Global analysis of VHHs framework regions with a structural alphabet

Floriane Noël, Alain Malpertuy, Alexandre de Brevern

Preprint – published in *Biochimie* 2016

Global analysis of VHHs framework regions with a structural alphabet.

Floriane Noel[1,2,3,4,#,+], Alain Malpertuy[5] & Alexandre G. de Brevern[1,2,3,4,*]

[1] INSERM, U 1134, DSIMB, F-75739 Paris, France.
[2] Univ Paris Diderot, Sorbonne Paris Cité, UMR_S 1134, F-75739 Paris, France.
[3] Institut National de la Transfusion Sanguine (INTS), F-75739 Paris, France.
[4] Laboratoire d'Excellence GR-Ex, F-75739 Paris, France.
[5] Atragene, F-94200 Ivry-sur-Seine, France.
[#] present adress : Institut Curie, PSL Research University, INSERM, UMR 932, F-75005, Paris, France.
[+] present adress : Université Paris Sud, Université Paris-Saclay, F-91405 Orsay, France.

Short title: VHH FRs structures

* Corresponding author:
Mailing address: Dr. de Alexandre G. de Brevern, INSERM UMR_S 1134, DSIMB, Université Paris Diderot, Institut National de Transfusion Sanguine (INTS), 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France
e-mail : alexandre.debrevern@univ-paris-diderot.fr

Abstract

The VHHs are antigen-binding region/domain of camelid heavy chain antibodies (HCAb). They have many interesting biotechnological and biomedical properties due to their small size, high solubility and stability, and high affinity and specificity for their antigens. HCAb and classical IgGs are evolutionary related and share a common fold. VHHs are composed of regions considered as constant, called the frameworks (FRs) connected by Complementarity Determining Regions (CDRs), a highly variable region that provide interaction with the epitope. Actually, no systematic structural analyses had been performed on VHH structures despite a significant number of structures. This work is the first study to analyse the structural diversity of FRs of VHHs. Using a structural alphabet that allows approximating the local conformation, we show that each of the four FRs do not have a unique structure but exhibit many structural variant patterns. Moreover, no direct simple link between the local conformational change and amino acid composition can be detected. These results indicate that long-range interactions affect the local conformation of FRs and impact the building of structural models.

Keywords: secondary structure / sequence structure relationship / structural alphabet / antibodies / frameworks.

Introduction

The antibodies (Ab) or immunoglobulins (Ig) are glycoproteins that play a central role in the immune response. They allow the recognition of antigens, the recruitment of cells and stimulation of immune defence mechanisms. They have a similar structure in all vertebrates [1]. These large molecules (~ 150 kDa) are composed of two identical heavy (H) and two identical light (L) chains, linked by disulphide bridges. The type of heavy chain of the antibody determines the type of Ig. The most common Ig is the Ig type G (IgG). These chains are arranged in variable and constant domains. The L chains are composed of a variable domain (VL) and a constant domain (CL). The H chains are composed of a variable domain (VH) and three constant domains (CH1 to CH3). Antibodies have been widely used in biotechnological applications [2]. The number of medical treatments based on the use of such macromolecules (*e.g*. oncology, infectious diseases or against autoimmune diseases) increases greatly [3-5]. However the difficulty of producing them and their costs limit their use.

In camelid family (*camelidae*), which includes camels and llamas, conventional antibodies are found, but in addition these species have particular antibodies where L chain and CH1 domain are missing. These antibodies are called Heavy Chain Antibodies (HCAbs [6]). C-terminal VH region derived from HCAbs are called VHH and Nanobody^{TM}. Interestingly, even without their VL counterparts, the VHHs have an affinity and specificity at least as efficient as IgGs. VHH have also a good stability and solubility that lead to their use for biotechnological and biomedical applications [7].

These VHHs are composed of 4 regions whose sequences and structures are defined as conserved (called *Framework Regions*, *FRs*). In addition, VHHs contain three connecting regions showing high variability both in sequence content and structure conformation. These regions are complementary to the antigen surface and are called *Complementarity Determining Regions* or CDRs). Figure 1 underlines in 3D (see Figure 1A [8-16]) and 2D (see

Figure 1B) similarities of VHHs to conventional antibodies. However, some difference can be noticed in the length of the CDRs and in the FR2 residue composition. The VHHs have many interesting biotechnological properties [17]. VHHs are very small molecules (~ 15 kDa) and are very soluble and highly stable. VHHs have a high specificity and affinity for their target antigen. In addition, it is possible to humanize them by modifying few residues in FR2 [18] without altering their properties. The cloning and production of VHHs are easy to implement. In addition the VHHs are interesting alternative to the use of monoclonal antibodies for therapy as shown by recent studies of their use against the Dengue virus [19], H5N1 influenza [20], viral infection [21], aflatoxins in agro-products [22], head and neck cancers [23], vascular endothelial growth factor implicated in cancers [24], venom therapy [25] and *Plasmodium knowlesi* malaria vector [26]. Phase II clinical trials are currently underway to evaluate the efficacy and toxicity of a therapeutic Nanobody^{TM} in patients with thrombotic thrombocytopenic purpura [27] showing recent promising results [28]. VHHs may also be used as support materials for the crystallography of different proteins [29], as they prevent domain mobility, can bind to interface or cavities, and also stabilize loops [21, 30]. They have been widely used for membrane proteins [31-33] and also as biosensor [34].

The structure of proteins is the support of their interactions. It is essential to have access to the VHH structures when working on their use in biotechnology. In a previous study, we highlighted the difficulty, as well as interest, to obtain a structural model of VHH directed against a specific receptor for chemokines (DARC or Duffy Antigen / Receptor for Chemokines, see [35-36]). In the present study, we focused on the VHH FR regions considered as constant. We have used all structural VHH data available from the Protein DataBank (PDB, [37-38]) to underline this *consistency*. For this purpose, we have used a structural alphabet (the Protein Blocks or BPs, [39]), which allows analysing finely the local protein structure conformations (see [40] for a review). It was used in multiple cases to

analyse difference from particular proteins involved in diseases, e.g. integrins implicated into allo-immunisations [41-43] or a transmembrane receptor implicated into hypogonadism [44]. Our study shows that the FRs regions, previously considered as constant, present (unexpected) variations. We also define structural patterns for the different FRs which will help to improve the 3D structural models as the analyses of paratope / epitope.

Materials and Methods

Data sets. The dataset of protein structures is taken from the Protein DataBank [37-38]. It was selected using key-words search on the PDB website [45]. VHH structures with missing residues in the structure were not taken into account. Analyses were done using an approach developed in PTM-SD [46]. The collected dataset is composed of 114 PDB files, with 160 VHH structures. Duplicated structures were deleted according to only one VHH structure was conserved if identical structures with identical sequences were identified. Then the final dataset of VHH structures is composed of 133 unique complete structures (see Sup Data 1), 70% are from *llama glama*, 4.5% from *vicugna pacos*, rest being from *camelus dromadarius*. 132 structures are X-ray structures (with a mean resolution of 2.1 Å) and one NMR structural model.

Data analysis. Different analyses on VHH sequences and VHH structures were performed. The delineation of the FRs and CDRs was done using multiple alignments generated by the ClustalO software (version 1.1.0) [47]. Visualization of VHHs structures was done with the PyMOL software [48].

Scripts for the different analyses were programmed with Python language and R script language [49]. The VHH structures are superimposed with the PROFIT software [50] and mulPBA webserver [51]. Root mean square deviation (*rmsd*) was computed to compare VHH

entirely or partially; *rmsd* is the square root of the average of distances (in Å) between backbone atoms of a protein structure [52].

Secondary structure. Assignment was performed by using the most classic approach DSSP [53] (CMBI version 4.0). DSSP assign more than four secondary structural states, thus we have reduced them as: α-helix including α, 3_{10} and π- helices, the β-strand containing only the β-sheet, the turn involving the turn assignments and bends, and the coil including the rest of the assignments (β-bridges and coil), as done in previous studies [54]. Default settings had been used for all methods.

The predictions of secondary structure were performed with PSIPRED [55-56] and Jpred 4 [57] software. The accuracy of the prediction is given by the Q_3 score which is the percentage of residues predicted in their right state (α-helix, β-strand or coil).

Protein Blocks description. Protein Blocks (PBs [40]) correspond to a set of 16 local prototypes, labelled from *a* to *p* (e.g., see Figure 1 of [58]), of 5 residues length, clustered based on φ, ϕ dihedral angles description. They were obtained using an unsupervised classifier similar to Kohonen Maps [59] and Hidden Markov Models [60]. The PBs *m* and *d* can be roughly described as prototypes for central α-helix and central β-strand, respectively. PBs *a* through *c* primarily represent the N-cap region of β-strand while PBs *e* and *f* correspond to the C-caps; PBs *g* through *j* are specific to coils, *k* and *l* correspond to the N cap region of α-helix, and PBs *n* through *p* to that of C-caps. This structural alphabet allows a reasonable approximation of local protein 3D structures [39] with an average root mean square deviation (*rmsd*) of 0.42 Å [61]. PBs assignment was performed with PBxplore tool developed under the guidance of Pierre Poulain (https://github.com/pierrepo/PBxplore*, in preparation*). PBs are used to translate the 3D structures under a 1D shape like a sequence of

amino acids and are particularly more accurate than the secondary structures [40, 62-63] (see Sup Data 2 to see translation of a 3D VHH into a 1D sequence).

Analyses. Protein Blocks were used in this study to define specific patterns associated to each FR. For the main pattern associated to one FR is simply the succession of PBs that is the most occurring. Some PBs are close so the patterns can be slightly degenerated. We define the pattern as done in Prosite, e.g. the pattern "*r* [*st*] *u*" means that at position 1, the PB seen is PB *r*, in position 2 is it a mixture of PBs *s* and *t* that are close and at position 3, the PB *u*.

Intrinsic propensity fr_x^i of amino acids (or PBs) observed at a position *i* of a sequence are normalized by dividing the frequency f_x^i of this amino acid *x* (or BP) by the observed frequency of this amino acid (or PB) in a non-redundant protein structure databank. Thus, the intrinsic propensity is $fr_x^i = \frac{f_x^i}{f_x^{DB}}$. If this value is one, the frequency is expected at random, if it is higher, it is over-represented and if it is less, it is under-represented.

The equivalent number of PBs [or amino acids] [39, 64] (N_{eq}) is a statistical measurement similar to entropy and represents the average number of PBs [or amino acids] for a given residue. N_{eq}^i at position *i* is calculated as follows: $N_{eq}^i = exp(- \sum_{x=1}^{16} f_x^i \, ln \, f_x^i)$.

A N_{eq} value of 1 indicates that only one type of PB [or amino acids] is observed, while a value of 16 is equivalent to a random distribution for PBs [and 20 for amino acids].

For N-terminus FR (FR1) and C-terminus FR (FR4), it should be noted that only the positions actually found on all VHHs were analysed),

Results

Overall Analysis. As noted previously in material and methods section, search in PDB

website allowed the final selection of 133 different VHH chains. As expected, both the superposition of the structures and multiple alignments showed that VHHs topology is globally well conserved with conserved regions (FRs) but and other regions less conserved (CDRs).

To summarize, the succession of FRs and CDRs is clearly detectable. However, it must be initially well defined. Then the demarcation of the FRs and CDRs of conventional antibodies (IgG) has been studied and well characterized [17] and it can be transposed to VHHs. Nonetheless, this is not a trivial process [65-68]. To define a clear separation between FRs and CDRs, a multiple alignment was done using the ClustalO software [47]. The alignment also helped to select 10 VHHs representative of the entire data set. These VHHs represent the most diverse set of VHHs i.e. with highest sequence divergence. FRs and CDRs delineation corresponds to classical IMGT numbering for VHHs [69-70].

The FRs regions are commonly considered as 'constant' regions both in terms of sequence and structure while the CDRs are variable / hypervariable [17]. The visualization of the superimposition of representative VHHs confirms that topologies of FRs are quite similar and mainly composed of β-strands (see Figure 1A). It also highlights the structural diversity of CDRs, and at lesser extend to FRs, particularly regarding the size of β-strands or the conformation of the connecting loops. We were able to simplify the topology as one 2D projection see Figure 1B), which enables a simplified analysis of VHHs. This representation is based mainly on the interaction between β-strands. Please note that it had been postulated that VHH have a CDR “4” [71]. The region between residues 71–78 (according to IMGT numbering [69-70]) is close to the other CDRs, leading to a larger paratope [72-74]. In the present study we didn't take into account this potential CDR4 as a CDR but as intrinsically part of FR3 as the sequence identity of this region was high as expected in the FR regions.

FR analyses. These observations led to quantify without any *a priori* this 'constancy' of

FRs both in terms of sequence and of structure. Then we observed that the FRs are similar, but with some positions not conserved. Their percentages of identity range from 76.9% to 94.4% respectively for FR2 (14 residues, see Sup Data 3) and FR4 (9 residues). A specific analysis in terms of species (namely *llama glama*, *vicugna pacos*, and *camelus dromedarius*), show no specific tendencies from one species in regards to another, i.e. no sequence species specificity is observed.

We therefore analysed the 3D structure through a 1D representation via Protein Blocks. Then we clustered the succession of PBs leading to the description of FRs as patterns, this PB series can be seen as a classical Prosite patterns but not made from amino acids but with 3D information. These patterns are structural patterns as PBs represent 3D local conformation. For each FR, a distance matrix between these series of PBs has been made to group the closest structural patterns [75].

Table 1 shows the results for each FR. FR2 is the framework which has the structural pattern representing the largest number of structures (84%) while for the other FRs the most recurrent structural patterns only represents 40% (FR1), 63% (FR3) and 38% (FR4). We also observed no second highly recurrent patterns appear for each of the FRs. FR2 is the most directed structural pattern; it had the lowest number of variant patterns (5). The very long FR3 has 10 variants representatives of the remaining 37% of the structures. It is a limited number in regards to the length of this FR. No significant correlation between FR length and the percentage of structures corresponding to the variant patterns can be found, or even between this length and the number of variants patterns.

Focusing on the FRs amino acid sequence, we observed that 14 positions are constraint to one amino acid (see Figure 2). As shown in Table 1, 30 positions (out of 78) are exclusive to one type of amino acids while 14 positions show a N_{eq} on amino acids of these positions quite high (> 2). Thus, for FR2 and FR3, they represent only 3 residues on 14 and 11 on 32

respectively. For FR1 and FR4, the figure is higher but associated with greater structural variability, probably due to fewer structural constraints. We also noticed that non-exclusive positions are rare and usually have very different residues.

In terms of sequence – structure relationship, we evaluate the prediction of secondary structure of the whole dataset using PSIPRED software [55-56]. PSIPRED is the most widely used secondary structure prediction method with a third version leading to an expected average prediction rate (Q_3) of 82%. A striking result is the weak quality of the prediction for the VHHs; the Q_3 value is only of 72.2%. It is slightly better for FR2 (86.4%), but lower for other the FRs with FR1 equals to 75.5%, FR3 to 66,9% and FR4 to 72.0%. This observation may be related to the difficulty to predict β-strand, *i.e*. the most difficult repetitive structure to predict [76]. Jpred 4 [57] prediction leads to a relative similar prediction rate of 78%, with similar tends. In regards to VH of IgGs, the results do not show a direct correlation.

FR2 analyses. The analysis of the FR2 (see Figure 3) allows us to identify a main structural pattern (*ddd*[*de*]*ehia*[*cd*]*ddfbd*) and seven variant patterns (see Table 1 and Figure 3B). The first three sub-patterns present similar structure (see Figure 3C with variant patterns $FR2^{sm1.1}$, $FR2^{sm1.2}$ et $FR2^{sm1.3}$). The central region of the main structural pattern shows a specific series of PBs *ehia*, while these variants have a variation around a motif *f*[*kb*][*bl*]*c* which are clearly distinct from *ehia* as seen in [77]. The divergence of this framework encompasses especially the loop connecting the two β-strands of FR2, showing the existence of five distinct structural subunits (having a rmsd of 1 Å which is important for this short length sequence, see 3D examples in Figure 3C).

The differences are mostly found at the same positions. The $FR2^{sm2}$ is associated with PB series *dfbd* which causes 'contraction' of the loop; this loop is therefore more outwardly. The $FR2^{sm3}$ greatly diverges as loop often associated to β-strand is replaced by a typical loop

associated to helical structures, namely *fkop*. So the loop sticks up at the opposite of $FR2^{sm2}$. The PB motif *ehia* of $FR2^{sm4}$ is shifted and replace by *hiab* that is drastically different compared to the other motifs. Here the loop is much more extended, and not like any other FR2. $FR2^{sm5}$ is also different with specific features observed on the second β-strand of FR2. This allows the creation of more hydrogen bonds than other FR2 and extends the length of β-sheets.

For this framework along 14 residues, the positions 1 (residue W), 3 (residue R) and 7 (residue G) are associated to only one type of amino acids (see Figure 3A). For 4 other positions, the most represented amino acid does not exceed 80% (positions 2, 9, 12 and 14), and finally the most variable position is the position 7 with 12 residues (amino acids on N_{eq}> 4, Figure 3A).

Analysis of the various patterns of different FRs shows only one case of a sequence substitution clearly associated with a unique type of variant pattern, namely the $FR2^{sm3}$, characterized by the succession *fokp*, a Proline residue is observed; it is the only case observed.

The other FRs. Concerning FR3, which is the longest FR, the Figure 4A shows the occurrence of PBs inside the VHHs. The figure 4A underlines the variability of positions 6 to 13 in terms of structure and also in less extend the connection to CDR3. In the amino-acid sequence, 11 positions among 32 are only associated to 1 kind of residues. Five positions have equivalent residues (namely, residues are highly similar). The other positions represent half of the FR (16/32) and very different amino acid associations even with very different properties are found (see Sup Data 4).

We observed that the most variable local structure conformation is also the most variable in terms of sequence. Nonetheless, the amino acid sequences do not allow predicting

the associated PBs. Only position 7 is well-conserved compare to the other positions (positions 6, 10 and 12 being associated to higher amino acid N_{eq} values, see Figure 2C).

A precise analysis of the 10 variant patterns underlines that the main differences are found in positions of the sequence that are not strictly conserved in main structural pattern. In addition, no direct correlation can be seen with amino acid content and the PBs constructed sequence. As for FR3, the connecting loop is associated to various modifications (see Figure 4B). Most of them shorten the first β-strand of FR3 and we observed they are associated to high B-factors.

FR1 and FR4 are shorter than FR3 and FR2. Their main structural patterns are less represented (40 and 38% respectively). For this two FRs, only 4 positions among 23 have amino acid composition encompassing very diverse types of residues (big, aromatic, polar, charged…). This percentage is largely lower compared to FR2 and FR3. For FR1, the less determined structural region is the N-terminus region. It is also found with higher B-factors, and a low number of contacts. The more stable region starts with the second β-strand. We observed just before the second β-strand a sharp determined turn that is a PBs series *dehiac;* a very stable element that is important for protein structures [77]. The connection to CDR1, even if the length of CDR1 is short and constant, shows a high diversity in terms of local conformations. It highlights the impact of CDR1 on the final fold of FR1.

FR4 shows similar features. Its end is often short with high B-factors or unresolved, a typical case of most of the proteins [78-79]. Only the first nine positions are always found in all VHHs and have been analysed. The striking point is that 5 positions among 9 are totally conserved and only two have distinct amino acid kinds (position 1 with aromatic and positively charged residues, and position 3 with positively charged, polar uncharged and breakers). Two points must be underline here: (i) at first, these results have nothing to do with a potential problem of the PCR primer region. Indeed, it is mainly after the analysed positions

[80]; (ii) the number of number of J segments is limited (e.g. 6 in llama) and so have not really provided a strong divergence between the FR4s.

The high number of structural variant patterns (19 corresponding to about two third of the total number of FR4) encompasses a larger twisted conformation directly guided by the CDR3. FR4 first positions are not flexible region, so it underlines the impact of CDR3 to guide its local conformation. Here again, no correlation between the variant patterns in terms of PBs and amino acid contents.

Discussion

In the recent years, antibody uses have shown an impressive success for biotechnological and biomedical applications [5]. Use of bioinformatics approaches is an essential tool for engineering proteins and *a fortiori* for antibodies. A good knowledge of the sequence – structure relationship, which controls the protein folding, is essential. This is especially true for antibodies for which modifications are important questions. Humanization technology was fundamental for the remarkable progress of antibodies use for therapeutic area [81]. This optimisation needs a small set of variants. These variants design are based on the antibody structure and/or sequence information, and could impact folding, fold, fold stability and specificity. A recent *in silico* approach to perform the crafting of frameworks to accommodate other CDR regions had been proposed [82]. It combines homology modelling with simulated annealing to humanize mouse antibodies using computationally derived antibody homologous structures. However, *in silico* approaches for antibodies design for drug discovery have numerous drawbacks [83].

Compared to others antibodies, camelid VHHs have a short but rich story of experimental laboratory usage, biotechnology and biomedical purposes, *e.g*. nanobody-based cancer therapy of solid tumours [84]. Indeed, clear advantages of nanobodies compared to

conventional antibodies include their size, solubility and stability. Because of these characteristics, nanobodies can be formulated as a long shelf-life, ready-to-use solution [85].

Nonetheless, in terms of structure, few works have analysed their characteristics. Many studies have been done on a limited number of VHH structures. For instance, a potential universal VHH framework was tested to graft various loops of VHHs of subfamily 2 (representing 75% of all antigen-specific VHHs), but only 5 chimeras were tested [86]. Similarly, the convexity of paratope defined by CDRs of VHHs was mainly extracted from an analysis of eight VHHs binding lysozymes [87].

Comparative molecular modelling had shown the difficulty to propose pertinent structural model of VHHs [26, 35-36], even with very sophisticated approaches as RosettaAntibody [71]. In fact, the results can be appreciated as less efficient than for classical antibodies as tested in AMA I and AMA II [88-91].

From our experiences, we have underlined a precise characteristic of VHH modelling [36], which is the difficulty to select a correct series of structural templates. At the first sight frameworks seem all highly stable, and we need mainly to focus on CDRs. In this study, we have done the first attempt to analyse precisely the structural diversity of FRs of VHHs. An important methodological asset is the use of Protein Blocks. PBs are a very useful tools as they allow a local comparison with higher precision than classical secondary tools, and are very useful to align protein structures [51, 92] and analyse protein flexibility [62, 93].

Each of the four FRs is associated to a main structural pattern, which can be considered as canonical, but represents only 40, 84, 63, and 38%, leading to the characterization to variant patterns ranging from 6 to 19. Among some of them, we observe some similarity and the final number of variant patterns could be slightly reduced. However, clearly some are really outliers at more than 1 Å from the main structural pattern and could have strong impact of the proposition of structural models. The molecular modelling performed on these extreme

cases show that the structural models obtained are of low quality even for the FR regions (*data not show*). This result is in agreement with previous modelling [71]. It is also linked to the amino acid conservation of numerous FR positions, which is sometimes limited. In addition we underline that no correlation is found between the local conformational change and amino acid composition. These results indicate that long-range interactions affect the local conformation of this constrained topology and have strong implication (i) for comparative structural modelling and (ii) for antibody informatics for drug discovery. We have already observed a direct effect on the first point with some dedicated examples.

Acknowledgments

We would like to thanks Nicolas Shinada, Akhila Melarkode Vattekatte, Jean-Philippe Meyneil and Jean-Christophe Gelly for fruitful discussions, and Pierrick Craveur for his help with the VHH structures. This work was supported by grants from the French Ministry of Research, University Paris Diderot, Sorbonne Paris Cité, French National Institute for Blood Transfusion (INTS), French Institute for Health and Medical Research (INSERM). AdB also acknowledges the Indo-French Centre for the Promotion of Advanced Research / CEFIPRA for collaborative grants (numbers 5302-2). This study was supported by grants from the Laboratory of Excellence GR-Ex, reference ANR-11-LABX-0051. The labex GR-Ex is funded by the programme “Investissements d’avenir” of the French National Research Agency, reference ANR-11-IDEX-0005-02. Calculations were performed on an SGI cluster granted by Conseil Régional Ile de France and INTS (SESAME Grant).

Legends

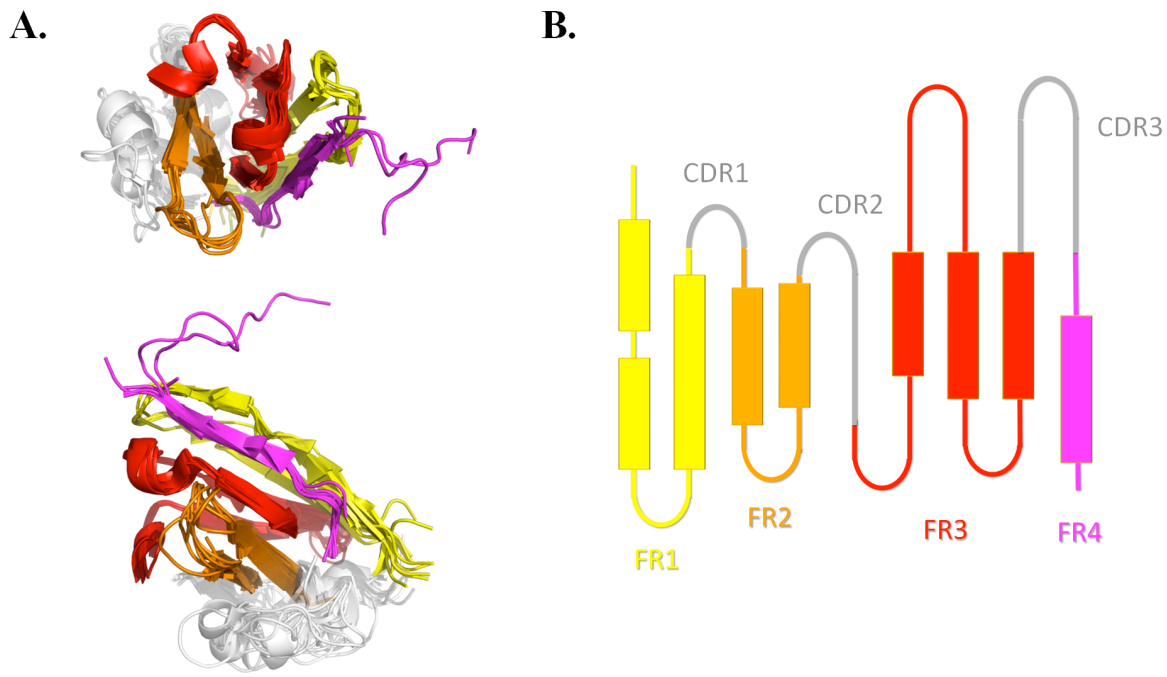


Figure 1. *VHH topology*. **A**) Superimposition of 10 representative VHH structures with two orientations. These 10 VHH represents the most diverse VHHs of the dataset in terms of sequences (PDB codes 3QXU chain D [8], 4GFT chain B [9], 3V0A chain C [10], 4C57 chain C [11], 4BFB chain E [12], 3CFI chain I [13], 4C58 chain B [11], 1KXQ chain F [14], 2X6M chain A [15], 2WZP chain E [16]). **B**) A 2D projection of the same information. In grey are indicated the three variable CDRs, while FR1 is in yellow, FR2 in orange, FR3 in red and FR4 in magenta. Superimposition was done using mulPBA [51] and visualisation with PyMOL software [48].

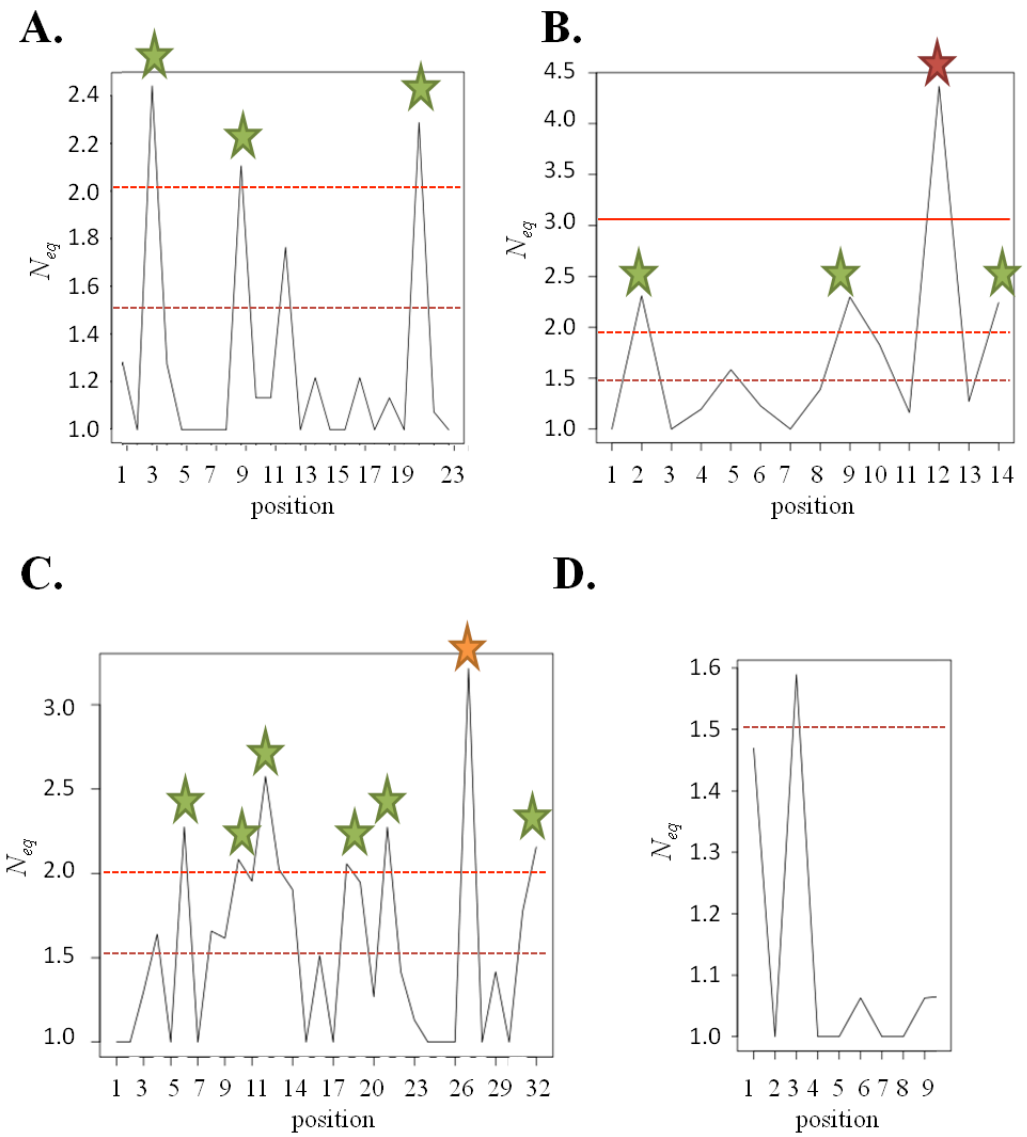


Figure 2. *Amino acid N_{eq} distribution of the four 4 FRs*. **A.** FR1, **B**. FR2, **C**. FR3 and **D**. FR4. Stars represent N_{eq} higher than 2.

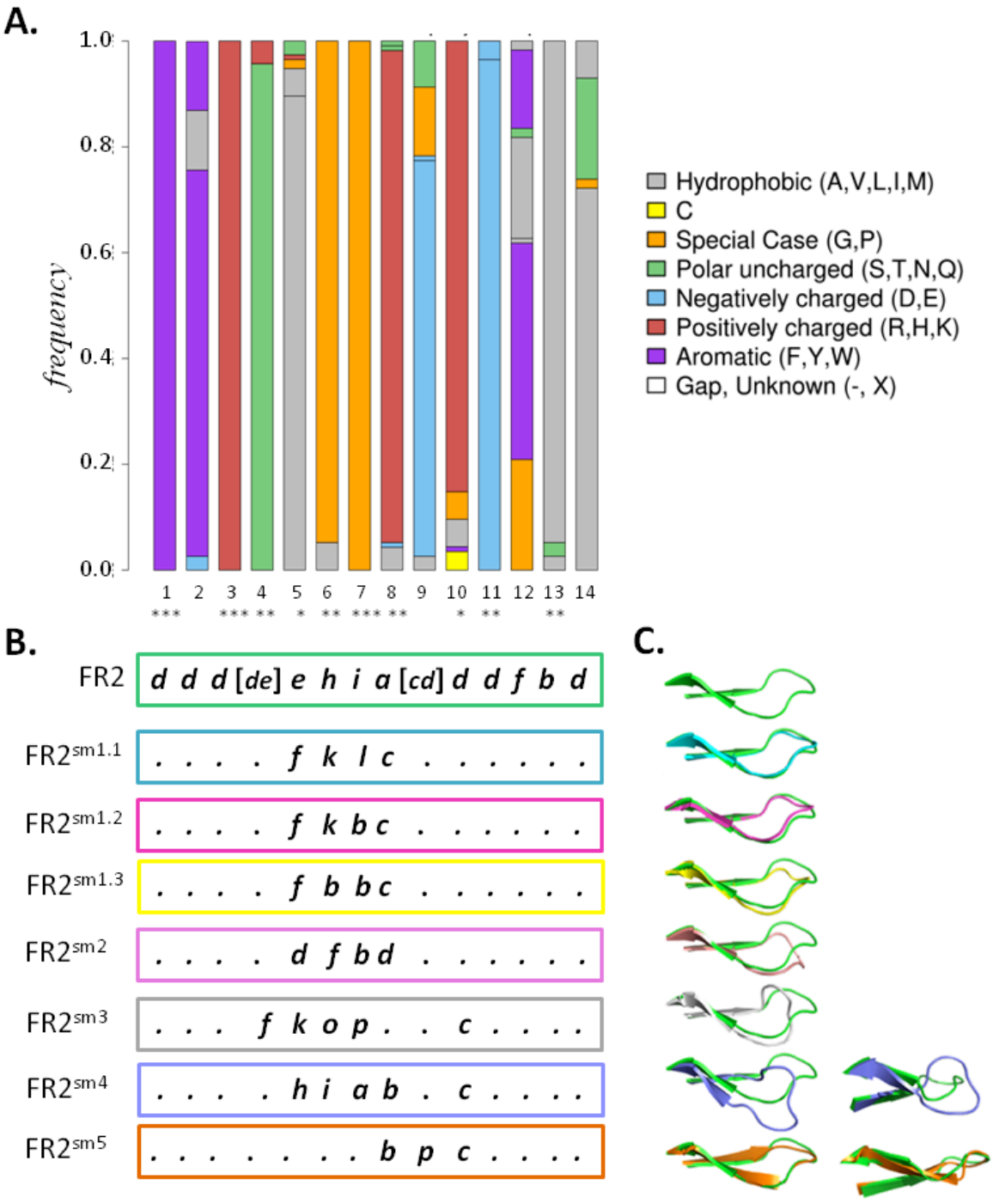


Figure 3. *FR2 characteristics*. **A**. amino acids frequencies at each position; **B**: Protein Blocks Representation of different structural variants of the FR2 (the dots correspond to the sequence identity with the main structural pattern); **C**. 3D representation of structural variants of the FR2. Following IMGT numbering [69-70], it corresponds to positions 36-49 which defines in IMGT this FR.

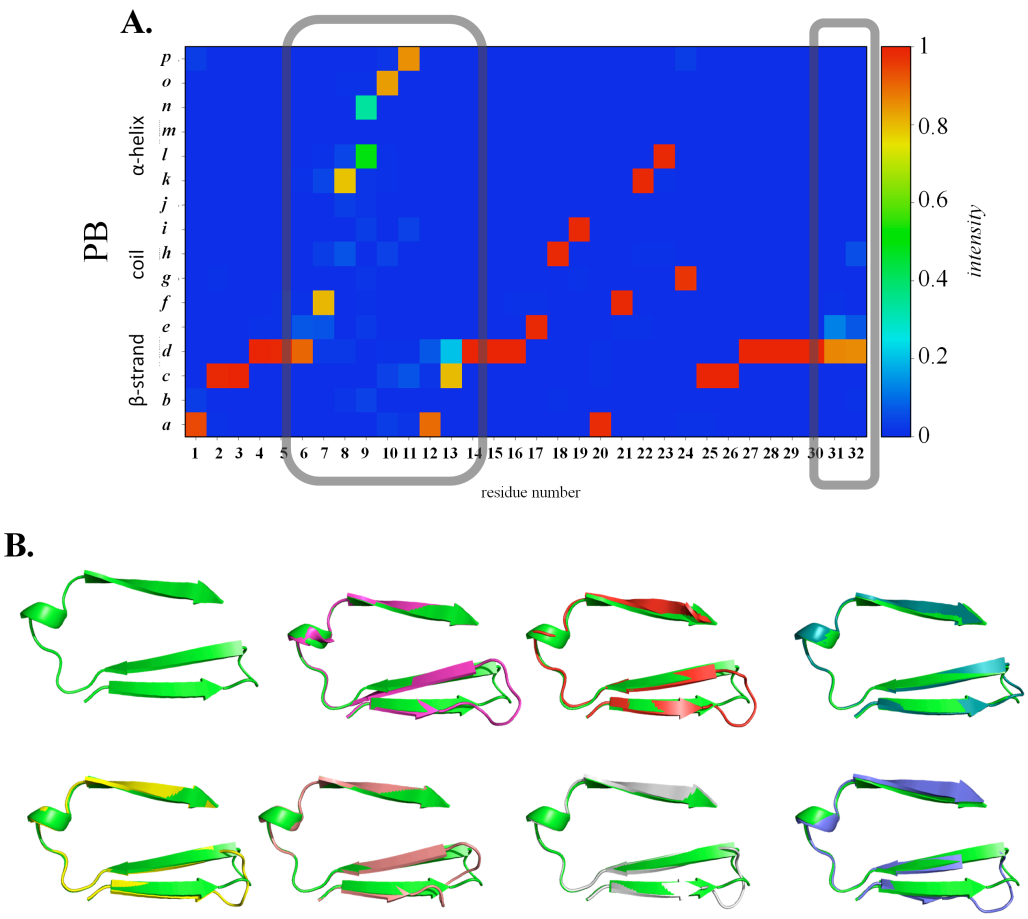


Figure 4. *FR3 characteristics*. **A**. PB frequencies along FR3, on the right is provided the corresponding gradient, the two more variable regions in terms of PBs are bordered in grey; **B**. 3D representation of the main structural pattern and some structural variants of the FR3.

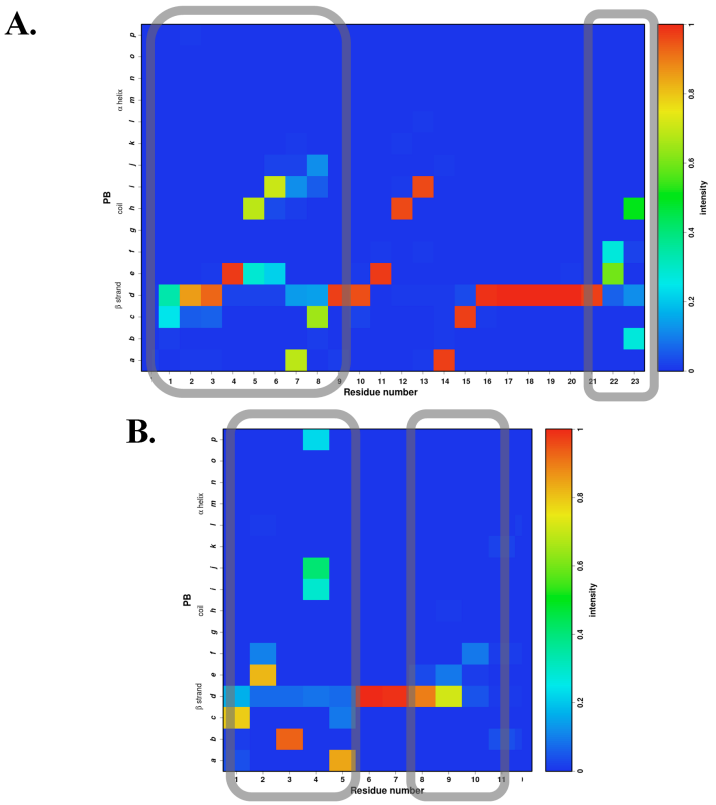


Figure 5. *FR1 and FR4 characteristics*. PB frequencies along **A**. FR1 and **B**. FR4, on the right is provided the corresponding gradient, the more variable regions in terms of PBs are bordered in grey.

Reference

[1] C.S. Kaetzel, Coevolution of Mucosal Immunoglobulins and the Polymeric Immunoglobulin Receptor: Evidence That the Commensal Microbiota Provided the Driving Force, ISRN Immunology, 2014 (2014) Article ID 541537.

[2] J.L. Teillaud, Engineering of monoclonal antibodies and antibody-based fusion proteins: successes and challenges, Expert Opin Biol Ther, 5 Suppl 1 (2005) S15-27.

[3] P. Holliger, P.J. Hudson, Engineered antibody fragments and the rise of single domains, Nat Biotechnol, 23 (2005) 1126-1136.

[4] J.D. Unciti-Broceta, T. Del Castillo, M. Soriano, S. Magez, J.A. Garcia-Salcedo, Novel therapy based on camelid nanobodies, Ther Deliv, 4 (2013) 1321-1336.

[5] J.B. Evans, B.A. Syed, From the analyst's couch: Next-generation antibodies, Nat Rev Drug Discov, 13 (2014) 413-414.

[6] C. Hamers-Casterman, T. Atarhouch, S. Muyldermans, G. Robinson, C. Hamers, E.B. Songa, N. Bendahman, R. Hamers, Naturally occurring antibodies devoid of light chains, Nature, 363 (1993) 446-448.

[7] C. Su, V.K. Nguyen, M. Nei, Adaptive evolution of variable region genes encoding an unusual type of immunoglobulin in camelids, Mol Biol Evol, 19 (2002) 205-215.

[8] S.W. Fanning, J.R. Horn, An anti-hapten camelid antibody reveals a cryptic binding site with significant energetic contributions from a nonhypervariable loop, Protein Sci, 20 (2011) 1196-1207.

[9] S. Khamrui, S. Turley, E. Pardon, J. Steyaert, E. Fan, C.L. Verlinde, L.W. Bergman, W.G. Hol, The structure of the D3 domain of Plasmodium falciparum myosin tail interacting protein MTIP in complex with a nanobody, Mol Biochem Parasitol, 190 (2013) 87-91.

[10] S. Gu, S. Rumpel, J. Zhou, J. Strotmeier, H. Bigalke, K. Perry, C.B. Shoemaker, A. Rummel, R. Jin, Botulinum neurotoxin is shielded by NTNHA in an interlocked complex, Science, 335 (2012) 977-981.

[11] A. Chaikuad, T. Keates, C. Vincke, M. Kaufholz, M. Zenn, B. Zimmermann, C. Gutierrez, R.G. Zhang, C. Hatzos-Skintges, A. Joachimiak, S. Muyldermans, F.W. Herberg, S. Knapp, S. Muller, Structure of cyclin G-associated kinase (GAK) trapped in different conformations using nanobodies, Biochem J, 459 (2014) 59-69.

[12] D.W. Banner, B. Gsell, J. Benz, J. Bertschinger, D. Burger, S. Brack, S. Cuppuleri, M. Debulpaep, A. Gast, D. Grabulovski, M. Hennig, H. Hilpert, W. Huber, A. Kuglstatter, E. Kusznir, T. Laeremans, H. Matile, C. Miscenic, A.C. Rufer, D. Schlatter, J. Steyaert, M. Stihle, R. Thoma, M. Weber, A. Ruf, Mapping the conformational space accessible to BACE2 using surface mutants and cocrystals with Fab fragments, Fynomers and Xaperones, Acta Crystallogr D Biol Crystallogr, 69 (2013) 1124-1137.

[13] A.Y. Lam, E. Pardon, K.V. Korotkov, W.G. Hol, J. Steyaert, Nanobody-aided structure determination of the EpsI:EpsJ pseudopilin heterodimer from Vibrio vulnificus, J Struct Biol, 166 (2009) 8-15.

[14] A. Desmyter, S. Spinelli, F. Payan, M. Lauwereys, L. Wyns, S. Muyldermans, C. Cambillau, Three camelid VHH domains in complex with porcine pancreatic alpha-amylase. Inhibition and versatility of binding topology, J Biol Chem, 277 (2002) 23645-23650.

[15] E.J. De Genst, T. Guilliams, J. Wellens, E.M. O'Day, C.A. Waudby, S. Meehan, M. Dumoulin, S.T. Hsu, N. Cremades, K.H. Verschueren, E. Pardon, L. Wyns, J. Steyaert, J. Christodoulou, C.M. Dobson, Structure and properties of a complex of alpha-synuclein and a single-domain camelid antibody, J Mol Biol, 402 (2010) 326-343.

[16] G. Sciara, C. Bebeacua, P. Bron, D. Tremblay, M. Ortiz-Lombardia, J. Lichiere, M. van Heel, V. Campanacci, S. Moineau, C. Cambillau, Structure of lactococcal phage p2 baseplate and its mechanism of activation, Proc Natl Acad Sci U S A, 107 (2010) 6852-6857.

[17] S. Muyldermans, T.N. Baral, V.C. Retamozzo, P. De Baetselier, E. De Genst, J. Kinne, H. Leonhardt, S. Magez, V.K. Nguyen, H. Revets, U. Rothbauer, B. Stijlemans, S. Tillib, U. Wernery, L. Wyns, G. Hassanzadeh-Ghassabeh, D. Saerens, Camelid immunoglobulins and nanobody technology, Vet Immunol Immunopathol, 128 (2009) 178-183.

[18] C. Vincke, R. Loris, D. Saerens, S. Martinez-Rodriguez, S. Muyldermans, K. Conrath, General strategy to humanize a camelid single-domain antibody and identification of a universal humanized nanobody scaffold, J Biol Chem, 284 (2009) 3273-3284.

[19] A. Fatima, H. Wang, K. Kang, L. Xia, Y. Wang, W. Ye, J. Wang, X. Wang, Development of VHH antibodies against dengue virus type 2 NS1 and comparison with monoclonal antibodies for use in immunological diagnosis, PLoS One, 9 (2014) e95263.

[20] F.M. Cardoso, L.I. Ibanez, S. Van den Hoecke, S. De Baets, A. Smet, K. Roose, B. Schepens, F.J. Descamps, W. Fiers, S. Muyldermans, A. Depicker, X. Saelens, Single-domain antibodies targeting neuraminidase protect against an H5N1 influenza virus challenge, J Virol, 88 (2014) 8278-8296.

[21] A. Desmyter, C. Farenc, J. Mahony, S. Spinelli, C. Bebeacua, S. Blangy, D. Veesler, D. van Sinderen, C. Cambillau, Viral infection modulation and neutralization by camelid nanobodies, Proc Natl Acad Sci U S A, 110 (2013) E1371-1379.

[22] Y. Wang, P. Li, Q. Zhang, X. Hu, W. Zhang, A toxin-free enzyme-linked immunosorbent assay for the analysis of aflatoxins based on a VHH surrogate standard, Anal Bioanal Chem, (2016).

[23] P.B. van Driel, M.C. Boonstra, M.D. Slooter, R. Heukers, M.A. Stammes, T.J. Snoeks, H.S. de Bruijn, P.J. van Diest, A.L. Vahrmeijer, P.M. van Bergen En Henegouwen, C.J. van de Velde, C.W. Lowik, D.J. Robinson, S. Oliveira, EGFR targeted nanobody-photosensitizer conjugates for photodynamic therapy in a pre-clinical model of head and neck cancer, J Control Release, 229 (2016) 93-105.

[24] M. Qasemi, M. Behdani, M.A. Shokrgozar, V. Molla-Kazemiha, H. Mohseni-Kuchesfahani, M. Habibi-Anbouhi, Construction and expression of an anti-VEGFR2 Nanobody-Fc fusionbody in NS0 host cell, Protein Expr Purif, 123 (2016) 19-25.

[25] N.D. Prado, S.S. Pereira, M.P. da Silva, M.S. Morais, A.M. Kayano, L.S. Moreira-Dill, M.B. Luiz, F.B. Zanchi, A.L.

Fuly, E.F.H. M, C.F. Fernandes, L.A. Calderon, J.P. Zuliani, L.H. Pereira da Silva, A.M. Soares, R.G. Stabeli, F.C.F. C, Inhibition of the Myotoxicity Induced by Bothrops jararacussu Venom and Isolated Phospholipases A2 by Specific Camelid Single-Domain Antibody Fragments, PLoS One, 11 (2016) e0151363.

[26] D. Smolarek, C. Hattab, A. Buczkowska, R. Kaczmarek, A. Jarzab, S. Cochet, A.G. de Brevern, J. Lukasiewicz, W. Jachymek, T. Niedziela, M. Grodecka, K. Wasniowska, Y. Colin Aronovicz, O. Bertrand, M. Czerwinski, Studies of a Murine Monoclonal Antibody Directed against DARC: Reappraisal of Its Specificity, PLoS One, 10 (2015) e0116472.

[27] J.B. Holz, The TITAN trial--assessing the efficacy and safety of an anti-von Willebrand factor Nanobody in patients with acquired thrombotic thrombocytopenic purpura, Transfus Apher Sci, 46 (2012) 343-346.

[28] F. Peyvandi, M. Scully, J.A. Kremer Hovinga, S. Cataland, P. Knobl, H. Wu, A. Artoni, J.P. Westwood, M. Mansouri Taleghani, B. Jilma, F. Callewaert, H. Ulrichts, C. Duby, D. Tersago, T. Investigators, Caplacizumab for Acquired Thrombotic Thrombocytopenic Purpura, N Engl J Med, 374 (2016) 511-522.

[29] E. Pardon, T. Laeremans, S. Triest, S.G. Rasmussen, A. Wohlkonig, A. Ruf, S. Muyldermans, W.G. Hol, B.K. Kobilka, J. Steyaert, A general protocol for the generation of Nanobodies for structural biology, Nat Protoc, 9 (2014) 674-693.

[30] A. Desmyter, S. Spinelli, A. Roussel, C. Cambillau, Camelid nanobodies: killing two birds with one stone, Curr Opin Struct Biol, 32 (2015) 1-8.

[31] G. Hassaine, C. Deluz, L. Grasso, R. Wyss, M.B. Tol, R. Hovius, A. Graff, H. Stahlberg, T. Tomizaki, A. Desmyter, C. Moreau, X.D. Li, F. Poitevin, H. Vogel, H. Nury, X-ray structure of the mouse serotonin 5-HT3 receptor, Nature, 512 (2014) 276-281.

[32] A.M. Ring, A. Manglik, A.C. Kruse, M.D. Enos, W.I. Weis, K.C. Garcia, B.K. Kobilka, Adrenaline-activated structure of beta2-adrenoceptor stabilized by an engineered nanobody, Nature, 502 (2013) 575-579.

[33] K.R. Schmitz, A. Bagchi, R.C. Roovers, P.M. van Bergen en Henegouwen, K.M. Ferguson, Structural evaluation of EGFR inhibition mechanisms for nanobodies/VHH domains, Structure, 21 (2013) 1214-1224.

[34] D. Saerens, F. Frederix, G. Reekmans, K. Conrath, K. Jans, L. Brys, L. Huang, E. Bosmans, G. Maes, G. Borghs, S. Muyldermans, Engineering camel single-domain antibodies and immobilization chemistry for human prostate-specific antigen sensing, Anal Chem, 77 (2005) 7547-7555.

[35] D. Smolarek, C. Hattab, G. Hassanzadeh-Ghassabeh, S. Cochet, C. Gutierrez, A.G. de Brevern, R. Udomsangpetch, J. Picot, M. Grodecka, K. Wasniowska, S. Muyldermans, Y. Colin, C. Le Van Kim, M. Czerwinski, O. Bertrand, A recombinant dromedary antibody fragment (VHH or nanobody) directed against human Duffy antigen receptor for chemokines, Cell Mol Life Sci, 67 (2010) 3371-3387.

[36] D. Smolarek, O. Bertrand, M. Czerwinski, Y. Colin, C. Etchebest, A.G. de Brevern, Multiple interests in structural models of DARC transmembrane protein, Transfus Clin Biol, 17 (2010) 184-196.

[37] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res, 28 (2000) 235-242.

[38] F.C. Bernstein, T.F. Koetzle, G.J. Williams, E.F. Meyer, Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, The Protein Data Bank: a computer-based archival file for macromolecular structures, J Mol Biol, 112 (1977) 535-542.

[39] A.G. de Brevern, C. Etchebest, S. Hazout, Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks, Proteins, 41 (2000) 271-287.

[40] A.P. Joseph, G. Agarwal, S. Mahajan, J.C. Gelly, L.S. Swapna, B. Offmann, F. Cadet, A. Bornot, M. Tyagi, H. Valadie, B. Schneider, C. Etchebest, N. Srinivasan, A.G. De Brevern, A short survey on protein blocks, Biophys Rev, 2 (2010) 137-147.

[41] V. Jallu, P. Poulain, P.F. Fuchs, C. Kaplan, A.G. de Brevern, Modeling and molecular dynamics simulations of the V33 variant of the integrin subunit beta3: Structural comparison with the L33 (HPA-1a) and P33 (HPA-1b) variants, Biochimie, 105 (2014) 84-90.

[42] V. Jallu, P. Poulain, P.F. Fuchs, C. Kaplan, A.G. de Brevern, Modeling and molecular dynamics of HPA-1a and -1b polymorphisms: effects on the structure of the beta3 subunit of the alphaIIbbeta3 integrin, PLoS One, 7 (2012) e47304.

[43] V. Jallu, G. Bertrand, F. Bianchi, C. Chenet, P. Poulain, C. Kaplan, The alphaIIb p.Leu841Met (Cab3(a+)) polymorphism results in a new human platelet alloantigen involved in neonatal alloimmune thrombocytopenia, Transfusion, 53 (2013) 554-563.

[44] L. Chevrier, A. de Brevern, E. Hernandez, J. Leprince, H. Vaudry, A.M. Guedj, N. de Roux, PRR repeats in the intracellular domain of KISS1R are important for its export to cell membrane, Mol Endocrinol, 27 (2013) 1004-1014.

[45] H. Berman, K. Henrick, H. Nakamura, Announcing the worldwide Protein Data Bank, Nat Struct Biol, 10 (2003) 980.

[46] P. Craveur, J. Rebehmed, A.G. de Brevern, PTM-SD: a database of structurally resolved and annotated posttranslational modifications in proteins, Database (Oxford), 2014 (2014).

[47] F. Sievers, D.G. Higgins, Clustal Omega, accurate alignment of very large numbers of sequences, Methods Mol Biol, 1079 (2014) 105-116.

[48] W.L. Delano, The PyMOL Molecular Graphics System on World Wide Web. http://www.pymol.org, (2013).

[49] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria., 2013.

[50] A.C.R. Martin, C.T. Porter, PROFIT (v 4.0), http://www.bioinf.org.uk/software/profit/, 2010.

[51] S. Leonard, A.P. Joseph, N. Srinivasan, J.C. Gelly, A.G. de Brevern, mulPBA: an efficient multiple protein structure alignment method based on a structural alphabet, J Biomol Struct Dyn, 32 (2014) 661-668.

[52] I. Kufareva, R. Abagyan, Methods of protein structure comparison, Methods Mol Biol, 857 (2012) 231-257.

[53] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers, 22 (1983) 2577-2637.
[54] L. Fourrier, C. Benros, A.G. de Brevern, Use of a structural alphabet for analysis of short loops connecting repetitive structures, BMC Bioinformatics, 5 (2004) 58.
[55] L.J. McGuffin, K. Bryson, D.T. Jones, The PSIPRED protein structure prediction server, Bioinformatics, 16 (2000) 404-405.
[56] D.W. Buchan, F. Minneci, T.C. Nugent, K. Bryson, D.T. Jones, Scalable web services for the PSIPRED Protein Analysis Workbench, Nucleic Acids Res, 41 (2013) W349-357.
[57] A. Drozdetskiy, C. Cole, J. Procter, G.J. Barton, JPred4: a protein secondary structure prediction server, Nucleic Acids Res, 43 (2015) W389-394.
[58] M. Tyagi, A. Bornot, B. Offmann, A.G. de Brevern, Analysis of loop boundaries using different local structure assignment methods, Protein Sci, 18 (2009) 1869-1881.
[59] T. Kohonen, Self-organized formation of topologically correct feature maps, Biol. Cybern, 43 (1982) 59-69.
[60] L.R. Rabiner, A tutorial on hidden Markov models and selected application in speech recognition, Proceedings of the IEEE, 77 (1989) 257-286.
[61] A.G. de Brevern, New assessment of a structural alphabet, In Silico Biol, 5 (2005) 283-289.
[62] P. Craveur, A.P. Joseph, J. Esque, T.J. Narwani, F. Noel, N. Shinada, M. Goguet, L. Sylvain, P. Poulain, O. Bertrand, G. Faure, J. Rebehmed, A. Ghozlane, L.S. Swapna, R.M. Bhaskara, J. Barnoud, S. Téletchéa, V. Jallu, J. Cerny, B. Schneider, C. Etchebest, N. Srinivasan, J.-C. Gelly, A.G. de Brevern, Protein flexibility in the light of structural alphabets, Frontiers in Molecular Biosciences, 2 (2015).
[63] A.G. de Brevern, A.P. Joseph, Species specific amino acid sequence-protein local structure relationships: An analysis in the light of a structural alphabet, J Theor Biol, 276 (2011) 209-217.
[64] C. Etchebest, C. Benros, S. Hazout, A.G. de Brevern, A structural alphabet for local protein structures: Improved prediction methods, Proteins, (2005) 810-827.
[65] V.K. Nguyen, R. Hamers, L. Wyns, S. Muyldermans, Camel heavy-chain antibodies: diverse germline V(H)H and specific mechanisms enlarge the antigen-binding repertoire, EMBO J, 19 (2000) 921-930.
[66] M.P. Lefranc, V. Giudicelli, P. Duroux, J. Jabado-Michaloud, G. Folch, S. Aouinti, E. Carillon, H. Duvergey, A. Houles, T. Paysan-Lafosse, S. Hadi-Saljoqi, S. Sasorith, G. Lefranc, S. Kossida, IMGT(R), the international ImMunoGeneTics information system(R) 25 years on, Nucleic Acids Res, 43 (2015) D413-422.
[67] M.P. Lefranc, V. Giudicelli, Q. Kaas, E. Duprat, J. Jabado-Michaloud, D. Scaviner, C. Ginestoux, O. Clement, D. Chaume, G. Lefranc, IMGT, the international ImMunoGeneTics information system, Nucleic Acids Res, 33 (2005) D593-597.
[68] V.K. Nguyen, S. Muyldermans, R. Hamers, The specific variable domain of camel heavy-chain antibodies is encoded in the germline, J Mol Biol, 275 (1998) 413-418.
[69] M.P. Lefranc, IMGT, the International ImMunoGeneTics Information System, Cold Spring Harb Protoc, 2011 (2011) 595-603.
[70] M.P. Lefranc, V. Giudicelli, C. Ginestoux, D. Chaume, IMGT, the international ImMunoGeneTics information system, http://imgt.cines.fr: the reference in immunoinformatics, Stud Health Technol Inform, 95 (2003) 74-79.
[71] A. Sircar, K.A. Sanni, J. Shi, J.J. Gray, Analysis and modeling of the variable region of camelid single-domain antibodies, J Immunol, 186 (2011) 6357-6367.
[72] C.J. Bond, C. Wiesmann, J.C. Marsters, Jr., S.S. Sidhu, A structure-based database of antibody variable domain diversity, J Mol Biol, 348 (2005) 699-709.
[73] J.D. Capra, J.M. Kehoe, Variable region sequences of five human immunoglobulin heavy chains of the VH3 subgroup: definitive identification of four heavy chain hypervariable regions, Proc Natl Acad Sci U S A, 71 (1974) 845-848.
[74] P. Carter, L. Presta, C.M. Gorman, J.B. Ridgway, D. Henner, W.L. Wong, A.M. Rowland, C. Kotts, M.E. Carver, H.M. Shepard, Humanization of an anti-p185HER2 antibody for human cancer therapy, Proc Natl Acad Sci U S A, 89 (1992) 4285-4289.
[75] M. Tyagi, A.G. de Brevern, N. Srinivasan, B. Offmann, Protein structure mining using a structural alphabet, Proteins, 71 (2008) 920-937.
[76] J. Andreani, J. Soding, bbcontacts: prediction of beta-strand pairing from direct coupling patterns, Bioinformatics, 31 (2015) 1729-1737.
[77] A.G. de Brevern, H. Valadie, S. Hazout, C. Etchebest, Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship, Protein Sci, 11 (2002) 2871-2886.
[78] G. Faure, A. Bornot, A.G. de Brevern, Protein contacts, inter-residue interactions and side-chain modelling, Biochimie, 90 (2008) 626-639.
[79] E. Jacob, R. Unger, A tale of two tails: why are terminal residues of proteins exposed?, Bioinformatics, 23 (2007) e225-230.
[80] D.R. Maass, J. Sepulveda, A. Pernthaner, C.B. Shoemaker, Alpaca (Lama pacos) as a convenient source of recombinant camelid heavy chain antibodies (VHHs), J Immunol Methods, 324 (2007) 13-25.
[81] J.C. Almagro, J. Fransson, Humanization of antibodies, Front Biosci, 13 (2008) 1619-1633.
[82] V.B. Kurella, R. Gali, Structure guided homology model based design and engineering of mouse antibodies for humanization, Bioinformation, 10 (2014) 180-186.
[83] H. Shirai, C. Prades, R. Vita, P. Marcatili, B. Popovic, J. Xu, J.P. Overington, K. Hirayama, S. Soga, K. Tsunoyama, D. Clark, M.P. Lefranc, K. Ikeda, Antibody informatics for drug discovery, Biochim Biophys Acta, 1844 (2014) 2002-2015.
[84] M. Kijanka, B. Dorresteijn, S. Oliveira, P.M. van Bergen en Henegouwen, Nanobody-based cancer therapy of solid tumors, Nanomedicine (Lond), 10 (2015) 161-174.

[85] M. Dumoulin, K. Conrath, A. Van Meirhaeghe, F. Meersman, K. Heremans, L.G. Frenken, S. Muyldermans, L. Wyns, A. Matagne, Single-domain antibody fragments with high conformational stability, Protein Sci, 11 (2002) 500-515.

[86] D. Saerens, M. Pellis, R. Loris, E. Pardon, M. Dumoulin, A. Matagne, L. Wyns, S. Muyldermans, K. Conrath, Identification of a universal VHH framework to graft non-canonical antigen-binding loops of camel single-domain antibodies, J Mol Biol, 352 (2005) 597-607.

[87] E. De Genst, K. Silence, K. Decanniere, K. Conrath, R. Loris, J. Kinne, S. Muyldermans, L. Wyns, Molecular basis for the preferential cleft recognition by dromedary heavy-chain antibodies, Proc Natl Acad Sci U S A, 103 (2006) 4586-4591.

[88] J.K. Maier, P. Labute, Assessment of fully automated antibody homology modeling protocols in molecular operating environment, Proteins, 82 (2014) 1599-1610.

[89] J.C. Almagro, A. Teplyakov, J. Luo, R.W. Sweet, S. Kodangattil, F. Hernandez-Guzman, G.L. Gilliland, Second antibody modeling assessment (AMA-II), Proteins, 82 (2014) 1553-1562.

[90] A. Teplyakov, J. Luo, G. Obmolova, T.J. Malia, R. Sweet, R.L. Stanfield, S. Kodangattil, J.C. Almagro, G.L. Gilliland, Antibody modeling assessment II. Structures and models, Proteins, 82 (2014) 1563-1582.

[91] J.C. Almagro, M.P. Beavers, F. Hernandez-Guzman, J. Maier, J. Shaulsky, K. Butenhof, P. Labute, N. Thorsteinson, K. Kelly, A. Teplyakov, J. Luo, R. Sweet, G.L. Gilliland, Antibody modeling assessment, Proteins, 79 (2011) 3050-3066.

[92] J.C. Gelly, A.P. Joseph, N. Srinivasan, A.G. de Brevern, iPBA: a tool for protein structure comparison using sequence alignment strategies, Nucleic Acids Res, 39 (2011) W18-23.

[93] A.G. de Brevern, A. Bornot, P. Craveur, C. Etchebest, J.C. Gelly, PredyFlexy: flexibility and local structure prediction from sequence, Nucleic Acids Res, 40 (2012) W317-322.

Region	#positions	Main pattern (%structures)	#positions with unique AA	#variant patterns (%structures)
FR1	23	1 (40)	11	14 (60)
FR2	14	1 (84)	3	6 (16)
FR3	32	1 (63)	11	10 (37)
FR4	9	1 (38)	5	19 (62)
Total	78	1 (*56*)	30	39 (*44*)

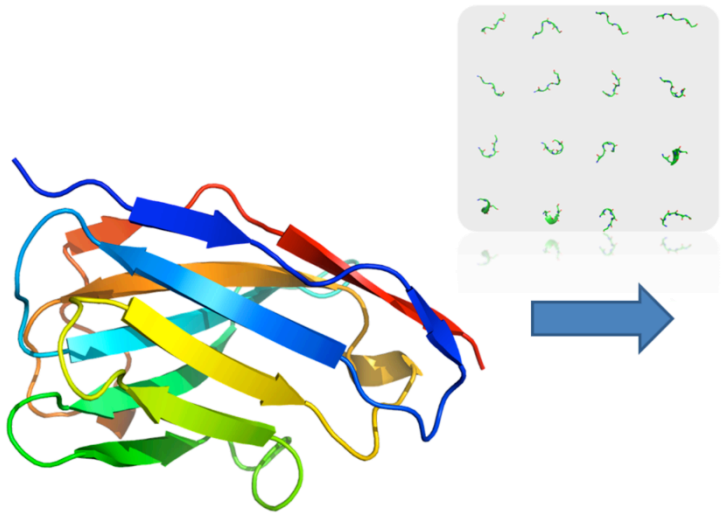
Table 1. *Summary of the difference variant patterns.* The following features are noted for each FR : The FR length, the percentage of structures corresponding to the main structural pattern, the number of positions associated to only one kind of amino acid, and the number of structural variants (with corresponding percentage). The last line corresponds to the summary of all FRs, in italics is provided the mean values of structure occurrences.

SUPPLEMENTARY DATA

Sup Data 1. *Protein dataset.* The 133 protein structures are described by species and types of experiments.

Number of structures	X-ray diffraction	Solution NMR	Total
Camelidae	1	0	1
Camelus dromedarius	32	0	32
Lama glama	93	1	94
Vicugna pacos	6	0	6
Total	132	1	133

Sup Data 2. *Translation of a VHH structures in a sequence of Protein Blocks*. On the left is the 3D structure of a camelid VHH [which is in complex with porcine pancreatic α-Amylase, http://www.rcsb.org/pdb/explore.do?structureId=1kxq]. The protein is encoded as a series of (φ, ψ) dihedral angles. Each consecutive fragment of five residues is compared to the series of (φ, ψ) dihedral angles of the 16 canonical Protein Blocks [1-3]. The one with the minimal distance is associated to the central residues. The protein fragments are overlapping. Please note that the final PB sequences are 4 residues shorter than the amino acid sequences (corresponding to the *ZZ* at N and C termini).

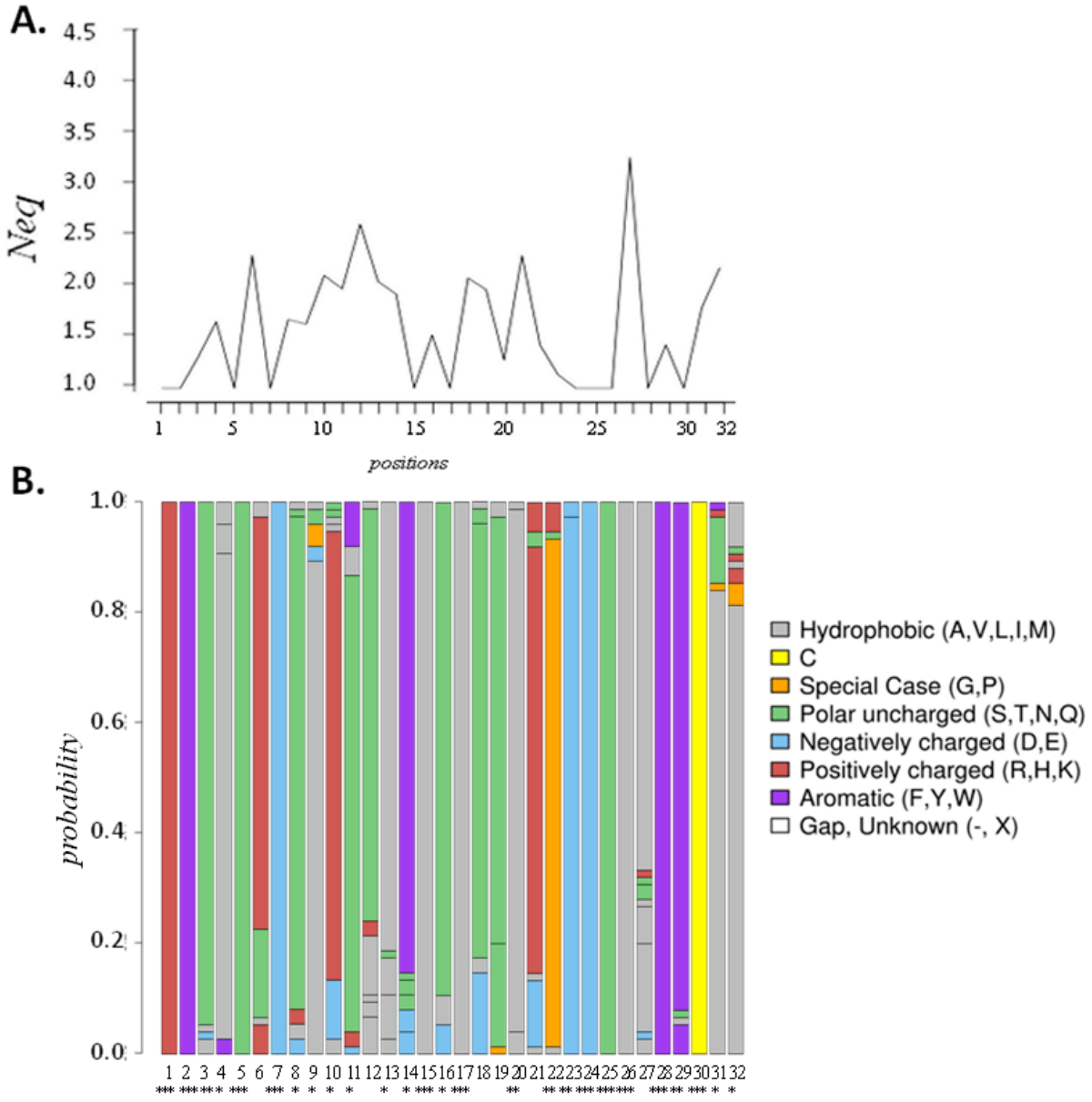


>1KXQ_F
ZZdddeeediddehiacdd
ddddfbdcfbdcddddde h
iadddfbdcdfknopacdd
fklgokaccdddfkbccdd
dddehiafklgccddddd d
fbfblknogoiabdcebia
ddddZZ

Sup Data 3. *Redundancy of the dataset.* Identity (%ID) mean rates for each region (FRs and CDRs) between 133 VHHs sequences obtained from the PDB (and associated standard-deviation, sd).

	% ID (mean)	% ID (sd)
FR1	88.70	4.06
FR2	76.91	8.54
FR3	79.35	4.51
FR4	94.43	4.70
CDR1	33.47	10.32
CDR2	52.13	6.19
CDR3	19.31	4.32

Sup Data 4. *FR3 characteristics*. **A**. Amino acid N_{eq} of FR3. **B**. Occurrence of amino acids at each position. Following IMGT numbering [4, 5], it corresponds to positions 66-94.



1. Joseph, A.P., et al., *A short survey on protein blocks.* Biophys Rev, 2010. **2**(3): p. 137-147.
2. de Brevern, A.G., *New assessment of a structural alphabet.* In Silico Biol, 2005. **5**(3): p. 283-9.
3. de Brevern, A.G., C. Etchebest, and S. Hazout, *Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks.* Proteins, 2000. **41**(3): p. 271-87.
4. Lefranc, M.P., *IMGT, the International ImMunoGeneTics Information System.* Cold Spring Harb Protoc, 2011. **2011**(6): p. 595-603.
5. Lefranc, M.P., et al., *IMGT, the international ImMunoGeneTics information system,* http://imgt.cines.fr: *the reference in immunoinformatics.* Stud Health Technol Inform, 2003. **95**: p. 74-9.