420862.2016.1205773.

Malik A, Imtong C, Sookrung N, Katzenmeier G, Chaicumpa W, Angsuthanasombat C. 2016. Structural characterization of humanized nanobodies with neutralizing activity against the bordetella pertussis CyaA-hemolysin: implications for a potential epitope of toxin-protective antigen. *Toxins* **8**:99 DOI 10.3390/toxins8040099.

Marcatili P, Rosi A, Tramontano A. 2008. PIGS: automatic prediction of antibody structures. *Bioinformatics* **24**:1953–1954 DOI 10.1093/bioinformatics/btn341.

Martin ACR, Thornton JM. 1996. Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *Journal of Molecular Biology* **263**:800–815 DOI 10.1006/jmbi.1996.0617.

Massa S, Xavier C, De Vos J, Caveliers V, Lahoutte T, Muyldermans S, Devoogdt N. 2014. Site-specific labeling of cysteine-tagged camelid single-domain antibody-fragments for use in molecular imaging. *Bioconjugate Chemistry* **25**:979–988 DOI 10.1021/bc500111t.

McLachlan AD, IUCr. 1982. Rapid comparison of protein structures. *Acta Crystallographica Section A* **38**:871–873 DOI 10.1107/S0567739482001806.

Melo F, Sali A. 2007. Fold assessment for comparative protein structure modeling. *Protein Science* **16**:2412–2426 DOI 10.1110/ps.072895107.

Mitchell LS, Colwell LJ. 2018a. Comparative analysis of nanobody sequence and structure data. *Proteins* **86(7)**:697–706 DOI 10.1002/prot.25497.

Mitchell LS, Colwell LJ. 2018b. Analysis of nanobody paratopes reveals greater diversity than classical antibodies. *Protein Engineering, Design and Selection* **31**:267–275 DOI 10.1093/protein/gzy017.

Muyldermans S. 2013. Nanobodies: natural single-domain antibodies. *Annual Review of Biochemistry* **82**:775–797 DOI 10.1146/annurev-biochem-063011-092449.

Muyldermans S, Atarhouch T, Saldanha J, Barbosa JA, Hamers R. 1994. Sequence and structure of VH domain from naturally occurring camel heavy chain immunoglobulins lacking light chains. *Protein Engineering* **7**:1129–1135 DOI 10.1093/protein/7.9.1129.

Muyldermans S, Baral TN, Retamozzo VC, De Baetselier P, De Genst E, Kinne J, Leonhardt H, Magez S, Nguyen VK, Revets H, Rothbauer U, Stijlemans B, Tillib S, Wernery U, Wyns L, Hassanzadeh-Ghassabeh G, Saerens D. 2009. Camelid immunoglobulins and nanobody technology. *Veterinary Immunology and Immunopathology* **128**:178–183 DOI 10.1016/j.vetimm.2008.10.299.

Nabuurs RJA, Rutgers KS, Welling MM, Metaxas A, De Backer ME, Rotman M, Bacskai BJ, Van Buchem MA, Van der Maarel SM, Van der Weerd L. 2012. In vivo detection of amyloid-β deposits using heavy chain antibody fragments in a transgenic mouse model for alzheimer's disease. *PLOS ONE* **7**:e38284 DOI 10.1371/journal.pone.0038284.

Noël F, Malpertuy A, de Brevern AG. 2016. Global analysis of VHHs framework regions with a structural alphabet. *Biochimie* **131**:11–19 DOI 10.1016/j.biochi.2016.09.005.

North B, Lehmann A, Dunbrack RL. 2011. A new clustering of antibody CDR loop conformations. *Journal of Molecular Biology* **406**:228–256 DOI 10.1016/j.jmb.2010.10.030.

Peyvandi F, Scully M, Kremer Hovinga JA, Cataland S, Knöbl P, Wu H, Artoni A, Westwood J-P, Mansouri Taleghani M, Jilma B, Callewaert F, Ulrichts H, Duby C, Tersago D. 2016. Caplacizumab for acquired thrombotic thrombocytopenic purpura. *New England Journal of Medicine* **374**:511–522 DOI 10.1056/NEJMoa1505533.

Prado NDR, Pereira SS, Da Silva MP, Morais MSS, Kayano AM, Moreira-Dill LS, Luiz MB, Zanchi FB, Fuly AL, Huacca MEF, Fernandes CF, Calderon LA, Zuliani JP, Da Silva LHP, Soares AM, Stabeli RG, Fernandes CFC. 2016. Inhibition of the myotoxicity induced by Bothrops jararacussu venom and isolated phospholipases A 2 by specific camelid single-domain antibody fragments. *PLOS ONE* **11**:e0151363 DOI 10.1371/journal.pone.0151363.

Rashidian M, Keliher EJ, Bilate AM, Duarte JN, Wojtkiewicz GR, Jacobsen JT, Cragnolini J, Swee LK, Victora GD, Weissleder R, Ploegh HL. 2015. Noninvasive imaging of immune responses. *Proceedings of the National Academy of Sciences of the United States of America* **112**:6146–6151 DOI 10.1073/pnas.1502609112.

Rasmussen SGF, Choi H-J, Fung JJ, Pardon E, Casarosa P, Chae PS, Devree BT, Rosenbaum DM, Thian FS, Kobilka TS, Schnapp A, Konetzki I, Sunahara RK, Gellman SH, Pautsch A, Steyaert J, Weis WI, Kobilka BK. 2011. Structure of a nanobody-stabilized active state of the β(2) adrenoceptor. *Nature* **469**:175–180 DOI 10.1038/nature09648.

Rohl CA, Strauss CEM, Misura KMS, Baker D. 2004. Protein structure prediction using rosetta. *Methods in Enzymology* **383**:66–93 DOI 10.1016/S0076-6879(04)83004-0.

Rutgers KS, Nabuurs RJA, Van den Berg SAA, Schenk GJ, Rotman M, Verrips CT, Van Duinen SG, Maat-Schieman ML, Van Buchem MA, De Boer AG, Van der

Maarel SM. 2011. Transmigration of beta amyloid specific heavy chain antibody fragments across the in vitro blood–brain barrier. *Neuroscience* **190**:37–42 DOI 10.1016/j.neuroscience.2011.05.076.

Salam NK, Adzhigirey M, Sherman W, Pearlman DA. 2014. Structure-based approach to the prediction of disulfide bonds in proteins. *Protein Engineering, Design & Selection* **27**:365–374 DOI 10.1093/protein/gzu017.

Schmidt B, Ho L, Hogg PJ. 2006. Allosteric disulfide bonds †. *Biochemistry* **45**:7429–7433 DOI 10.1021/bi0603064.

Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research* **18**:6097–6100 DOI 10.1093/nar/18.20.6097.

Shahangian SS, H. Sajedi R, Hasannia S, Jalili S, Mohammadi M, Taghdir M, Shali A, Mansouri K, Sariri R. 2015. A conformation-based phage-display panning to screen neutralizing anti-VEGF VHHs with VEGFR2 mimicry behavior. *International Journal of Biological Macromolecules* **77**:222–234 DOI 10.1016/j.ijbiomac.2015.02.047.

Shen M-Y, Sali A. 2006. Statistical potential for assessment and prediction of protein structures. *Protein Science* **15**:2507–2524 DOI 10.1110/ps.062416606.

Shirai H, Kidera A, Nakamura H. 1999. H3-rules: identification of CDR-H3 structures in antibodies. *FEBS Letters* **455**:188–197 DOI 10.1016/S0014-5793(99)00821-2.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**:539 DOI 10.1038/msb.2011.75.

Sircar A, Kim ET, Gray JJ. 2009. RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Research* **37**:W474–W479 DOI 10.1093/nar/gkp387.

Sircar A, Sanni KA, Shi J, Gray JJ. 2011. Analysis and modeling of the variable region of camelid single-domain antibodies. *Journal of Immunology* **186**:6357–6367 DOI 10.4049/jimmunol.1100116.

Smolarek D, Bertrand O, Czerwinski M, Colin Y, Etchebest C, de Brevern AG. 2010a. Multiple interests in structural models of DARC transmembrane protein. *Transfusion Clinique et Biologique* **17**:184–196 DOI 10.1016/J.TRACLI.2010.05.003.

Smolarek D, Hattab C, Hassanzadeh-Ghassabeh G, Cochet S, Gutiérrez C, de Brevern AG, Udomsangpetch R, Picot J, Grodecka M, Wasniowska K, Muyldermans S, Colin Y, Le Van Kim C, Czerwinski M, Bertrand O. 2010b. A recombinant dromedary antibody fragment (VHH or nanobody) directed against human Duffy antigen receptor for chemokines. *Cellular and Molecular Life Sciences* **67**:3371–3387 DOI 10.1007/s00018-010-0387-6.

Spinelli S, Frenken LGJ, Hermans P, Verrips T, Brown K, Tegoni M, Cambillau C. 2000. Camelid heavy-chain variable domains provide efficient combining sites to haptens †. *Biochemistry* **39**:1217–1222 DOI 10.1021/bi991830w.

Steeland S, Puimège L, Vandenbroucke RE, Van Hauwermeiren F, Haustraete J, Devoogdt N, Hulpiau P, Leroux-Roels G, Laukens D, Meuleman P, De Vos M, Libert C. 2015. Generation and characterization of small single domain antibodies

inhibiting human tumor necrosis factor receptor 1. *Journal of Biological Chemistry* **290**:4022–4037 DOI 10.1074/jbc.M114.617787.

Stefan E, Cambillau C, Conrath K, Plückthun A. 2002. Biophysical properties of camelid VHH domains compared to those of human VH3 domains †. DOI 10.1021/BI011239A.

Tarr AW, Lafaye P, Meredith L, Damier-Piolle L, Urbanowicz RA, Meola A, Jestin J-L, Brown RJP, McKeating JA, Rey FA, Ball JK, Krey T. 2013. An alpaca nanobody inhibits hepatitis C virus entry and cell-to-cell transmission. *Hepatology* **58**:932–939 DOI 10.1002/hep.26430.

Tolbert WD, Wu X, Pazgier M. Crystal structure of N26_i1 Fab, an ADCC mediating anti-HIV-1 antibody. Worldwide Protein Data Bank. DOI 10.2210/pdb4fze/pdb.

Van der Linden RH, Frenken LG, De Geus B, Harmsen MM, Ruuls RC, Stok W, De Ron L, Wilson S, Davis P, Verrips CT. 1999. Comparison of physical chemical properties of llama VHH antibody fragments and mouse monoclonal antibodies. *Biochimica Et Biophysica Acta* **1431**:37–46 DOI 10.1016/S0167-4838(99)00030-8.

Vu KB, Ghahroudi MA, Wyns L, Muyldermans S. 1997. Comparison of llama V(H) sequences from conventional and heavy chain antibodies. *Molecular Immunology* **34**:1121–1131 DOI 10.1016/S0161-5890(97)00146-6.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**:1189–1191 DOI 10.1093/bioinformatics/btp033.

Wilkinson IC, Hall CJ, Veverka V, Shi JY, Muskett FW, Stephens PE, Taylor RJ, Henry AJ, Carr MD. 2009. High resolution NMR-based model for the structure of a scFv-IL-1beta complex: potential for NMR as a key tool in therapeutic antibody design and development. *The Journal of Biological Chemistry* **284**:31928–31935 DOI 10.1074/jbc.M109.025304.

Yamashita K, Ikeda K, Amada K, Liang S, Tsuchiya Y, Nakamura H, Shirai H, Standley DM. 2014. Kotai antibody builder: automated high-resolution structural modeling of antibodies. *Bioinformatics* **30**:3279–3280 DOI 10.1093/bioinformatics/btu510.

Zabetakis D, Olson MA, Anderson GP, Legler PM, Goldman ER. 2014. Evaluation of disulfide bond position to enhance the thermal stability of a highly stable single domain antibody. *PLOS ONE* **9**:e115405 DOI 10.1371/JOURNAL.PONE.0115405.

Zemla A. 2003. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Research* **31**:3370–3374 DOI 10.1093/nar/gkg571.

Zhu K, Day T, Warshaviak D, Murrett C, Friesner R, Pearlman D. 2014. Antibody structure determination using a combination of homology modeling, energy-based refinement, and loop prediction. *Proteins: Structure, Function, and Bioinformatics* **82**:1646–1655 DOI 10.1002/prot.24551.

Zuo J, Li J, Zhang R, Xu L, Chen H, Jia X, Su Z, Zhao L, Huang X, Xie W. 2017. Institute collection and analysis of nanobodies (iCAN): a comprehensive database and analysis platform for nanobodies. *BMC Genomics* **18**: Article number 797 DOI 10.1186/s12864-017-4204-6.