Contributions of CDR3 to V<sub>H</sub>H Domain Stability and the Design of Monobody Scaffolds for Naive Antibody Libraries

Christopher J. Bond<sup>1</sup>*, James C. Marsters Jr<sup>1</sup> and Sachdev S. Sidhu<sup>2</sup>*

<sup>1</sup>Departments of Medicinal Chemistry, 1 DNA Way, South San Francisco, CA 94080, USA
<sup>2</sup>Protein Engineering Genentech Inc., 1 DNA Way South San Francisco, CA 94080 USA

Camelids produce functional antibodies devoid of light chains, and as a result, the heavy chain variable (V<sub>H</sub>) domain folds autonomously<sup>2,3</sup> and functions independently as an antigen-binding unit.<sup>4</sup> These autonomous V<sub>H</sub> (V<sub>H</sub> H) domains can be considered “monobodies”<sup>5</sup> because they consist of a single folding unit that is the smallest natural antibody fragment retaining antigen-binding activity. The absence of the light chain has shifted the entire binding epitope of V<sub>H</sub>H domains to the heavy chain and, in so doing, has imposed a greater reliance on the third complementarity-determining region (CDR3). The binding surface of V<sub>H</sub>H domains involve only three CDRs as compared to the six in classical antigen-binding fragments (Fab’s) or single chain variable fragments (scFv’s) and yet, camelid antibodies are capable of attaining binding affinities comparable to those of conventional antibodies.<sup>4</sup>

Their small size and simplicity make monobodies attractive scaffolds for use as starting points for peptidomimetic and small molecule design, as reagents for cell biology and high-throughput protein analysis, or as potential therapeutic agents.<sup>6</sup> The average CDR3 length of camelid antibodies is significantly greater than that of classical antibodies,<sup>7,8</sup> and this fact has led to the proposition...
that V_{HH} domains may be particularly successful in targeting enzyme active sites, and thus, they may be useful for the design of enzyme inhibitors.\textsuperscript{9} V_{HH} domains can also be used as modular binding units in the design of bi-specific antibodies\textsuperscript{10} or as intracellular antibodies.\textsuperscript{11} In addition, because they can be readily produced in high yields in \textit{Escherichia coli}, monobodies may be particularly suited for protein arrays and other proteomics applications which require rapid and cost-effective production of binding reagents. Finally, appropriately designed V_{HH} domains could be used as scaffolds for the presentation of constrained peptide libraries.\textsuperscript{12,13}

Over the past decade, phage display has emerged as a powerful technology, which has had major impacts on the field of antibody engineering.\textsuperscript{14–17} Phage display permits the presentation of antibody fragments on the surfaces of phage particles that also contain the encoding DNA. In this way, a physical linkage is established between phenotype (the displayed protein) and genotype (the encapsulated DNA). In early experiments, natural antibody repertoires were cloned and used to produce phage-displayed libraries of Fabs or scFvs which proved quite successful in the generation of affinity reagents.\textsuperscript{18–21} The utility of antibody-phage libraries was further advanced by the development of libraries with synthetic CDRs displayed on defined scaffolds.\textsuperscript{22–28} These systems offer great potential for high-precision antibody engineering because preferred frameworks can be used and diversity can be tailored and introduced in a site-specific manner. Furthermore, V_{HH} domains may be particularly suited for use as scaffolds for synthetic library design, because only three CDRs are required for high-affinity binding and, due to their small size and simple folding mechanisms, monobodies are likely to be efficiently displayed on phage.

The success of a naïve, synthetic antibody library is dependent upon the choice of both an appropriate scaffold and the regions within the scaffold that are mutated to generate diversity. While the Kabat database\textsuperscript{29} can be used to delineate the approximate boundaries of the CDRs within an antibody scaffold, some residues within these regions may also contribute to the structural stability of the overall fold. The inappropriate inclusion of structural residues as sites of diversification reduces the functional size of the library by producing misfolded proteins that are not only incapable of generating specific binders, but may also contribute to non-specific background binding due to aggregation. The delineation of structural residues within CDRs is especially critical in the design of libraries based on V_{HH} domains where certain CDR3 residues may be required for stabilization of the former light chain interface.\textsuperscript{30,31} Thus, while the V_{HH} domain scaffold offers significant advantages for synthetic library design, the optimal design of a monobody library mandates the precise delineation of CDR3 residues involved in structural stabilization, so that these positions can be fixed and the remaining CDR3 positions can be mutated to generate a library that is both highly diverse and well folded.

There are several hallmark changes in the frameworks encoded by the germline V genes of V_{HH} domains, as compared to their classical VH counterparts.\textsuperscript{3} Most of these changes are at the former light chain interface, and presumably, are adaptations to the loss of the light chain. Most notable is position 37, which is exclusively occupied by a Phe or Tyr in V_{HH} domains, while in classical VH3 domains this position is usually occupied by the smaller aliphatic residues, Val or Leu. The use of a bulky aromatic residue at the center of the former light chain interface is initially surprising, since it adds hydrophobic character to a region that, upon loss of the light chain, one might expect to become more hydrophilic. However, crystal structures reveal that residues from CDR3 pack against Phe/Tyr37 to form a small hydrophobic core.\textsuperscript{30,31} Most V_{HH} domains also substitute an Arg at position 45 in place of Leu or Val, which typically occur in the classical VH3 family; the long aliphatic side-chain of Arg45 packs against Phe/Tyr37 and contributes to the hydrophobic core while the guanidinium group is positioned facing solvent. Two additional differences between V_{HH} and VH domains are frequently observed; V_{HH} domains substitute Ser for Trp at position 47 and Ala for Lys/Arg at position 94, but the consequences of these changes are unclear.

Cys is also commonly observed in V_{HH} framework and CDR1 sequences (predominantly at positions 30, 32, 33 and 45), and in each of these occurrences, the Cys side-chain appears to be poised to make a disulfide bond with an appropriately positioned Cys within CDR3. Such an intrachain disulfide bond has been directly observed in the structure of an anti-lysozyme V_{HH} domain\textsuperscript{30} (Figure 1). The presence of an additional intrachain disulfide bridge may have two consequences; it probably anchors CDR3 against the former interface and also predisposes the orientation of CDR3 for appropriate presentation to antigen.

Previous analyses of cameld monobodies\textsuperscript{30,31} and soluble heavy chain domains\textsuperscript{32} revealed that loss of the light chain binding partner is compensated by interactions between CDR3 and the former light chain interface. In all of the structures reported so far, residues from CDR3 pack against Phe37 to form a small hydrophobic core, which presumably stabilizes the erstwhile interface and prevents aggregation. The types of interactions employed in the different V_{HH} structures vary considerably, and the spectrum of structural solutions is most clearly demonstrated by comparing the structure of a llama anti-human chorionic gonadotropin V_{HH} domain (α-HCG)\textsuperscript{31} with that of the camel anti-lysozyme V_{HH} domain (α-lysozyme)\textsuperscript{30} (Figure 1).

In the α-HCG structure, Trp100 from CDR3 packs against Phe37, and a small hydrophobic
core is formed between Phe37, the carbon atoms of the Trp100 side-chain and the aliphatic portion of the Arg45 side-chain. In this configuration, the guanidinium group of Arg45 and the nitrogen of the Trp100 ring are exposed to the solvent. Thus, the erstwhile interface is characterized by a hydrophobic cluster around Phe37, which is ringed by hydrophilic amide moieties. The interface solution in the \( \alpha \)-lysozyme structure is more modular; the C-terminal end of CDR3 forms a small \( \alpha \)-helix that caps the former interface and is held in place by an additional disulfide bond between Cys109 in CDR3 and Cys33 in CDR1. Despite the different means, however, the end effect is the same, in that the hydrophobic framework residues are sequestered from solvent.

Recently, a phage-displayed library yielded soluble llama V\( _{H} \) domains that did not contain the hallmark camelid residues Phe37 and Arg45, but rather, had sequences that are typical of classical V\( _{H} \) domains. Thus, while natural V\( _{H} \) domains typically use Phe/Tyr37 for repacking of the former light chain interface, alternative solutions are possible. However, the NMR structure of one of these soluble V\( _{H} \) domains revealed a clustering of the CDR3 residues Leu100 and Trp103 with the framework residues Val37 and Leu45. This implies that the types of residues at the former light chain interface may vary, but stabilization of the V\( _{H} \) domain is nonetheless reliant upon the packing of CDR3 residues against framework positions central to the former light chain interface. The role of CDR3 in the stabilization of soluble V\( _{H} \) domains was also highlighted by a biophysical study of V\( _{H} \) germline family members in which representative members from each family were expressed in a soluble form, and solubility was dependent upon the insertion of the same 17-residue CDR3 into each of the V\( _{H} \) domains. While it is not clear how this particular CDR3 contributes to heavy chain stability, the results imply that CDR3 is a fundamental determinant mediating heavy chain solubility.

Taken together, these studies clearly demonstrate an additional structural role for CDR3 in the stability and solubility of V\( _{H} \) domains. This imposes constraints on CDR3 diversity, and these constraints are likely to be somewhat particular to each scaffold, as they depend on which CDR3 residues have been recruited for stabilization of the former light chain interface.

We wanted to engineer V\( _{H} \) scaffolds capable of displaying long CDR3 loops that made only minor contributions to the overall stability of the scaffold. Ideally, any structural interactions with the framework would be restricted to the borders of CDR3, thus enabling the presentation of long, continuous loops with sequences unconstrained by global structural requirements. Herein, we have used rapid phage display methods to develop highly stable scaffolds based on the \( \alpha \)-HCG template. While the \( \alpha \)-HCG CDR3 contains only seven residues, the phage-derived scaffolds display 17 residue CDR3 sequences, and furthermore, the structural stabilities of the scaffolds are dependent on only a few CDR3 residues located at either end of the loop. The phage-derived scaffolds should prove useful in the design of synthetic antibody libraries, and the results provide important insights into the relationships between CDR3 sequence and structural stability within the V\( _{H} \) domain fold. In addition, the general methods and
strategies should prove useful for the rapid analysis of antibody structure and function.

Results

Selection of synthetic CDR3s that stabilize the V_{\text{H}} domain fold

\(\alpha\)-HCG is a VH3 family member and, as such, binds to protein A. More importantly, association with protein A is mediated by contacts distal to the light chain interface and does not involve the CDRs.\(^{40-42}\) Thus, the interaction can be used to select for correctly folded V_{\text{H}} domains\(^{29}\) or, as we describe herein, to monitor the contributions of residues in the former light chain interface and CDRs to the overall structural stability and expression of the V_{\text{H}} domain. We exploited the protein A-V_{\text{H}} domain interaction to investigate the structural role of CDR3 in a synthetic library displayed on the \(\alpha\)-HCG scaffold. This allowed us to delineate the structural boundaries within CDR3 and also enabled the design of scaffolds with desirable properties for the phage display of naïve antibody libraries.

\(\alpha\)-HCG was chosen as the parent scaffold because it had already been expressed in bacteria and purified using a protein A column, and the availability of an X-ray crystal structure\(^{31}\) aided the analysis of our results. A synthetic gene encoding \(\alpha\)-HCG was synthesized with an optimal nucleotide sequence for bacterial expression (Figure 2). This gene provided for efficient display of \(\alpha\)-HCG on M13 bacteriophage using a previously described phagemid\(^{43}\) that directed the secretion of \(\alpha\)-HCG fused to the C-terminal domain of the gene-3 minor coat protein (P3C). In addition, the same vector could be used to purify soluble \(\alpha\)-HCG following insertion of a stop codon between the \(\alpha\)-HCG gene and the P3C gene.

CDR3 of \(\alpha\)-HCG contains only seven amino acid residues, but we wanted to develop and characterize scaffolds capable of presenting long CDRs. For the design of our CDR3 library, we chose a 17 residue loop to reflect the average V_{\text{H}} CDR3 length.\(^{8}\) Positions Gly95 and Trp103 were chosen as the N and C-terminal borders of CDR3, as they seemed the most conservative choices based upon the 152 available cameld monobody sequences.\(^{44}\) The seven residues between these boundaries in \(\alpha\)-HCG were replaced by a library encoding 17-residue loops of random sequence (library NNS17). The randomized positions were encoded by “NNS” degenerate codons that contain 32 unique codons that together encode all 20 natural amino acids, although some amino acids are represented by more than one codon (see Materials and Methods). The 17 residues within the randomized region were numbered according to the Kabat nomenclature.\(^{29}\)

We cycled library NNS17 through four rounds of binding selection to enrich for library members that were capable of binding to protein A, and thus, were likely to be properly folded, highly expressed V_{\text{H}} domains. To obtain a broad view of the selected populations, we sequenced hundreds of selected clones (335 or 324 from rounds 3 or 4, respectively). As we describe below, this broad sampling of the selected clones enabled us to identify general trends in sequence conservation as revealed by an aggregate analysis of many unique clones, and to identify the most fit scaffolds by analysis of the dominant clones within the population.

An aggregate analysis of CDR3 reveals general trends in sequence conservation

Of the 335 clones sequenced after three rounds of binding selection against protein A, 222 were unique sequences that constituted a non-redundant data set which we used to identify general trends in sequence conservation. The 222 sequences were

---

Figure 2. The synthetic gene encoding for \(\alpha\)-HCG. The DNA sequence of the coding strand is shown along with the protein translation below. The CDRs and the C-terminal Flag epitope are underlined. The protein is numbered according to the Kabat nomenclature\(^{29}\) as shown on the left, and insertions are numbered directly above the inserted amino acids (52a, 82a, 82b, and 82c).
aligned, and the occurrence of each amino acid at each of the 17 positions within the loop was tabulated. The totals were then normalized by dividing by the number of times each amino acid was encoded by the redundant NNS codon; for example, the NNS codon contains three unique codons for Arg, and thus, the Arg total at each position was divided by 3 to correct for the bias. The resulting normalized data set (Figure 3a) was then analyzed for significant deviations from a random distribution.

An overall test for independence between amino acid frequency and residue was performed using a chi-squared test for independence (Figure 3b). Pearson residuals were used to identify specific amino acid and residue combinations that were observed significantly more or less frequently than one would expect by chance under the hypothesis of independence. A Pearson residual was considered statistically significant if its magnitude exceeded 3.8. Based on the approximate normality of Pearson residuals, this cut-off corresponds to a p-value < 0.05 even after a Bonferroni adjustment to account for the fact that 340 hypotheses were tested (20 amino acids at each of 17 residues).

A Pearson residual, which is defined as the difference between observed and expected counts normalized by the square-root of the expected count, above 3.8 is strong evidence for selection bias for the particular amino acid at the given position. A Pearson residual less than 3.8 is strong evidence for selection bias against the amino acid at the given position.

Amino acids that deviated most significantly from random (p < 0.05) showed a strong selection bias for incorporation into the loop. The N-terminal end of the loop was biased towards the sequence motif R(L/I/M)XR. Near the central portion of the loop the preference seemed to be for either glycine or hydrophilic residues. The C-terminal end of CDR3 (positions 100g–100l) was characterized by an over-representation of hydrophobes (Phe, Val, Ile and Trp) at particular positions. Both Trp and Gly occurred frequently throughout much of the loop. However, at only a few positions did the occurrence of either of these residues rise significantly above background. In particular, the occurrence of glycine was most significant at positions 100c and 100d near the central portion of the loop. It is possible that Gly at these positions enables some flexibility for turn formation. Trp occurred throughout the loop, but more strongly near the C-terminal end of CDR3.
and may be involved in a number of non-specific interactions. Yet, only at position 100g was the number of tryptophan residues significantly above background.

Identification of optimal V_{HH} scaffolds

The aggregate analysis provided information about general trends in CDR3 sequences that stabilize V_{HH} domains, but we also wanted to identify the most stable scaffolds for detailed analysis. For this purpose, we analyzed the sequence data from the fourth round of selection, reasoning that further selective pressure should favor particularly stable clones and cause the population to converge around these solutions. Indeed, the sequence diversity was significantly reduced by round 4, as approximately 50% of the population was represented by the ten most abundant sequences, and in fact, the top four sequences alone accounted for more than one-third of the population (Figure 4(a)).

While each of the top ten sequences had elements of the preferred amino acid distribution observed in the aggregate analysis (Figure 3), there were significant differences, both amongst the top clones and in comparison with the aggregate consensus. This is not surprising, since the aggregate consensus represents the average characteristics of several hundred clones, while each of the top clones represents a particular solution to the stabilization of the V_{HH} domain fold. It is notable that, while Trp was highly abundant throughout the loop in the aggregate analysis, the

<table>
<thead>
<tr>
<th>Percent Abundance</th>
<th>CDR3 Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Round 3</td>
<td>Round 4</td>
</tr>
<tr>
<td>3.9</td>
<td>10.5</td>
</tr>
<tr>
<td>3.3</td>
<td>9.8</td>
</tr>
<tr>
<td>2.1</td>
<td>8.3</td>
</tr>
<tr>
<td>2.1</td>
<td>6</td>
</tr>
<tr>
<td>1.2</td>
<td>4.1</td>
</tr>
<tr>
<td>1.2</td>
<td>3.8</td>
</tr>
<tr>
<td>0.6</td>
<td>2.2</td>
</tr>
<tr>
<td>1.2</td>
<td>1.9</td>
</tr>
<tr>
<td>0.6</td>
<td>1.3</td>
</tr>
<tr>
<td>0.3</td>
<td>1.3</td>
</tr>
</tbody>
</table>

Figure 4. The ten most abundant CDR3 sequences in library NNS17 following four rounds of selection for binding to protein A. The top ten sequences (a) are shown in rank order along with the percentage of the total population (Percent Abundance) that each represented following three or four rounds of selection. Sequences that match the aggregate consensus (Figure 3) are in underlined bold text. The results of the shotgun alanine-scanning analysis (b) are shown for the top four scaffolds. The wt/Ala ratios for each residue in CDR3 are shown for the scaffolds RIG (white bars), LLR (cross-hatched bars), VLK (gray bars), and RLV (black bars). The name of each scaffold is derived from the sequence at positions 96, 97 and 98. The numbering follows the Kabat nomenclature.29
top four clones are not rich in Trp. Instead, Trp is prevalent in some of the sequences in the lower half of the top ten list (Figure 4a), and even more so amongst the less dominant clones at round 4 (data not shown). We speculate that the large hydrophobic Trp side-chain may shield the former light chain interface through contacts that are fairly non-specific. As a result, there may be numerous low-energy solutions to V_{H}H domain stabilization that utilize Trp side-chains throughout CDR3. However, highly stable solutions may require precise interactions that can only arise from the concerted effects of several specific side-chains at specific sites within the loop. Thus, the numerous low-energy, non-specific solutions appear to be prevalent under conditions of low selective pressure, but the rare, highly stable solutions are able to accumulate and eventually dominate as a result of further selection for binding to protein A.

**Shotgun alanine-scanning of selected scaffolds**

We reasoned that the most dominant sequences after four rounds of selection for binding to protein A were likely to represent particularly stable V_{H}H domain scaffolds. Importantly, the top four sequences exhibited significant identity with the round 3 aggregate consensus mainly at positions on the boundaries of CDR3. Thus, we were hopeful that the stability of these scaffolds was dependent upon key side-chains at either end of the CDR3 loop but would be independent of sequences in the central region, thus making them ideal for the presentation of CDR3 libraries. To map the structural requirements for the CDR3 sequences of the top four V_{H}H domain scaffolds, we subjected each to alanine-scanning mutagenesis. For all amino acids except Gly, Ala substitutions result in truncated side-chains, and thus, the effect of Ala mutations can be used to infer the contributions of individual side-chains to protein structure and function.

To expedite the analysis, we used an extremely rapid combinatorial mutagenesis strategy, termed "shotgun alanine-scanning", which had previously been used to study several other proteins. For each scaffold, we constructed a library using special degenerate codons designed to allow each CDR3 residue to vary as the wild-type (wt) or Ala with equal probability. Binding to protein A was again used as an indicator of stability and expression. Following two rounds of protein A binding selection, approximately 100 clones were sequenced and the wt/Ala ratio was determined for each CDR3 position.

We assumed that residues integral to the stability of the V_{H}H domain fold would not tolerate Ala substitutions, and the wt sequence would be strongly conserved amongst selected clones. Thus, wt/Ala ratios greater than 1 should represent side-chains that contribute to protein stability while side-chains with wt/Ala ratios less than 1 may actually destabilize the protein in comparison with the Ala side-chain. Each of the scaffolds showed a distinct pattern of conserved wt residues and, in each case, sequence conservation was restricted to residues near the boundaries of CDR3 (Figure 4b). In general, the patterns of wt sequence conservation were in good agreement with the consensus obtained from the aggregate analysis (Figure 3). Leu and Ile were the aggregate consensus at position 97, and this side-chain was highly conserved in comparison with Ala amongst all four scaffolds (Figure 4). Scaffolds VLK, LLR, and RLV (scaffolds are named after the sequence at positions 96, 97, and 98) all contain either Trp or Phe at position 100g, which appears intolerant to Ala substitution. In contrast, the RIG scaffold contains a Glu at this position and, while Glu does match the aggregate consensus, it does not appear to be important in the shotgun scan. However, the RIG scaffold exhibited high-level conservation of a Trp100i residue in the alanine-scan, and we speculate that this residue may play a structural role that is similar to the role of Trp/Phe100g in the other scaffolds. The RIG and VLK scaffolds also showed conservation of Val100j in the alanine-scan, and this also agreed with the aggregate consensus. The only notable disagreement between the aggregate consensus and the individual alanine-scanning data occurs at position 99, where an Arg occurs in both the aggregate consensus and three of the top four scaffolds, and yet, the Arg99 was not conserved in comparison with Ala in any of the shotgun scans.

**Biophysical characterization of selected V_{H}H domains**

Overall, the shotgun alanine-scanning results for the top four V_{H}H domains indicated that the stability of each protein was dependent upon only a few key side-chains on the boundaries of CDR3. To further ascertain the suitability of these V_{H}H domains as scaffolds for naïve antibody libraries, we expressed each as a soluble protein and measured the thermal stability, aggregation state, and the binding affinity and specificity for protein A (Table 1). All four proteins were monomeric, as

**Table 1. Biophysical analysis of V_{H}H domains**

<table>
<thead>
<tr>
<th>V_{H}H domain*</th>
<th>t_{m} (°C)</th>
<th>Reversible folding</th>
<th>Protein A K_{d} (μM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>α-HCG</td>
<td>57</td>
<td>Yes</td>
<td>1.6</td>
</tr>
<tr>
<td>RIG</td>
<td>62</td>
<td>Yes</td>
<td>0.8</td>
</tr>
<tr>
<td>VLK</td>
<td>56</td>
<td>Yes</td>
<td>0.9</td>
</tr>
<tr>
<td>LLR</td>
<td>59</td>
<td>Yes</td>
<td>3.3</td>
</tr>
<tr>
<td>RLV</td>
<td>62</td>
<td>Yes</td>
<td>1.8</td>
</tr>
<tr>
<td>RIG(R96A)</td>
<td>NDab</td>
<td>No</td>
<td>&gt;10</td>
</tr>
<tr>
<td>RIG(S100A)</td>
<td>60</td>
<td>Yes</td>
<td>0.6</td>
</tr>
<tr>
<td>RIG(W100A)</td>
<td>51</td>
<td>No</td>
<td>&gt;10</td>
</tr>
<tr>
<td>RIG(V100A)</td>
<td>57</td>
<td>Yes</td>
<td>0.8</td>
</tr>
</tbody>
</table>

* Data are shown for the α-HCG parent, the top four scaffolds in Figure 4, and Ala mutants of the RIG scaffold.

**ND indicates that this value could not be determined.**
determined by size, exclusion chromatography and the melting curves were all fully reversible and indicated a two-state folding transition (data not shown). The thermal stabilities of the phage-derived V\textsubscript{HH} domains, as measured by the temperature of unfolding transition (melting temperature, \( T_m \)), were greater than or comparable to that of the \( \alpha \)-HCG parent. The affinity of each V\textsubscript{HH} domain for protein A was measured by surface plasmon resonance (Biacore) and found to be essentially the same as that of \( \alpha \)-HCG; no detectable binding was observed between any of the proteins and two negative control proteins (bovine serum albumin (BSA) and human vascular endothelial growth factor (VEGF)).

To test the accuracy of the shotgun alanine-scanning predictions, we purified several point mutants of the RIG scaffold as soluble proteins. Based on the conservation of wt sequence (Figure 4b), Ala substitutions for Arg96, Trp100i or Val100j were predicted to be destabilizing to the RIG scaffold. Therefore, we purified these point-mutated proteins, and the mutant S100A, which was predicted to have a stability only slightly lower than that of the wt. All of the mutant proteins were well expressed in \( E. coli \) and all were monomeric as determined by size-exclusion chromatography. The measured thermal stabilities were consistent with the shotgun scan data (Table 1). W100iA was predicted to be the most destabilizing mutation, and indeed, this mutation abolished the reversible folding behavior seen in the wt RIG scaffold and reduced the apparent \( T_m \) by 10 deg.C. The mutation R96A also abolished the reversible denaturation profile, and in this case, the apparent \( T_m \) could not be determined. The V100jA mutant retained a reversible denaturation profile, but the \( T_m \) was reduced by 5 deg.C. In contrast, the S100A mutant exhibited a \( T_m \) almost indistinguishable from that of the wt RIG scaffold and also exhibited reversible denaturation behavior.

We also measured the affinity of each of the mutant proteins for protein A (Table 1). The affinities of the mutants S100A and V100jA were essentially the same as that of wt. In contrast, the affinities of the proteins containing either an R96A or W100iA mutation were severely reduced (\( K_d \) > 10 \( \mu \)M) and no saturable binding could be measured. Taken together, these data indicate that the mutations R96A and W100iA are extremely destabilizing for the structural integrity of the RIG scaffold and reduced the apparent \( T_m \) of the wt RIG scaffold and also exhibited reversible denaturation behavior.

Finally, we directly perturbed the protein A binding site in the RIG scaffold to ensure against the unlikely possibility that the selection process had generated CDR3 sequences with affinity for protein A. In classical V\textsubscript{H} domains, the mutation T57E abolishes affinity for protein A.\textsuperscript{53} and Thr57 is conserved in the \( \alpha \)-HCG sequence. The mutation T57E was introduced into the RIG scaffold, and we could not detect any binding interaction between the mutated protein and protein A by Biacore analysis. The CD spectrum of the mutated protein was indistinguishable from that of the wt, indicating that the molecule was well folded (data not shown).

**Discussion**

The antibody variable domain plays a difficult role. It maintains a stable tertiary fold while supporting the extreme sequence diversities that supply the myriad binding specificities of the immune system. Statistical analyses of antibody sequences in combination with structural studies have been quite successful in distinguishing between the hypervariable sites that mediate antigen binding and the conserved frameworks that maintain the global fold.\textsuperscript{54-57} In addition, site-directed mutagenesis methods have been instrumental in quantifying the contributions of individual residues to antibody structure and function.\textsuperscript{48,58-60} In the present study, we sought to develop combinatorial protein engineering methods to rapidly explore the structural consequences of antibody diversity, and to evolve monobody scaffolds optimized for the presentation of synthetic antibody libraries.

We first screened a completely random CDR3 library for stability and analyzed hundreds of members to identify common features (Figure 3). Despite the extreme diversity of the naïve library and the large number of sequences analyzed, we were able to employ statistical filters that identified recurring sequence motifs. While the large hydrophobic amino acid Trp was highly abundant in the data set, this over-representation was not site-specific, but rather, resulted from a general prevalence of Trp throughout the loop. However, there were definite, site-specific over-representations of particular residues at positions near the N and C-terminal boundaries of CDR3. These over-abundant amino acids were almost exclusively hydrophobes or Arg, which is positively charged but also possesses a long, aliphatic side-chain. The trends towards hydrophobic character are consistent with the idea that CDR3 residues might pack against the former light chain interface to form a small, hydrophobic core that stabilizes the \( \text{V}_{\text{HH}} \) domain fold. Furthermore, it appears that Trp side-chains can accomplish this through interactions that are rather non-specific and scattered throughout the loop, while in contrast, other hydrophobic side-chains mediate interactions that are more precise and specific to particular locations near the boundaries of the loop. As we discuss below, further analysis of the fittest library members suggests that the site-specific interactions depend on synergy between several precisely...
positioned side-chains. While stable V_H domains that depend on several site-specific interactions are rarer than those stabilized by non-specific Trp interactions, they also appear to be the most stable library members, as they clearly dominate under more stringent selection conditions (Figure 4).

Comparison of the aggregate consensus with the sequences of the most dominant clones suggested that the most stable V_H domains depend on side-chains located at key sites near the boundaries of CDR3. To verify this hypothesis, we subjected the CDR3 loops of the four most dominant clones to comprehensive alanine-scans to pinpoint the side-chains that contribute to stability. With traditional methods, such an analysis would have required thermal stability measurements for 68 purified, point-mutated proteins; a tedious task that would have taken several months. By using shotgun alanine-scanning mutagenesis, we were able to accomplish the task in approximately two weeks with simple molecular biology methods and statistical analyses. In each scaffold, several residues near the boundaries of CDR3 were highly conserved in comparison with Ala, indicating that these side-chains contributed significantly to stability. The accuracy of these predictions was directly confirmed for the RIG scaffold; three side-chains (Arg96, Trp100i, and Val100j) were predicted to be important for stability by shotgun alanine-scanning, and an Ala substitution at each of these sites significantly reduced the thermal stability of point-mutated proteins (Table 1).

Thus, our use of protein A binding as a surrogate for V_H domain stability measurements was proven valid, at least for semi-quantitative comparisons. The use of an invariant protein-protein interaction as an indicator of protein stability has been exploited previously; determinants of interaction can be testing against proteins of therapeutic interest. Our results support the hypothesis that CDR3 is a stability determinant in V_H domains. While the specific interactions between a given CDR3 and framework are likely to depend on the length and sequence of the CDR, it is quite remarkable that a completely naïve library converged upon four scaffolds that used similar CDR3 side-chains for stabilizing the V_H scaffold but presented very different loop sequences. Thus, we are optimistic that these novel monobody scaffolds will be able to support CDR3 sequence diversities sufficient for the evolution of synthetic antibodies against diverse protein antigens, but of course, the confirmation of our hopes must await the outcome of selection experiments with naïve, synthetic antibody libraries, which we are constructing and will be testing against proteins of therapeutic interest.

Materials and Methods

Oligonucleotides

Equimolar DNA degeneracies are represented in the IUB code (M = A/C, N = A/C/G/T, R = A/G, S = G/C, W = A/T, Y = C/T). Degenerate codons are shown in bold text. The following mutagenic oligonucleotides were used for library constructions:

**Oligonucleotides**

- **NNS17**: GCC GTC TAT ACT TGT GGT GCT GGT
- **RIGg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **VLKg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **LLRg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **RIGg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **VLKg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **LLRg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **RIGg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **VLKg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **LLRg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **RIGg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **VLKg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **LLRg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **RIGg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **VLKg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **LLRg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **RIGg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **VLKg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **LLRg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **RIGg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **VLKg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **LLRg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **RIGg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **VLKg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **LLRg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **RIGg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **VLKg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **LLRg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **RIGg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **VLKg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **LLRg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **RIGg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **VLKg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **LLRg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **RIGg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **VLKg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **LLRg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **RIGg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **VLKg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **LLRg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **RIGg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **VLKg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **LLRg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **RIGg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **VLKg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **LLRg**: GCC GTC TAT ACT TGT GGT GCT GGT
- **RIGg**: GCC GTC TAT ACT TGT GGT GCT GGT
Construction of a phagemid for phage display of α-HCG

A synthetic gene encoding for α-HCG (Figure 2) was synthesized using nested oligonucleotide PCR. To ensure efficient expression, an optimal nucleotide sequence for bacterial expression was generated by back translating the amino acid sequence with a perlscript routine (CODEIT, C.J.B.). For phage display of α-HCG, we modified a phagemid (pS1602) that was previously used for the display of human growth hormone.43 Standard molecular biology techniques were used to replace the fragment of pS1602 encoding for human growth hormone with the synthetic gene encoding for α-HCG. The resulting phagemid (pCB36624) encoded the following fusion with the synthetic gene encoding for human growth hormone:61 the maltose-binding protein signal peptide, followed by a Flag epitope tag, followed by a Gly/Ser-rich linker peptide, followed by α-HCG, followed by a Flag epitope tag, followed by Gly/Ser-rich linker peptide, followed by P3C.

Construction of phage-displayed libraries

Libraries were constructed using previously described methods43 with appropriately designed “stop template” phagemids. For library NNS17, a derivative of pCB36624 that contained TAA stop codons in place of codons 93, 94, 100 and 101 was used as the template for the Kunkel mutagenesis method,45 with mutagenic oligonucleotide NNS17 designed to simultaneously repair the stop codons and introduce 17 NNK degenerate codons between the codons encoding Gly95 and Trp103. Library NNS17 contained 5 x 10^10 unique members.

For shotgun alanine-scanning, libraries were constructed with appropriate “stop template” versions of each parent phagemid. In each case, codons for residues to be scanned were replaced by degenerate shotgun codons46 designed to encode an equimolar distribution of alanine and the wt at each position. Libraries to scan RIG, VLK, LLR, or RLV scaffolds were constructed using oligonucleotide RIGsg, VLKsg, LLKsg, or RLVsg, respectively. Each library contained at least 5 x 10^10 unique members.

Sorting of phage-displayed libraries

NUNC 96-well Maxisorp immunoplates were coated overnight at 4 °C with protein A (5 μg/ml) (Sigma) and blocked for one hour with BSA (Sigma). Phage from the libraries described above were propagated in E. coli XL1-blue (Stratagene) with the addition of M13-K07 helper phage (New England Biolabs). After overnight growth at 37 °C, phage were concentrated by precipitation with PEG/NaCl and resuspended in phosphate-buffered saline (PBS), 0.5% BSA, 0.1% Tween 20 (Sigma), as described.46 Phage solutions (~10^12 phage/ml) were added to the coated immunoplates. Following a two hour incubation to allow for phage binding, the plates were washed 12 times with PBS, 0.05% Tween20. Bound phage were eluted with 0.1 M HCl for ten minutes and the eluant was neutralized with 1.0 M Tris base. Eluted phage were amplified in E. coli XL1-blue and used for further rounds of selection.

DNA sequencing and analysis

Individual clones from each round of selection were grown overnight at 37 °C, in a 96-well format, in 500 μl of 2YT broth supplemented with carbenicillin and M13-K07 helper phage. Culture supernatants containing phage particles were used as templates for PCRs that amplified the DNA fragment encoding the VλH domain. The PCR primers were designed to add M13-(21) and M13R universal sequencing primers at either end of the amplified fragment, thus facilitating the use of these primers in sequencing reactions. Amplified DNA fragments were sequenced using Big-Dye terminator sequencing reactions, which were analyzed on an ABI Prism 3700 96-capillary DNA analyzer (PE Biosystems, Foster City, CA). All reactions were performed in a 96-well format.

The sequences were analyzed with the program SGCOUNT as described.49 SGCOUNT aligned each DNA sequence against the wild-type DNA sequence by using a Needleman–Wunsch pairwise alignment algorithm, translated each aligned sequence of acceptable quality, and tabulated the occurrence of each natural amino acid at each position. Additionally, SGCOUNT reported the presence of any sequences containing identical amino acids at all mutated positions.

Analysis of library NNS17

To determine if there was any bias in the distribution of amino acids in library NNS17 following selection for binding to protein A, we calculated the Pearson residuals for the entire data set. The Pearson residual is defined as:

\[
\begin{equation}
\epsilon_{ij} = (n_{ij} - \mu_{ij})/\mu_{ij}^{1/2}
\end{equation}
\]

where \(n_{ij}\) is the measured number of occurrences at residue \(j\) of amino acid \(i\), and \(\mu_{ij}\) is the expected distribution of amino acid \(i\) at position \(j\) and is defined as:

\[
\mu_{ij} = \frac{N(A_\lambda/N)(R_{\lambda j}/N)}
\]

\((A_\lambda\) is the total number of amino acids of type \(i\), \(R_{\lambda j}\) is the total number of codons measured at position \(j\), and \(N\) is the total number of codons measured in the entire data set. \(P\)-values were calculated using the method of Bonferroni.46 The standard value of \(p < 0.05\) was adopted as a cutoff for statistical significance.

Protein purification

To permit the expression of soluble protein, phage display vectors were modified by the insertion of a stop codon following the sequence encoding the C-terminal Flag epitope (Figure 2). Individual VλH domain proteins were secreted in E. coli BL21 cells (Stratagene) by IPTG (0.3 mM) induction at 30 °C for three hours. Frozen cell pellets were resolubilized in periplasmic extraction buffer (20% (w/v) sucrose, 50 mM Tris (pH 7.5), 1 mM EDTA). VλH domains were purified from the periplasmic supernatant by protein A affinity chromatography. The eluant from the protein A column was buffer-exchanged into PBS and concentrated using a Centricon YM-3 (Amicon). Concentrated VλH domains were then further purified by gel-filtration using a Superdex 75 gel column (Amersham Pharmacia). All proteins ran as monomers with an apparent molecular mass of 14 kDa.

Thermal stability measurements

Thermal stabilities of the purified VλH domains were measured using an Aviv CD spectrometer model 202 (Protein Solutions, Lakewood, NJ). The signal at 207 nm
was used to monitor unfolding. A 0.5 deg.C temperature step was used during thermal melts and the temperature range was 30–80°C. Melting temperatures were determined for both the unfolding and folding transitions, and all thermal melts were performed using 1 mM protein solutions in PBS.

**Affinity measurements**

The affinity of the interaction between protein A and each V₃₉ domain was determined by surface plasmon resonance, using a Biacore 3000. Protein A was coated onto CM5 chips, and BSA or VEGF were coated separately as negative controls. Binding isotherms were calculated using the steady-state response differentials for a series of V₃₉ domain concentrations. The equilibrium constant (Kₑ) was determined by fitting the resulting curves (KaleidaGraph, Synergy Software) to the following equation:

\[
R_a = R_{\text{max}} + (R_{\text{max}} - R_{\text{min}})/(1 + C/K_e)
\]

where \(R_a\) is the measured response differential, \(R_{\text{max}}\) is the maximal response differential, \(R_{\text{min}}\) is the minimal response differential and C is the ligand concentration.

**Additional material**

Additional sequence data is available from S.S. at sidhu@gene.com by request.

**Acknowledgements**

We thank Bridget Currell for DNA sequencing, the Genentech DNA synthesis group for oligonucleotides, and Michael Ostland for help with statistical analysis.

**References**


Edited by I. Wilson

(Received 3 June 2003; received in revised form 17 July 2003; accepted 18 July 2003)