Analysis and Modeling of the Variable Region of Camelid Single-Domain Antibodies

Aroop Sircar, Kayode A. Sanni, Jiye Shi and Jeffrey J. Gray

J Immunol published online 27 April 2011
http://www.jimmunol.org/content/early/2011/04/27/jimmunol.1100116

Supplementary Material http://www.jimmunol.org/content/suppl/2011/04/27/jimmunol.1100116.DC1
Subscription Information about subscribing to The Journal of Immunology is online at:
http://jimmunol.org/subscription
Permissions Submit copyright permission requests at:
http://www.aai.org/About/Publications/JI/copyright.html
Email Alerts Receive free email-alerts when new articles cite this article. Sign up at:
http://jimmunol.org/alerts
Analysis and Modeling of the Variable Region of Camelid Single-Domain Antibodies

Aroop Sircar,*† Kayode A. Sanni,‡ Jiye Shi,* and Jeffrey J. Gray*†

Camelids have a special type of Ab, known as heavy chain Abs, which are devoid of classical Ab light chains. Relative to classical Abs, camelid heavy chain Abs (cAbs) have comparable immunogenicity, Ag recognition diversity and binding affinities, higher stability and solubility, and better manufacturability, making them promising candidates for alternate therapeutic scaffolds. Rational engineering of cAbs to improve therapeutic function requires knowledge of the differences of sequence and structural features between cAbs and classical Abs. In this study, amino acid sequences of 27 cAb variable regions (VH) were aligned with the respective regions of 54 classical Abs to detect amino acid differences, enabling automatic identification of cAb VHH CDRs. CDR analysis revealed that the H1 often (and sometimes the H2) adopts diverse conformations not classifiable by established canonical rules. Also, although the cAb H3 is much longer than classical H3 loops, it often contains common structural motifs and sometimes a disulide bond to the H1. Leveraging these observations, we created a Monte Carlo-based cAb VHH structural modeling tool, where the CDR H1 and H2 loops exhibited a median root-mean-square deviation to natives of 3.1 and 1.5 Å, respectively. The protocol generated 8–12, 14–16, and 16–24 residue H3 loops with a median root-mean-square deviation to natives of 5.7, 4.5, and 6.8 Å, respectively. The large deviation of the predicted loops underscores the challenge in modeling such long loops. cAb VH homology models can provide structural insights into interaction mechanisms to enable development of novel Abs for therapeutic and biotechnological use.

The Journal of Immunology, 2011, 186: 000–000.

The unique properties of the cAbs can be attributed to changes in amino acid compositions at key positions (1, 7, 9–12). Most of these mutations change hydrophobic residues to polar residues and occur at VH positions that would have interacted with either the VL or CH1 domains had they been present in a classical Ab-like orientation (1, 10). cAb VH X-ray crystal structures show the usual Ig fold, typically most similar to the human VH of family III (13). However, considerable differences have been observed in the CDRs, and some long CDR H3s bend and make contacts with the framework region of the cAb VH, which, in a classical Ab, would have been in contact with VL (12). CDRs play a central role in Ab–Ag recognition; thus, cAb VH structures with the
biologically relevant conformations of the unique CDR loops are required to understand cAb V_{H}–Ag interactions.

Unfortunately, experimental structure determination using X-ray crystallography or nuclear magnetic resonance is laborious, time-consuming, and expensive, resulting in a gap between the number of available protein sequences and structures. Furthermore, of ~65,000 protein structures present in the Protein Data Bank (PDB) (14), there are only ~1,100 Ab structures, of which ~50 are cAb V_{H} structures. The paucity in cAb V_{H} structures combined with the reliance on homology modeling for computational design of humanized Abs for production of at least 11 marketed classical Abs (5), including Herceptin (trastuzumab or humanized anti-HER2), Zenapax (daclizumab or humanized anti-Tac), and Avastin (bevacizumab or humanized anti-vascular endothelial growth factor), highlights the need for a high-resolution cAb V_{H} homology modeling tool.

We previously created RosettaAntibody (15), a homology modeling tool for classical Ab variable regions (fragment variable region of Ig). RosettaAntibody assembles the sequence-match-based templates for the heavy and light chain framework and the canonical CDRs L1, L2, L3, H1, and H2 templates, followed by ab initio modeling of the CDR H3 loop and subsequent optimization of the V_{L}–V_{H} relative orientation and all CDR loop conformations. Although the median global root-mean-square deviation (rmsd) for short CDR–H3 loops (<10 residues) was <2.0 Å, the prediction for longer CDR H3 loops was worse with median global rmsds up to 6.0 Å. Some of the best loop modeling protocols like Protein Local Optimization Program (16) and inverse kinematic loop building (17) generate 8–13 residue loops of sub-Angstrom to 3-Å accuracy. However, these algorithms are computationally expensive and limited to short loops in a native environment. Given the poor performance in modeling long loops and the nonnative environment of the CDR H3 loop in a homology model, building cAb V_{H} CDR H3 loops, which average in length of 16 residues (2) (human and murine CDR H3s average 14 and 12, respectively (2)), is expected to be quite challenging. Another fragment variable region of Ig homology modeling program, the Prediction of Ig Structure server (18), efficiently grafts a CDR H3 structure with the highest sequence homology but cannot predict novel conformations. Despite uncertainties in CDR H3 predictions, we have demonstrated that flexible backbone docking strategies like EnsembleDock (19) and SnugDock (20) can sometimes compensate for errors in RosettaAntibody homology models by optimizing the paratope for the successful prediction of high-resolution Ab–Ag interaction complexes.

In this paper, we analyze the sequences and structures of known cAb V_{H} domains and develop a RosettaAntibody-based cAb V_{H} homology modeling tool. We test whether the canonical numbering schemes can be applied to cAb V_{H} domains, and where they fail, we seek to identify new distinguishing markers. Similarly, we test the classification of canonical loop conformations and seek to update the repertoire appropriately. For H3 loops, we seek conserved structural features that can alleviate the challenge of the long lengths. Finally, we test a combined homology modeling procedure and comment on the usefulness of the models.

Materials and Methods

Test set

All cAb V_{H} structures were downloaded from the PDB (14) as of November 10, 2009. The 46 downloaded structures were filtered for redundancies (three or fewer point mutants). For cases where both the bound and unbound forms of the cAb V_{H} were present, only the unbound was retained, resulting in 27 unique cAb V_{H} structures (Supplemental Fig. 1). Shark, camelized human, and humanized camelid HCabs were not included in the test set. For each downloaded structure, the cAb V_{H} domain was manually identified and extracted for subsequent analysis. The test set contains 17 camel and 10 llama cAbs, with CDR H3 loop lengths ranging from 8 to 24 residues (Table 1). For comparison with V_{H}, we extracted the heavy chain from the 54 Abs in the RosettaAntibody test set (15).

Sequence and structure analysis

The amino acid sequences in FASTA (21) format were derived from the PDB files of the cAb V_{H} and V_{D} domains and aligned using the MUSCLE (version 3.7) (22) multiple sequence alignment feature in multiple sequence alignment view (version 4.2) (23) graphical multiple alignment tool. The alignment was manually edited to ensure that reported conserved Ab residues were aligned correctly, especially for the regions immediately preceding and following the CDR loops. Sequence features were identified by visual inspection of the multiple alignment (Supplemental Fig. 1). The cAb V_{H} structures were visualized by using PyMOL (24) to identify the structural features. PyMOL was also used to compute the C_{α}–C_{α} distances of the cysteine residues in CDR H1 and H3 that form a disulfide bond. For all other distance and dihedral angle measurements, the Rosetta biomolecular modeling software was used.

We considered supplementing the amino acid sequences of the test set with cAb sequl for which crystal structures were not available. Although we found an additional 136 llama Ab sequences at http://ncbi.nlm.nih.gov, it was not clear whether they were from classical Abs or cAbs. To avoid confusing detection of cAb signatures, only the sequences from PDB structures of cAb V_{H} domains were used.

Homology modeling

The homology modeling protocol follows RosettaAntibody to create models by 1) identifying homologous framework and loop templates from the RosettaAntibody database by maximum basic local alignment search tool (BLAST) (25) bit score, 2) grafting CDR templates onto the framework, 3) building the CDR H3 loop, and 4) globally refining the paratope. The main differences from the standard RosettaAntibody protocol are highlighted. The cAb V_{H} structures were appended to the RosettaAntibody Ab database. Except where noted, we ensured that the homology modeling protocol was blind by removing query crystal structures from the database. Similar to the four-residue C-terminal CDR H3 fragment library used in modeling CDR H3 loops in V_{H} Abs (15), we created a six-residue C-terminal CDR H3 fragment library specific to cAbs. The six-residue fragments have been classified as “stretched” or “twisted” as described in Results, and those that could not be classified are referred to as “neutral”. The template identification is similar to that in RosettaAntibody. Because of the absence of the light chain, templates for the light chain framework and CDRs are omitted, and V_{L}–V_{H} assembly is unnecessary.

On successful identification of the respective templates, the CDRs are grafted into the framework as described previously (15). Side-chain conformations of the grafted loop and the neighboring residues are optimized by rotamer packing (26). In a few cases, grafting creates a broken loop because of framework deviations [also observed in a few cases in canonical Abs submitted to the RosettaAntibody server (27)]. When grafting breaks loops, the loops are repaired by minimal refinement using a combination of small (28), shear (28), and cyclic coordinate descent (29) moves with side-chain packing following the high-resolution CDR H3 loop refinement in RosettaAntibody (15) without side-chain minimization (30).

CDR H3 loop modeling and H1 refinement

The loop modeling algorithm follows that in the RosettaAntibody protocol composed of 1) a centroid pseudo-atom side-chain representation (31) low-resolution Monte Carlo stage where diverse loop conformations are sampled by large perturbations by fragment insertion (including the stretched-twisted and twisted fragments), and 2) an all-atom high-resolution Monte Carlo-plus-minimization stage where all the side-chain conformations are optimized and the loop backbone dihedral angles are perturbed minimally to relieve steric clashes. To model cAb V_{H} domains with extremely long CDR H3 loops and the larger diversity of CDR H1 loops, the RosettaAntibody protocol was enhanced as follows.

Two bounded harmonic potential terms are added to the scoring function. The first constraint enforces the disulfide bond if cystines are present in CDR H1 and H3 loops. The second ensures that stretched-twisted structures fold such that the n − 5 residue of CDR H3 is near to residue 46 in the heavy chain framework. The constraint term is:
ideal CDR H3 bond lengths and bond angles and stretched torsion angles are included for final discrimination and ranking of homology models. Observed structural features during the course of the search but is not in high-resolution stages, respectively. The term penalizes deviations from the maximum over the range of observed variable, and $\sigma$ is the half-width of the well. For the disulfide bond formation, $x$ is the distance between the C atoms of the cystine residues forming the disulfide bond, and $S_{\alpha}$, $S_{\beta}$, and $S_{\gamma}$ are 4.0, 6.1, and 0.6 Å, respectively (Supplemental Table I). For the stretched-twist conformation, $x$ is the distance between the C atoms of the $n = 5$ CDR H3 residue and the 46th residue, and $x_{\min}$, $x_{\max}$, and $\sigma$ are 6.5, 9.1, and 0.7 Å, respectively (Table II, last column, "Stretch Tw Std. Dev."). In each case, $x_{\min}$ and $x_{\max}$ are the respective minimum and maximum over the range of observed variable, and $\sigma$ is the SD of the variable. The weight of the constraint term is 100 means that a sequence-homology–based template was not identified, a template CDR of the same length was grafted, and 100 means that the native CDR was grafted. Draft rmsd measures the lower bound after loop coordinates to the homology framework, and Homology LowE rmsd measures loop accuracy of the LowE model after CDR H1 refinement. For the CDR H3, measures are given for building loops In Native Context and on the Homolog framework with predicted CDRs. LowE indicates the lowest-energy model. LowRMS indicates the model with the lowest global rmsd in the 10 lowest-energy models. LowALL indicates the lowest global rmsd observed among all models.

The glycine at residue 26 was disordered in the crystal structure of 1RJC; thus, the H1 rmsd is computed against the ordered residues of the loop and excluded from the median summary.

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>CDR Length</th>
<th>CDR H1</th>
<th>CDR H2</th>
<th>CDR H3</th>
<th>Seq.ID (%)</th>
<th>Graft rmsd (Å)</th>
<th>Homology LowE rmsd (Å)</th>
<th>Seq.ID (%)</th>
<th>Graft rmsd (Å)</th>
<th>Homology LowE rmsd (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1IEH</td>
<td>10</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>88</td>
<td>3.8</td>
<td>3.9</td>
<td>88</td>
<td>4.1</td>
<td>3.1</td>
</tr>
<tr>
<td>1G9E</td>
<td>10</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>88</td>
<td>4.0</td>
<td>3.3</td>
<td>88</td>
<td>2.1</td>
<td>1.6</td>
</tr>
<tr>
<td>1HCV</td>
<td>10</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>88</td>
<td>3.1</td>
<td>3.1</td>
<td>88</td>
<td>0.6</td>
<td>1.3</td>
</tr>
<tr>
<td>3CPP</td>
<td>10</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>67</td>
<td>1.2</td>
<td>1.2</td>
<td>67</td>
<td>0.7</td>
<td>1.8</td>
</tr>
<tr>
<td>1YC7</td>
<td>8</td>
<td>7</td>
<td>10</td>
<td>0</td>
<td>71</td>
<td>5.6</td>
<td>5.3</td>
<td>71</td>
<td>3.5</td>
<td>5.5</td>
</tr>
<tr>
<td>1ISX</td>
<td>10</td>
<td>7</td>
<td>11</td>
<td>1</td>
<td>93</td>
<td>0.7</td>
<td>1.8</td>
<td>93</td>
<td>0.7</td>
<td>1.8</td>
</tr>
<tr>
<td>3EZ7</td>
<td>10</td>
<td>7</td>
<td>12</td>
<td>1</td>
<td>93</td>
<td>1.2</td>
<td>1.2</td>
<td>93</td>
<td>0.7</td>
<td>1.8</td>
</tr>
<tr>
<td>1BZQP</td>
<td>10</td>
<td>8</td>
<td>12</td>
<td>0</td>
<td>76</td>
<td>3.5</td>
<td>3.4</td>
<td>76</td>
<td>1.5</td>
<td>3.0</td>
</tr>
<tr>
<td>2P49</td>
<td>10</td>
<td>8</td>
<td>12</td>
<td>0</td>
<td>76</td>
<td>3.4</td>
<td>3.1</td>
<td>76</td>
<td>1.7</td>
<td>3.2</td>
</tr>
<tr>
<td>1IZV</td>
<td>10</td>
<td>8</td>
<td>12</td>
<td>0</td>
<td>88</td>
<td>5.3</td>
<td>4.9</td>
<td>88</td>
<td>3.2</td>
<td>2.5</td>
</tr>
<tr>
<td>1KXQ</td>
<td>7</td>
<td>8</td>
<td>14</td>
<td>0</td>
<td>65</td>
<td>0.7</td>
<td>1.1</td>
<td>65</td>
<td>0.7</td>
<td>1.1</td>
</tr>
<tr>
<td>2RRP</td>
<td>10</td>
<td>8</td>
<td>14</td>
<td>0</td>
<td>90</td>
<td>1.2</td>
<td>1.2</td>
<td>90</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>1OPP</td>
<td>7</td>
<td>8</td>
<td>15</td>
<td>0</td>
<td>76</td>
<td>0.9</td>
<td>0.8</td>
<td>76</td>
<td>0.9</td>
<td>0.8</td>
</tr>
<tr>
<td>1JLYL</td>
<td>10</td>
<td>13</td>
<td>15</td>
<td>0</td>
<td>100</td>
<td>1.6</td>
<td>1.1</td>
<td>100</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>1IZVP</td>
<td>10</td>
<td>8</td>
<td>16</td>
<td>8</td>
<td>36</td>
<td>3.3</td>
<td>3.3</td>
<td>36</td>
<td>0.8</td>
<td>4.5</td>
</tr>
<tr>
<td>1U0Q</td>
<td>10</td>
<td>8</td>
<td>16</td>
<td>7</td>
<td>27</td>
<td>2.7</td>
<td>2.7</td>
<td>27</td>
<td>0.7</td>
<td>2.8</td>
</tr>
<tr>
<td>3DWT</td>
<td>11</td>
<td>8</td>
<td>16</td>
<td>6</td>
<td>4.5</td>
<td>3.6</td>
<td>3.6</td>
<td>4.5</td>
<td>1.1</td>
<td>3.3</td>
</tr>
<tr>
<td>1QBC</td>
<td>13</td>
<td>8</td>
<td>16</td>
<td>100</td>
<td>0.4</td>
<td>1.8</td>
<td>1.8</td>
<td>100</td>
<td>1.4</td>
<td>0.9</td>
</tr>
<tr>
<td>1MVPF</td>
<td>8</td>
<td>7</td>
<td>19</td>
<td>3</td>
<td>29</td>
<td>5.2</td>
<td>4.4</td>
<td>29</td>
<td>5.2</td>
<td>4.4</td>
</tr>
<tr>
<td>1IR8</td>
<td>10</td>
<td>7</td>
<td>17</td>
<td>1</td>
<td>73</td>
<td>1.0</td>
<td>1.0</td>
<td>73</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>1IZVY</td>
<td>10</td>
<td>8</td>
<td>18</td>
<td>2</td>
<td>2.7</td>
<td>2.3</td>
<td>2.3</td>
<td>2.7</td>
<td>1.6</td>
<td>4.3</td>
</tr>
<tr>
<td>1FZX</td>
<td>10</td>
<td>6</td>
<td>19</td>
<td>7</td>
<td>4.4</td>
<td>3.6</td>
<td>3.6</td>
<td>4.4</td>
<td>3.6</td>
<td>3.6</td>
</tr>
<tr>
<td>1RIC</td>
<td>10</td>
<td>7</td>
<td>19</td>
<td>60</td>
<td>4.5</td>
<td>3.2</td>
<td>3.2</td>
<td>60</td>
<td>0.7</td>
<td>6.7</td>
</tr>
<tr>
<td>1TQ0</td>
<td>10</td>
<td>5</td>
<td>24</td>
<td>70</td>
<td>4.1</td>
<td>3.3</td>
<td>3.3</td>
<td>70</td>
<td>0.7</td>
<td>6.7</td>
</tr>
<tr>
<td>1MEL</td>
<td>10</td>
<td>8</td>
<td>24</td>
<td>70</td>
<td>3.9</td>
<td>2.9</td>
<td>2.9</td>
<td>70</td>
<td>0.7</td>
<td>6.7</td>
</tr>
<tr>
<td>1XFQ</td>
<td>10</td>
<td>8</td>
<td>24</td>
<td>70</td>
<td>3.9</td>
<td>2.9</td>
<td>2.9</td>
<td>70</td>
<td>0.7</td>
<td>6.7</td>
</tr>
<tr>
<td>1ZMYW</td>
<td>10</td>
<td>8</td>
<td>24</td>
<td>70</td>
<td>3.9</td>
<td>2.9</td>
<td>2.9</td>
<td>70</td>
<td>0.7</td>
<td>6.7</td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.6</td>
<td>3.1</td>
<td></td>
<td>1.5</td>
<td>3.9</td>
</tr>
</tbody>
</table>

where $x_{\min}$ and $x_{\max}$ are the lower and upper bounds, and $\sigma$ is the half-width of the well. For the disulfide bond formation, $x$ is the distance between the C atoms of the cystine residues forming the disulfide bond, and $S_{\alpha}$, $S_{\beta}$, and $S_{\gamma}$ are 4.0, 6.1, and 0.6 Å, respectively (Supplemental Table I). For the stretched-twist conformation, $x$ is the distance between the C atoms of the $n = 5$ CDR H3 residue and the 46th residue, and $x_{\min}$, $x_{\max}$, and $\sigma$ are 6.5, 9.1, and 0.7 Å, respectively (Table II, last column, "Stretch Tw Std. Dev."). In each case, $x_{\min}$ and $x_{\max}$ are the respective minimum and maximum over the range of observed variable, and $\sigma$ is the SD of the variable. The weight of the constraint term is 100 means that a sequence-homology–based template was not identified, a template CDR of the same length was grafted, and 100 means that the native CDR was grafted. Draft rmsd measures the lower bound after loop coordinates to the homology framework, and Homology LowE rmsd measures loop accuracy of the LowE model after CDR H1 refinement. For the CDR H3, measures are given for building loops In Native Context and on the Homolog framework with predicted CDRs. LowE indicates the lowest-energy model. LowRMS indicates the model with the lowest global rmsd in the 10 lowest-energy models. LowALL indicates the lowest global rmsd observed among all models.

The glycine at residue 26 was disordered in the crystal structure of 1RJC; thus, the H1 rmsd is computed against the ordered residues of the loop and excluded from the median summary.

1 and 2, CDR H2 canonical class 1 and 2, respectively; c, cystine bond between CDR H1 and H3; l, bound to small molecule; p, bound to protein Ag; s, stretched-twist; t, twist; u, unbound structure.
cAb V_{H}H structures and compared them with the 54 classical V_{H} structures used for benchmarking RosettaAntibody. First, we summarize our observations based on sequence and then those based on structure.

**Sequence analysis of cAb V_{H}H**

The cAb V_{H}H test set, comprising a modest 27 members, is nevertheless the largest test set used to date for sequence alignment to detect amino acid patterns unique to cAb V_{H}H domains. These 27 unique cAb V_{H}H sequences make up 22, 23, and 23 unique CDR H1, H2, and H3 sequences, respectively (Supplemental Fig. 1). The low average pairwise sequence identities for CDRs H1, H2, and H3 in our test set of 37, 32, and 20%, respectively, suggest that the test set comprises diverse CDR sequences. In addition, the highest pairwise sequence identity of nonredundant CDR loops for H1, H2, and H3 is 70, 86, and 42%, respectively.

Although many differences have been previously reported in the analysis of individual cases (7, 9–12), some new differences arise by visual inspection of the sequence alignment of the cAb V_{H}H and V_{H} test sets (summarized in Table III, Supplemental Fig. 1). cAb V_{H}H positions, which differ, include residues 23, 49, 84, 87, 105, 108, and 109 (all specific residue numbers in this paper follow the Chothia convention (32)). Their structural locations are shown in Fig. 1 along with a superposition with a classical Ab to show the location with respect to the classical light chain and the classical C_{L} domain (absent in cAbs). Several positions have been reported before, and some new sequence differences are now apparent. At position 29, previously reported to be aspartic acid, glycine, asparagine, or serine (10), we see wider variation and, in a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine. Position 45 was previously noted as a quarter of cAb V_{H}Hs, tyrosine.
contributes to the enhanced solubility of the cAb V_hH domain (1).

Relatedly, classical Abs frequently have a surface-exposed lysine 23 in the first framework region, but the cAb V_hH test set did not have any lysines in this position. The Eris server (33) predicts an average ΔΔG of 1.8 kcal/mol upon mutation of alanine to lysine, suggesting that the mutation away from lysine stabilizes the cAb V_hH relative to a V_hH (11). The lysine would be also expected to aid solubility, but apparently, it is not necessary in cAbs perhaps because of other solubility-enhancing mutations.

Alanine occupies cAb V_hH position 49 (one llama V_hH [1U0Q] had a glycine) in the middle of an anti-parallel β-sheet, whereas classical Abs frequently have glycine there, proximal to the apex of the CDR L3 loop. If position 49 is alanine or glycine, the correlated mutation position 69 can be either [IVLT] or [IVLFJM], respectively. Thus, when position 69 is phenylalanine or methionine (observed only in classicals), position 49 must be glycine to accommodate the larger volume. Another probable pair of correlated mutations is positions 14 and 84 (Cα–Cα distance ∼6 Å), which exhibit, respectively, predominantly alanine in cAb V_hHs or proline in V_hHs and proline in cAb V_hHs or serine in V_hHs. An anti-parallel sheet is formed by the β-strands for proper hydrogen-bond formation.

Residue 109 in the last (C-terminal) β-sheet of the variable heavy region is occupied by valine in cAb V_hHs, whereas classical Abs exhibit mostly valine and some leucines. If leucine occurred at position 109 in a cAb V_hH domain, it would clash with the cAb V_hH-conserved leucine at position 18. In contrast, in classical V_hH domains, leucines in position 109 are accommodated by a valine at position 18. Thus, the 18–109 pair is always leucine–valine in cAb V_hHs, whereas it can be valine–leucine, leucine–valine, or valine–valine in classical V_hHs. A few classical V_hH domains (1VFA, 2ADG, 2AEP, and 2H1P) have an apparent sterically unfavorable leucine–leucine pair. To compensate for the larger volume requirement of the two leucines, the neighboring position 82 for such V_hHs is always methionine. Position 82 can be isoleucine, leucine, or methionine in classical V_hHs, but it is always a methionine in cAb V_hHs. In summary, the three neighboring residues 18, 82, and 109 are conserved leucine, methionine, and valine in cAb V_hH domains, whereas the classical V_hH domains vary. The variation in classical V_hH domains is possible because of a slightly larger separation between the

| Table III. Observed sequence differences in cAb V_hH domains relative to V_hH domains |
|-----------------|--------------|-----------------|
| Chothia Residue | Classical     | cAbs            | Summary (Reported and Newly Discovered) |
| No.             | Amino Acids   |                 |                                |
| 11^FR           | A/LV/–       | S/L/–           | L–S^10                         |
| 14^FR           | P/–/AL       | A/P/–           | P–A, only two V_hHs have P      |
| 18^FR           | A/LV/–       | T/–/–           | [LV]–L                         |
| 23^CDR          | G/AT/PQSV    | –/AT/E^1 Q^1 S^1 | A, K most common in classicals; K absent in camelids |
| 29^CDR          | –/FILVY     | –/ADPGNSVY/–   | [FILV[→][DGNS]_V]               |
| 37^FR           | –/VH/AF     | –/FYNYIP       | V→E^11 V→[FY]                      |
| 44^FR           | G/AR/KS/E   | E/GOA^1 A^1 D^1 | G→E^11, G→[EQ], other           |
| 45^FR           | –/LF/LF     | R/LF/LF        | L→R; Occasional C reported in cDNA^10 |
| 47^FR           | –/W/FLY     | G/AFLSW/γ^1 y   | W→G, W→[GSLF]^12                |
| 49^FR           | G/AVWS      | A/SV/γ^1       | G→A                            |
| 82^FR           | –/ILM/–     | M/–/–          | [ILM]→M, I occurs rarely        |
| 84^FR           | S/ANT/S     | P/SG^1 L^1 R^1 | S→P, only two V_hHs have S      |
| 87^FR           | –/ST/–      | –/ST/–         | Camelds contain only T           |
| 94^FR           | R/KS/GHPV   | A/K/G^1 T^1 T^1 | R→noncharged, disrupts salt bridge to residue 101^10 |
| 103^FR          | –/W/–       | –/RW^1         | W→R^2                          |
| 105^FR          | –/AQ/T/HR   | Q/–/–          | [AQ]/→Q, Camelds contain only Q |
| 108^FR          | T/L/S/M     | Q/–/–          | [LTS]→Q                        |
| 109^FR          | V/L/–       | V/–/–          | L→V                            |

Underlined residues indicate new observations, and those that occur in most cAb V_hH are summarized in bold. Camel (c) or llama (l) residues observed only once.

→, absence of a particular amino acid in its respective occupancy category; CDR, CDR; FR, FR. framework.
strands containing positions 18 and 109 relative to cAb V_{1H} strands.

The reasons for the following mutations are less apparent. The 87th residue lying in a surface exposed loop is always threonine in cAbs but is equally occupied by both serine and threonine in classical Abs. Residue 105 in a surface-exposed loop C-terminal to the CDR H3 loop (that ends in residue 102) has only glutamines in cAb V_{1H}s but frequently exhibits alanine, glutamine, and threonine in classicals.

Loop identification

Standardized numbering systems (32, 34) serve as alignments for clear identification of CDR loop locations. However, the classical numbering systems have not been tested on cAb V_{1H} sequences, and servers like Abnum (35) that routinely number Ab sequences fail on most cAb V_{1H} sequences because of differences in the loop stems. On the basis of the visual inspection of the sequences and similarity to the canonical loop definitions, we revised the rules for cAbs. Our new rules are detailed in Table IV.

Structural analysis of CDR loops

Fig. 2 compares the structural features of whole cAb V_{1H} domain and similarity to the canonical loop definitions, we revised the loop stems. On the basis of the visual inspection of the sequences and similarity to the canonical loop definitions, we revised the rules for cAbs. Our new rules are detailed in Table IV.

CDR H1 and H2

CDR H1 and H2 loop conformations in classical Abs can typically be identified using the canonical Chothia rules (36) (http://www.bioinf.org.uk/abs/chothia.html); however, these rules have not been tested on cAbs. Upon testing these rules for CDR H1, we first notice that some sequences preclude classification because of differences in loop length (Table IV): canonical CDR H1 classes 1, 2, and 3 have 10, 11, and 12 residues, respectively; however, cAb CDR H1s sometimes contain 7, 8, 9, or 13 residues. Even when loop length matches, the canonical conformations do not cover the span of H1 conformations: Fig. 2B shows the 20 of the 27 cAb V_{1H} CDR H1 loops with length of 10. Most deviate significantly from the two 10-residue canonical structures (class 1, 2FB1; class 1b, 7FAB/3HFM).

Like the H1 loop, the cAb H2 loop precludes classification based on length: some cAb V_{1H} H2 loops have 6 or 13 residues, whereas the classical CDR H2 loop classes 1, 2, and 4 make up 7, 8, 8, and 10 residues, respectively. In contrast to the H1 loop, when the H2 loop length does match, the canonical conformations for H2 loops are often useful. Twelve of the 19 cAbs that have eight-residue CDR H2 loops fit canonical class 2 (represented by 1BBD); four of the five cAbs that have seven-residue CDR H2 loops fit canonical class 1 (1GIG). Fig. 2C shows the structural diversity of cAb CDR H2 loops compared with those of classicals.

CDR H3

The CDR H3 is the most variable loop in terms of amino acid composition, length (37), and structure. Although CDR H3 loops largely elude structural classification, for classical Abs, Shirai et al. (38) detected conserved structural motifs in the C-terminal stem. We previously incorporated the rules for CDR H3 modeling in classical Abs (15). For cAbs, we analyzed the backbone torsion angles for the Shirai motifs. Unfortunately, neither the Shirai kink nor the extended conformation is present in cAb H3 loops. We did, however, note the previously observed disulfide bond between CDR loops H1 and H3 (10) and two structural motifs that occur in a subset of H3 loops.

Disulfide bond between CDRs H1 and H3. Nine Abs in the test set exhibit a disulfide bond between the CDR H1 and H3 loops (10) (Fig. 1, Table I), and as previously noted, all of these are camel Abs (7). The mean distance between the C_{\alpha} atoms of the cystine residues forming the disulfide bond is 5.6 ± 0.6 Å (Supplemental Table I). The CDR H1 cystine residue involved in the disulfide bond formation is always residue 33, with a single exception (anti-VSG cAbAn33, 1YC7 (39), where it is residue 32). Although no such conserved residue was observed for the corresponding cystine residue in the CDR H3 loop, for the seven CDR H3 loops longer than 17 residues, the cystine always occurred in or immediately N-terminal to the middle of the loop.

Stretched-twist structural feature. A structural superposition of the cAb V_{1H} framework region reveals that the C-terminal region of 17 of the 27 cAb V_{1H} CDR H3 loops exhibits a conserved structural motif (red segments in Fig. 2A). To quantify the observation, we calculated the φ and ψ backbone dihedral angles of the conserved subset (Table II). For residues n + 1 through n − 4, the backbone φ and ψ angles are very similar for all the members exhibiting the conserved structural motif (CDR H3 residues are numbered 1 to n, where H3 residue 1 corresponds to residue 95 and residue n to 102). We define the motif by residues with backbone dihedral angles with SD under 30˚ across the subset. All backbone dihedral angles of these residues meet this criterion, except φ of n − 3 and ψ of n − 4. The SD of the dihedral angles of residues n + 1 through n − 4 of the subset is one-third of the entire data set. Because the conserved feature has a sharp twist near the C terminus and then stretches to reach the heavy framework, we refer to the motif as a “stretched-twist” (Fig. 2A).

Stretched-twisted CDR H3s bend and contact the region of the framework that would have been in contact with the light framework in a classical Ab. The mean (and SD) C_{\alpha}–C_{\alpha} distances between the CDR H3 apex residues, n − 4 and n − 5, and the nearest cAb V_{1H} framework residues (residue 46) are 11.3 ± 0.8 and 7.9 ± 0.7 Å, respectively (black line in Fig. 1, Table II). The corresponding SDs over the entire test set were much higher, 3.5 and 5.5 Å, respectively, illustrating that the distances are not conserved over the entire test set. Because the mean (and SDs) were obtained from our data set, which includes both Ag-bound and unbound cAbs, the observed C_{\alpha}–C_{\alpha} distances do not rely on Ag binding. The similarly conserved C_{\alpha}–C_{\alpha} distances in the unbound llama cAb V_{1H} A52 structure (1I3V) (40) and the con-
responding bound structure with dye RR1 (1I3U) (40) further reinforces that the observed Cα-Cα distances are conserved for “stretched twisted” CDR H3s and are independent of Ag binding.

Twist structural feature. Six additional structures exhibit only the sharp turn of the stretched-twist motif in residues n + 1, n, n − 1, and n − 2 (green segments in Fig. 2A, Table II). All four residues have conserved backbone dihedrals with an SD > 30° across the set, except for the n − 1 φ angle. The SDs are higher in this structural feature than that in the stretched-twist, indicating more diversity. We refer to this feature alone as a “twist.” Structures that cannot be classified as either “stretched-twist” or “twist” are referred to as “neutral.”

Sequence–structure rules. Upon examining the sequences, structures, and environments of the subsets of H3 loops exhibiting conserved motifs, we found several useful signatures. Examining the sequences, structural feature than that in the stretched-twist, indicating more diversity. We refer to this feature alone as a “twist.” Structures that cannot be classified as either “stretched-twist” or “twist” are referred to as “neutral.”

Sequence–structure rules. Upon examining the sequences, structures, and environments of the subsets of H3 loops exhibiting conserved motifs, we found several useful signatures. Examining the sequences, structural feature than that in the stretched-twist, indicating more diversity. We refer to this feature alone as a “twist.” Structures that cannot be classified as either “stretched-twist” or “twist” are referred to as “neutral.”

Sequence–structure rules. Upon examining the sequences, structures, and environments of the subsets of H3 loops exhibiting conserved motifs, we found several useful signatures. Examining the sequences, structural feature than that in the stretched-twist, indicating more diversity. We refer to this feature alone as a “twist.” Structures that cannot be classified as either “stretched-twist” or “twist” are referred to as “neutral.”

...
V1H test set that has more than two consecutive aromatic amino acids in the CDR H3 C-terminal positions n − 2 through n + 1. By virtue of having a histidine at position n − 1 (in between two tyrosines), the steric force the CDR H3 loop to fold atypically toward the framework. The other structure to have more than two consecutive aromatic residues in positions n − 2 through n + 1 is the camel Ab CABAM9D (1KXQ) (43), with all four positions occupied by aromatic amino acids. However, although the three aromatic residues in 1F2X resulted in an atypical bulge, the presence of the fourth consecutive aromatic residue reverses the chain orientation, so 1KXQ adopts a stretched-twist conformation.

Rule 2d was included to incorporate the llama Ab (2BSE) (44), the only Ab in the cAb V1H test set that has a glycine at position n. The flexibility afforded by glycine in the CDR H3 loop’s C terminus enables the loop to extend away from the protein body, contrary to other stretched-twist structures, which bend toward framework residue 46. Interestingly, the three other twisted Abs that satisfy the first clause of rule 2 and have a tyrosine or tryptophan at n − 2 position (1F2X, 1G9E, and 1HCV) have a glycine in either of N-terminal CDR H3 stem residues 93 and 94. However, in addition to these three Abs, another Ab in the cAb V1H test set (1ZVY) has a glycine at position 93 and adopts the stretched-twist conformation.

Possible existence of CDR 4

Because the region between residues 71–78 is close to the other CDRs, it has been suggested that it combines with CDRs H1, H2, and H3 to form a larger paratope (45, 46). In addition, affinity maturation studies involving mutations in this region affected Ag binding (47, 48), suggesting that CDR 4 could be recruited in Ag binding. A structural alignment (Fig. 2E) shows that the loop formed by heavy chain residues 71–78 has a significantly larger structural divergence than in the V1H loops. To determine the importance of CDR 4 in Ag binding, we examined the CDR 4–Ag contacts in the Ab–Ag complexes in the test set. Of the 19 Ab–Ag complexes, only 2 (1KXQ and 1SIX) show CDR 4 contacts to Ag. Apparently, CDR 4 is capable of but not critical for Ag interactions.

Homology models

With the new rules and observations about cAb V1H domains, we created a tailored algorithm to model cAb V1H domains starting from their sequences. The method, based on RosettaAntibody, uses structural fragments from a database and a Monte Carlo-plus-minimization structure prediction and refinement algorithm, and it incorporates the newly observed structural features either as constraints or by selecting appropriate database fragments in model construction. cAb V1H coordinates were added to the RosettaAntibody database, and homology models are assembled with templates and loops (when available) from this database. Matching cAb CDR loops are grafted onto a selected framework. For testing each target in a “blind” manner, the native structure and templates with exact sequence match over the entire length of the query sequences (framework, CDRs H1 and H2) were removed from the database. The CDR H3 loop is built using RosettaAntibody’s low- and high-resolution Monte Carlo-plus-minimization-based loop building techniques (28). Because the CDR H1 also exhibits larger conformational diversity than that reported in classical Abs, the CDR H1 loop is also subjected to Monte Carlo perturbation and refinement.

Grafting CDRs

CDRs H1 and H2 are grafted to create cAb V1H homology models in a manner similar to that used for canonical CDR loops using RosettaAntibody. Because of the noncanonical nature of most cAb V1H CDR H1 and H2 loops, traditional canonical class-based template selection fails to identify template loops for such cases. Instead, RosettaAntibody selects template loops based on BLAST bit scores, and thus, it is not limited by class definitions and identifies templates as long as similar sequences can be found in the database. The median global rmsd for cAb V1H CDRs H1 and H2 (3.6 and 1.5 Å, respectively; Table 1) is higher than that obtained by using RosettaAntibody to graft canonical CDRs into V1H (0.84 and 0.93 Å, respectively) (15). The CDR V1H CDR H1 exhibits larger conformational diversity than classical Abs, and the available CDR H1 templates deviate from the native structures. To improve the CDR H1 prediction accuracy, we later subject it to explicit perturbations.

Although the selection strategy succeeds for most targets, in a few cases, the BLAST search for similar loop sequences in the database does not return any matches. Several cases (1Y7C and 1ZV5) can be addressed by grafting a length-matched sequence; these often have poor accuracy (median rmsd ~3 Å). The CDR V1H-R2 anti-RR6 Ab (1QD0) (49) still fails because it is the only Ab in the database to have a 13-residue CDR H1. Similarly, the CDRs with the shortest (6 residue) and the longest (13 residue) CDR H2 loops (camel Ab CAB-CAB-CA05, 1F2X (42); llama Ab A52, I13V (40)) are unique and do not have length-matched loop templates. For testing purposes, in these individual cases, we grafted the native loop structure. In truly blind tests, de novo methods and grafting methods, which do not require length matching, may be useful (50).

Modeling CDR H3

As in classical Abs, the cAb CDR H3 loop is hypervariable and must be modeled de novo. In this section, we isolate the loop modeling algorithm performance by building the CDR H3 in the native environment; that is, we start with the crystal structure, remove the CDR H3 residues, and rebuild the loop. We illustrate the CDR H3 modeling strategy with a representative cAb V1H, cabbcII-10 (1ZMY) (51). CabbcII-10 has a 24-residue CDR H3 loop with a disulfide bond between the CDR H1 and H3 loops and exhibits the stretched-twist conformation.

We tested four variants of algorithms for modeling the CDR H3 loop. Initially, we used a nine-residue fragment-insertion Monte Carlo strategy and did not use any constraints to model observed structural features. For cabbcII-10, this strategy sampled CDR H3 conformations as low as 4.8 Å global rmsd, with the lowest-energy (LowE) model having a global rmsd of 7.8 Å. Next, to test the usefulness of the observed structural features, we performed the loop building simulations while also incorporating constraints that bias formation of the disulfide bond and the approach of the apex of the stretched-twist loop to the framework (Materials and Methods). The inclusion of constraints improved the CDR H3 global rmsd of the LowE model to 4.6 Å and enabled sampling of conformations as close as 3.6 Å. To further improve the fine sampling, we incorporated additional Monte Carlo steps using three-residue fragment insertions following the nine-residue fragments. The finer sampling produced a more accurate LowE CDR H3 with global loop rmsd of 3.6 Å.

To test whether further improvement is limited by sampling or scoring, we compared the model energies to energies of native structures subjected to the same high-resolution refinement stages. Refined native loops scored better than all model loops (Supplemental Fig. 2), implying that better models, if sampled, could be identified by the energy function. Therefore, we tested whether increasing the number of models from 5,000 to 20,000 could produce more accurate models. The increased sampling slightly
improved the rmsd of structures sampled (lowest rmsd in the entire set of models improved from 3.0 to 2.5 Å), but the difference was not significant enough to improve the LowE performance, which had a median of 4.0 Å. Because of the computational expense of modeling long loops, we limited the number of models to 5000 loops in the final protocol. The time required to build one model is ~10 min on 1 CPU, which, for one simulation involving 5000 models, translates to ~35 CPU days or approximately 3 wall-clock hours on a 300-CPU cluster.

Results on the full test set are given in Supplemental Table II for the four variants of the algorithm, and Table I shows the results of CDR H3 modeling using the final protocol. With the final protocol, the low-energy model loops range from 0.9 to 6.9 Å rmsd, with a median of 3.9 Å. If the best loop in the 10 LowE models is allowed, the mean rmsd falls to 3.4 Å. The most native-like loop sampled during 5000 independent runs ranges from 0.9 to 3.9 Å rmsd, with a median of 2.6 Å. The test set contains 13 loops with CDR H3 loop lengths between 16 and 24 residues, and incorporating the additional sampling using three-residue fragments improved the median global rmsds for the LowE models from 4.5 to 3.9 Å (Supplemental Table II). In 23 of 27 cases, refined native structures score better (lower) than the LowE model, suggesting that loop sampling still limits the algorithm’s performance. Thus, the final CDR H3 loop building protocol uses the constraints to model the observed structural features, uses fragment-based loop building with 9-residue fragments and 3-residue fragments for loops over 16 residues, and generates 5000 candidate structures.

**Complete homology models**

Complete homology models were generated by building the CDR H3 loop in an environment where the framework and CDRs H1 and H2 have been assembled from sequence-homologous templates. The CDR H1 is also subjected to perturbations, because grafting alone often did not produce near-native structures. Table I shows that the median CDR H3 global rmsd for the test set is 5.4 Å for the LowE models. Fig. 3A shows the diversity of CDR conformations in the 10 lowest-scoring models, illustrating that the ab initio loop modeling generates a wide variety of CDR H3 conformations, whereas the other CDRs show minimal variation because of the absence of fragment insertions to cause significant backbone conformational change. Although these rmsd values are high, Fig. 3B shows that the homology model captures the rough topology of the H3 loop, including the enforced stretched-twisted C-terminal region and the distance between the n − 5 residue of the CDR H3 loop and residue 46 (8.7 Å, within the observed range of 6.5–9.1 Å). The worst-case deviations in CDR H3 loop predictions can be attributed to a poorly modeled environment: in the two structures for which the LowE CDR H3 loop rmsd is >10 Å, at least one of the other grafted CDRs in the respective structures deviate by more than 5.0 Å. In addition, the CDR H3 models closest to the native conformations among all models built have a median global rmsd of 2.7 Å, revealing that the most native-like models still deviate from the native structure. Lower refined-native scores suggest that better sampling is required for more accurate predictions.

The median global rmsds for the LowE models for 8–12, 14–16, and 17–24 residue CDR H3 loops were 5.7, 4.5, and 6.8 Å, respectively. The lower rmsd for the 14–16 residue loops compared with the shorter 8–12 residue loops is surprising because the available conformational space for longer loops is much larger than shorter loops, and loop prediction is expected to be more challenging. This anomaly can be explained by the more accurate environment for the 14–16 residue loops as is evident from the median global rmsds of 2.3 and 1.0 Å for the CDRs H1 and H2, respectively, as compared with higher deviations of 3.3 and 1.5 Å for the respective CDRs in 8–12 CDR H3s. Furthermore, seven of the eight CDR loops of length 14–16 residues can be classified as having the stretched-twisted motif, and thus, the H3 apex constraint can be applied.

To accurately model CDR H3s, we not only minimized the conformations of the neighboring CDR H1 and H2 loops but also explicitly perturbed the CDR H1 loop. The median global CDR H1 rmsd improves from 3.6 to 3.1 Å by incorporating perturbations in addition to grafting.

**Discussion**

We endeavored to create and test, to our knowledge, the first homology modeling protocol tailored for cAb V\(_{H}\) Abs. The protocol is an extension of RosettaAntibody with several unique features. Explicit perturbations to the CDR H1 loop during CDR H3 modeling enabling sampling of the diverse noncanonical CDR H1 structures found in cAb V\(_{H}\). The incorporation of the additional three-residue fragment-based low-resolution loop building following nine-residue fragment-based loop assembly allowed sampling of more native-like low-energy models. Finally, incorporation of constraints to capture the H1–H3 disulfide bond and the approach of the H3 loop apex to the framework improved native-like sampling by preventing predicted loops from adopting conformations with significant deviation from the observed features. The results demonstrate that structural models can be built with gross structural features of the native H3 loop but significant local conformational deviations.

The challenge in modeling very long CDR H3 loops is evident from the high global rmsds, and advancements in computationally efficient loop modeling techniques will be critical to create more accurate cAb V\(_{H}\) homology models. The diversity in the low-energy loop structures emphasizes the large conformational space available to the CDR H3. The challenge of adequately sampling H3 conformations may be amplified as a result of the manner in which the H3 and H1 loops compensate for the absence of the light chain and create Ag recognition properties comparable to classical (dimeric) Abs (7). Forcing the formation of disulfide bonds in applicable cAb V\(_{H}\) domains reduces the conformational space making more native-like predictions possible and probably...
also minimizes the entropic penalty of long loops on binding with Ag (12).

Perturbations to the CDR H1 produced more native-like CDR H1 loops as opposed to simple grafting, but differences still remain between the predicted and the native conformation. Building two neighboring loops using ab initio methods in a nonnative environment is challenging (52, 53), because multiple loop sampling can produce false-positive structures and requires significantly more computational time. The forcing of the disulfide bond between the CDR H3 and H1 loops in respective CDRs also reduced the conformational space of the CDR H1 loop. As the number of cAbs in the PDB increases, we anticipate better starting conformations for grafting homologous CDR H1 loops.

Ab surfaces have many charged residues (54), and particularly, the cAb VH H1 CDR H1 and H3 loops have many charged residues, resulting in dominant inter- and intraloop electrostatic interactions (55), which might not be captured correctly given the difficulty in modeling such interactions (56). However, the most significant hurdle in higher accuracy loop prediction is inadequate sampling. Other algorithms scale the amount of sampling with the number of residues in the loop (e.g., Protein Local Optimization Program (16) samples 2ⁿ loop conformations up to 10⁶, where n is the loop length). The high computational cost of such simulations renders them difficult and more efficient sampling strategies are required.

In contrast to previous studies involving multiple cAb VH H1 sequences where it is uncertain whether a sequence is from a VH H1 or VH H2 (10), the test set created in this study, by virtue of being derived from crystallized cAb VH H1 domains, guarantees that only VH H1 domains are present. cAb VH H1 sequence comparison revealed key residues unique to cAb VH H1 regions, which probably contribute to their desirable characteristics (7), although additional structures will help confirm or generalize the observations. Reconstructing human Abs to incorporate the identified key residues may result in more camel-like human cAbs. The observations will contribute to their desirable characteristics (7), although additional studies are required.

RosettaAntibody modeling as a feature in the RosettaAntibody server (27) and cAbs in the PDB increases, we anticipate better starting conformations for grafting homologous CDR H1 loops.

10 ANALYSIS AND MODELING OF CAMELID VHHs

Disclosures
The authors have no financial conflicts of interest.

References