Biochemistry

Article

pubs.acs.org/biochemistry

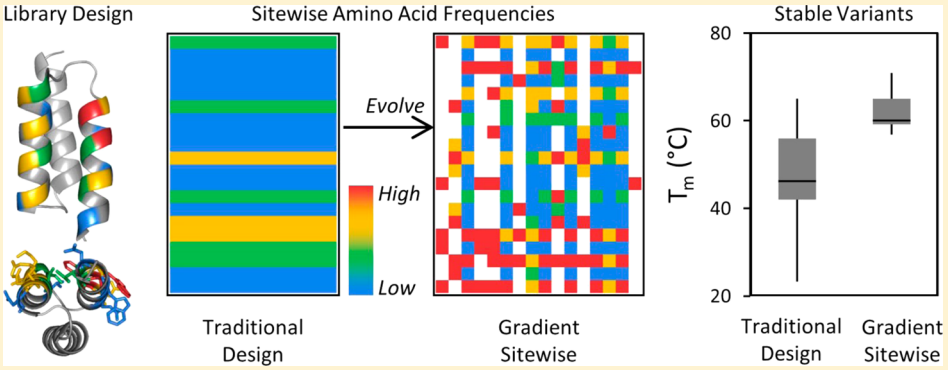
A Gradient of Sitewise Diversity Promotes Evolutionary Fitness for Binder Discovery in a Three-Helix Bundle Protein Scaffold

Daniel R. Woldring,[†] Patrick V. Holec,[†] Lawrence A. Stern,[†] Yang Du,[‡] and Benjamin J. Hackel*[,†]

[†]Department of Chemical Engineering and Materials Science, University of Minnesota—Twin Cities, 421 Washington Avenue Southeast, Minneapolis, Minnesota 55455, United States

[‡]Molecular and Cellular Physiology, Stanford University, 279 Campus Drive, Stanford, California 94305, United States

Supporting Information



ABSTRACT: Engineered proteins provide clinically and industrially impactful molecules and utility within fundamental research, yet inefficiencies in discovering lead variants with new desired functionality, while maintaining stability, hinder progress. Improved function, which can result from a few strategic mutations, is fundamentally separate from discovering novel function, which often requires large leaps in sequence space. While a highly diverse combinatorial library covering immense sequence space would empower protein discovery, the ability to sample only a minor subset of sequence space and the typical destabilization of random mutations preclude this strategy. A balance must be reached. At library scale, compounding several destabilizing mutations renders many variants unable to properly fold and devoid of function. Broadly searching sequence space while reducing the level of destabilization may enhance evolution. We exemplify this balance with affibody, a three-helix bundle protein scaffold. Using natural ligand data sets, stability and structural computations, and deep sequencing of thousands of binding variants, a protein library was designed on a sitewise basis with a gradient of mutational levels across 29% of the protein. In direct competition of biased and uniform libraries, both with 1×10^9 variants, for discovery of 6×10^4 ligands (5×10^3 clusters) toward seven targets, biased amino acid frequency increased ligand discovery 13 ± 3-fold. Evolutionarily favorable amino acids, both globally and site-specifically, are further elucidated. The sitewise amino acid bias aids evolutionary discovery by reducing the level of mutant destabilization as evidenced by a midpoint of denaturation (62 ± 4 °C) 15 °C higher than that of unbiased mutants (47 ± 11 °C; $p < 0.001$). Sitewise diversification, identified by high-throughput evolution and rational library design, improves discovery efficiency.

Molecular recognition ligands are valuable tools in fundamental biology, medicine, and industrial biotechnology. Engineered ligands allow control over the binding epitope, the affinity, the selectivity, and the biophysical properties of the ligand. Protein ligands are frequently engineered by modulating amino acids in a select region, known as the paratope, of a protein while conserving a stable underlying framework.[1] A variety of protein topologies have demonstrated efficacy as scaffolds for the evolution of novel binding function, including natural immune repertoires of antibodies[2] and variable lymphocyte receptors[3] as well as a multitude of synthetically diversified scaffolds.[1,4] One particular example, the affibody domain, has been effectively used as a ligand scaffold in processes such as evolution of binding to numerous targets, with affinities as strong as 20 pM, and application to diagnostics, molecular imaging, and therapy.[5,6] The affibody is a 58-residue, three-helix bundle derived from the Z domain of staphylococcal protein A. It is readily expressed recombinantly in bacteria and highly soluble and reversibly unfolds with a wild-type midpoint of 72 °C,[7] although engineered mutants have exhibited destabilization to denaturation midpoints of 37−65 °C (mean of 49 °C).[7−12] Mutants with novel binding activity have been discovered and evolved from combinatorial libraries with diversity at 13

Received: November 9, 2016
Revised: February 3, 2017
Published: March 1, 2017

DOI: 10.1021/acs.biochem.6b01142
Biochemistry 2017, 56, 1656−1671

residues on one face of the N-terminal and middle helices. Each of the 13 sites was diversified to the 20 natural amino acids using broadly distributed NNK codons. Library screening has been performed using phage display,[5,13,14] ribosome display,[15,16] and bacterial display;[12,17−19] yeast display was used for framework evolution.[20]

Evolution of novel binding function necessitates mutation of a sufficient paratope area to drive the new intermolecular interaction[21] while maintaining sufficient intramolecular stability. Mutation of intramolecularly critical sites or mutation of semitolerant sites to suboptimal amino acids can limit the evolutionary potential despite the introduction of an otherwise effective paratope.[22−27] Thus, identification of the mutational tolerance of each site, within the context of a diverse array of sequences possible within a combinatorial library, can aid evolution. Implementation of variable diversities at each site, in both entropy and specific amino acid preferences, has proven to enhance the evolutionary efficacy in synthetic fibronectin domain libraries[28,29] and natural[30] and synthetic[31] antibody repertoires. A sitewise constraint has also been implemented in designed ankyrin repeat proteins[32−34] and fibronectin domain sheet libraries[35,36] using rational bias. A sitewise constraint has not yet been published for the affibody domain. Amino acid bias, across sites, has been implemented using amino acids frequently observed in protein−protein interfaces, particularly tyrosine and serine,[28,37−40] as well as glycine in loop paratopes. Sitewise bias has been identified via natural antibody repertoire mimicry,[41] wild-type constraint,[28] structural analysis,[33] and high-throughput evolution and deep sequencing.[29] Identification of detrimental mutations via deep scanning strategies has also been studied.[42−46] Computational prediction of mutational impacts on stability[47−50] and functional maturation,[51,52] though not functional discovery, have been extensively studied and could be used to guide library design.

The study presented here aimed to identify the sitewise amino acid diversities consistent with efficient evolution of a broad array of binding functions (i.e., creation of a single library containing specific binders to a multitude of targets) and examine the drivers and implications of these amino acid preferences. We provide a platform for designing small protein libraries in terms of deciding which residues to include or avoid at individual positions. The effectiveness of this approach is demonstrated in the context of discovering high-affinity variants from an affibody library.

MATERIALS AND METHODS

Preliminary Library Design (first generation). As a preliminary attempt to identify the most beneficial diversification strategy for the three-helix affibody scaffold, a combinatorial library that incorporated a wide variety of mutations across select sites was designed. While, traditionally, the affibody scaffold is uniformly mutated at 13 positions using all 20 natural amino acids, it is largely unknown which amino acids are most effective at any particular site. To better understand these functional diversities, 15 solvent-exposed sites throughout helices 1 and 2 (classic 13 and sites E15 and I31) were mutated using five separate levels of diversity: (i) wild-type (WT) residue, (ii) WT or serine (small size and promotes neutral interaction[53]), (iii) WT, serine, or tyrosine (frequently drives binding affinity and specificity[39,54]), (iv) relaxed, moderate diversity (A, C, D, G, N, S, T, or Y), or (v) full diversity mimicking the chemical composition of the third antibody heavy chain complementarity-determining region (CDR-H3). These five combinatorial libraries were separately constructed on the DNA level and then pooled to form the first-generation library. Any single variant within the initial library exhibited sitewise mutations from only one of the five combinatorial sublibraries.

Gradient Sitewise (GS) Library Design (second generation). The GS library was designed in a sitewise manner by balancing numerous data inputs. Numerous sites were constrained in their amino acid diversity (details in the Supporting Information). Amino acid diversity at likely hot spot affibody positions was guided by the amino acid prevalence in natural antibody interfaces (Abysis database; CDR-H3 diversity, Kabat sites 95−102) and previously evolved affibodies found in the literature[5−7,9,11−15,18,55−78] as well as those generated in house with the first-generation library. The prevalence of each amino acid, except Gly and Arg, was calculated on the basis of equally weighted CDR-H3 diversity via Abysis and previously evolved binder data, yielding the codon design that we call B* hereafter (Figure S1). The B_{LC}* codon closely mimics this design, while also emphasizing the low cysteine content (0.5% in B_{LC}* vs 2.5% in B*).

Solvent Accessible Surface Area (SASA). Protein Data Bank (PDB) files for affibody (PBD entries 1H0T, 1LP1, 2B88, 2KZI, 2OKT, and 3MZW) were processed using GetArea[79] to calculate the SASA relative to random coil (probe radius, 1.4 Å). At each site, the median value across all six structures was reported. For PDBs that included multiple affibody chains, the mean SASA of the chains was determined prior to calculating median values among all structures.

Computational Stability. At each of the 13 classically diversified sites (9−11, 13, 14, 17, 18, 24, 25, 27, 28, 32, and 35) and for each naturally occurring amino acid, the change in stability ($\Delta\Delta G_{folding}$) upon mutation was calculated with FoldX.[80] First, an affibody structure was randomly mutated in accordance with the first-generation library design to calculate the baseline stability. Next, $\Delta\Delta G_{folding}$ was calculated for each single mutation (all 19 natural amino acids substituted, separately, into each site) versus the parental mutant. This process of saturation scanning was repeated for 312 randomly generated parental mutants across five affibody PDB entries (1H0T, 1LP1, 2B88, 2KZI, and 3MZW). The median $\Delta\Delta G_{folding}$ was then calculated for each residue site combination for the 13 classically diversified sites; 312 iterations were shown to be sufficient for convergence of the (de)stabilization values.

Natural Homologue Analysis. Pfam[81] family B (PF02216) was accessed in April 2013 to retrieve 1484 total sequences, 119 of which were unique. Rather than equally weighting all 1484 sequences, we gave the unique sequences a weight dictated by the square root of the number of occurrences for that sequence. Using this adjusted weighting, we calculated a sitewise amino acid distribution (Table S2).

Relative Helix Propensity. Empirical data from several published studies[82−87] were collectively used to calculate relative propensities of each natural amino acid within helical secondary structures. The values from the previous studies were linearly averaged, based on destabilizing energetics of folding relative to glycine (propensity of 0) and alanine (propensity of 1). Using this normalized scale, the propensity of proline is calculated to be −2.7.

Library Construction. Each combinatorial library was built using synthetic oligonucleotides (IDT DNA) with degenerative codons, which were assembled by overlap extension PCR.

DOI: 10.1021/acs.biochem.6b01142
Biochemistry 2017, 56, 1656−1671

Library genes were transformed into EBY100 *Saccharomyces cerevisiae* yeast via electroporation,[88] wherein the library fragments homologously recombined with the linearized pCT vector to yield a construct that allowed yeast surface display of the encoded proteins.[89] The transformation efficiency was quantified using dilution plating with SD-CAA selective medium. Sanger sequencing of initial libraries was performed for quality control.

Binder Selection. Yeast libraries were grown in SD-CAA selective medium (16.8 g/L sodium citrate dihydrate, 3.9 g/L citric acid, 20.0 g/L dextrose, 6.7 g/L yeast nitrogen base, and 5.0 g/L casamino acids) at 30 °C while being shaken at 250 rpm. Surface display was achieved by switching to SG-CAA medium (10.2 g/L sodium phosphate dibasic heptahydrate, 8.6 g/L sodium phosphate monobasic monohydrate, 19.0 g/L galactose, 1.0 g/L dextrose, 6.7 g/L yeast nitrogen base, and 5.0 g/L casamino acids) upon incubation at 30 °C with shaking at 250 rpm for 16 h. Induced libraries were enriched for affibody variants that specifically bound each of several protein targets [first generation, lysozyme and rabbit immunoglobulin G (IgG); second generation, death receptor 5, transferrin, cytochrome *c*, glucose-6-phosphate dehydrogenase, CD276, MET, and a G-protein-coupled receptor; separate enrichments performed in parallel] using both magnetic streptavidin-coated bead sorting[90] and fluorescence-activated cell sorting.[91] Each round of bead sorting consisted of two incubations (2 h each) of induced yeast with either bare beads or beads preincubated with an arbitrary biotinylated protein for depletion of nonspecific binding interactions. Following the depletions, the remaining yeast cells were incubated with beads with a biotinylated target for 2 h and washed twice with PBSA. Yields for both nonspecific binding and target binding were quantified using serial dilution plating. Populations that demonstrated a strong specificity (>10-fold) of target binding relative to nonspecific binding were isolated for sequence analysis. For all populations, after two rounds of enrichment using target-labeled beads, flow cytometry was conducted. Preparation for cytometry consisted of incubating an induced yeast population with 100 nM biotinylated target and 67 nM anti-c-Myc epitope tag antibody (9E10, Biolegend) in PBSA for 30 min at room temperature. Following the primary labeling step, cells were washed with PBSA, incubated with AlexaFluor647-conjugated goat anti-mouse antibody and AlexaFluor488-conjugated streptavidin for 15 min at 4 °C, and then washed with cold PBSA. If target binding was observed, all binding events above background, denoted as high-stringency binders, were isolated for sequence analysis. Ultra-high-stringency binding populations (Figures S5−S7) were obtained through one additional round of cytometry sorting using 5 nM biotinylated target and a stricter sorting gate. In the event that no binding was detected, all full-length, c-Myc positive events were isolated for evolution. Plasmid DNA was zymoprepped from yeast, subjected to dual error-prone PCR efforts on the full gene and on shuffled helices,[92] and electroporated into yeast for additional iterations of bead sorting and flow cytometry.

Affinity and Specificity Analysis. Representative variants were randomly chosen from the high- and ultra-high-stringency populations (panels A and B of Figure S6, respectively). Each variant was induced in yeast, incubated with anti-c-Myc antibody and at various concentrations of their respective target, washed with cold PBSA, incubated with AlexaFluor647-conjugated goat anti-mouse antibody and AlexaFluor488-conjugated streptavidin for 15 min at 4 °C, and then washed with cold PBSA. The extent of binding was measured via flow cytometry and normalized on the basis of the maximal signal strength associated with each target and the background fluorescence of target-free yeast. Dissociation constants (K_D) were calculated by fitting the data to a two-state binding curve (n = 3). Specificity was assessed by separately incubating each high-stringency variant with multiple nontarget biotinylated proteins at 100 nM or with a target protein at 50 nM. The extent of binding, normalized by the maximal possible fluorescence associated with each target, is then compared between the target and nontarget samples (Figure S8).

High-Throughput Sequence Analysis. The plasmid DNA from the initial libraries as well as the enriched populations, both high-stringency cytometry detectable and magnetic bead selective, was isolated for sequence analysis with a Zymoprep kit (Zymo Research). Illumina MiSeq adapter and indexing sequences were added via PCR. Paired-end (250 bp) analysis yielded 18×10^6 quality reads. Raw paired-end read output was groomed and assembled with PANDAseq using a quality score threshold of 0.99 and then converted to FASTA.[93] FASTA sequence files were processed using ScaffoldSeq.[94] A sequence homology threshold of 80% was used for clustering similar variants. Clusters were then dampened using a factor of 0.25 to account for enriched sequence frequency while gaining a diversity of information from less frequent variants.

Library of Origin Analysis. Sequence variants from the initial and evolved populations were assessed for having originated from each of the three library designs. The probability of finding a particular sequence within each of the three sublibraries (NNK, GS, and GS_{LC}) is calculated as the quantity $P(k)_S$. This equation takes into account the amino acid, i, present at each position, j, of the sequence, k. The frequency of i at position j within the sitewise amino acid design of sublibrary S is given by $f_{\bar{s},i,j}$. The library origin probability, $P(k)_S$, is then obtained by taking the product across each of the 17 potentially diversified sites (eq 1). Amino acids not offered within a particular sublibrary (e.g., Arg at site 11 within the GS design or Tyr at site 6 of the NNK design) were given a default value of 1×10^{-4} rather than zero to account for random errors during library construction, evolution of binders, and DNA sequencing.

$$P(k)_S = \prod_{j=1}^{17} f_{\bar{s},i,j} \tag{1}$$

Each sequence was then binned with the sublibrary that yielded the greatest library origin probability. Within each sublibrary bin, the number of occurrences, n, of each unique sequence, v, was tallied and square rooted to normalize the contribution of dominant variants. The total sequence count, N_S, within each of the three sublibraries is the summation of normalized counts for each unique variant (eq 2). Rare variants, having been observed fewer than 10 times, were excluded from the analysis.

$$N_S = \sum_S \sqrt{n_v} \tag{2}$$

Stability Measurements. Individual variants were randomly selected from cytometry detectable binding populations. Each gene was shuttled into a pET-24 derivative to include a C-terminal six-histidine tag, expressed in BL21(DE3) bacteria, and purified using cobalt resin columns.[29] Purified proteins

DOI: 10.1021/acs.biochem.6b01142
Biochemistry 2017, 56, 1656−1671

were diluted in PBS to a concentration of 1 mg/mL and assessed via circular dichroism using a Jasco J815 instrument. Temperature scans were conducted using a range of 20−90 °C (1 °C/min) while monitoring the 220 nm wavelength. Midpoints of thermal denaturation were calculated assuming a standard two-state unfolding curve. For reducing conditions, dithiothreitol (DTT) was prepared fresh and added to the samples to yield a final concentration of 2 mM DTT and then incubated at room temperature for 30 min.

RESULTS AND DISCUSSION

First-Generation Library. We sought to identify sites, within the typically diversified paratope, that would provide the most substantial benefit from constrained diversity. The main approach for identifying sites, and their amino acid preferences, consisted of high-throughput ligand engineering and deep sequencing analysis. Thus, we designed, constructed, and screened a combinatorial library of affibody domains to discover functional variants and direct their evolution. With an aim of improving the binder frequency to generate a diverse set of functional sequences, modest constraint at select sites was introduced in the initial library. Evaluation of the solvent accessible surface area (Table 1) and spatial orientation at the 13 classically diversified sites revealed three sites that were hypothesized to provide more evolutionary benefit from at least partial wild-type constraint. Q9 is only 19% accessible to solvent (Figure 1, orange). N11 is 51% accessible but predominantly facing away from the evolved binding surface (Figure 1, yellow). R27 is at the core of the paratope but only 27% accessible (Figure 1, blue). The other diversified sites range from 33 to 90% accessible (median of 71%). Two sites that were previously conserved, E15 and I31, were modestly accessible (38 and 27%, respectively) and oriented near the evolved binding surface (Figure 1, green and violet). The mild diversity of these sites was hypothesized to provide an evolutionary benefit.

Table 1. First-Generation Library Design[a]

site[b]	WT	SASA (%)	previous libraries[c]	first-generation library[d]
N6	A	78	N	A
K7	K	79	K	K
E8	E	39	E	E
Q9	Q	19	20	N/NS/NSYT/12/20*
Q10	Q	66	20	20*
N11	N	51	20	N/NS/NSYT/12/20*
A12	A	4	A	A
F13	F	33	20	20*
Y14	Y	72	20	20*
E15	E	38	E	E/EA/DSYA/12/20*
I16	I	0	I	I
L17	L	10	20	20*
H18	H	67	20	20*
E24	E	88	20	20*
E25	E	90	20	20*
Q26	Q	23	Q	Q
R27	R	27	20	R/RS/RSYCHP/12/20*
N28	N	71	20	20*
A29	A	74	A	A
F30	F	10	F	F
I31	I	27	I	I/IS/ISYCFN/12/20*
Q32	Q	72	20	20*
S33	S	44	K	A
L34	L	1	L	L
K35	K	76	20	20*
D36	D	66	D	D

[a]The site, wild-type (WT) amino acid, relative solvent accessible surface area (SASA), and diversity in the published and first-generation libraries are presented for sites in the N-terminal and middle helix of affibody. [b]Sites 1−5 and 37−58 were conserved as wild type (Table S1). [c]The numeral 20 represents a mixture of all 20 amino acids using NNK degenerate codons. [d]The term 20* denotes a mixture of all 20 amino acids weighted on the basis of their frequency in the antibody repertoire but with reduced glycine because of the helical structure (Figure S1). Slashes separate sublibrary designs.

Thus, a combinatorial library was constructed from five sublibraries in which these constrained sites were varied from wild-type conservation to full amino acid diversity (Table 1). The 10 other typical paratope sites were broadly diversified to all 20 amino acids using complementarity-biased amino acid frequencies.[28,37−40] The library, constructed by overlap extension PCR of degenerate oligonucleotides and homologous recombination in a yeast display system, contained 4×10^8 variants. The library generally matched the intended design (median absolute deviation from design, $|f_{observed} - f_{design}| = 0.8\%$), although the content of tyrosine and cysteine was modestly higher than desired while that of alanine and aspartic acid was lower than designed (Figure S2A).

Specific binders to hen egg lysozyme and rabbit IgG were discovered and evolved from the combinatorial library using yeast display with magnetic and flow cytometry selections. Deep sequencing of the evolved variants yielded 6×10^4 unique sequences in 523 diverse families. Numerous amino acids, at specific sites, exhibited substantial enrichment or depletion from the unselected to the evolved populations, which is indicative of the benefit or detriment of that amino acid at that site in functional affibodies (Figure 2). These data inform the effective design of an improved combinatorial library.

Second-Generation Library Design. *Overall.* A second-generation library was designed to further evaluate sitewise preferences in the context of a more focused design. On a sitewise basis, amino acids that (i) appeared frequently in binder sequences from the first-generation library (Figure 2) and 345 published binder sequences (Figure 3A), (ii) are computationally predicted to be stable in the context of diverse paratopes (Figure 3B), and (iii) occur naturally in affibody homologues (Figure 4A) were favored. Broad diversity was favored in sites that are (i) solvent accessible and oriented toward the proposed binding interface (Table 1 and Figure 1), (ii) broadly diverse in natural homologues (Figure 4B), and (iii) computationally predicted to be stable to multiple mutations (Figure 3B). Generally at all sites, amino acids that (i) appear frequently in antibody CDR-H3 (bioinf.org.uk/Abysis; Kabat sites 95−102) (Figure 5A), which has also been implemented in previous synthetic libraries,[28,37−40] and (ii) favor helix formation[82−87] (Figure 5B) were favored. Each undiversified framework position, where mutational diversity is unlikely to provide added quality to the overall library design, was conserved as wild type in the context of the optimized affibody framework.[8] Library design details are provided in Table S3. Commentary on the broadly diversified distribution, cysteine content, and nontraditionally diversified sites follows.

Broadly Diversified Sites. At sites that were predicted to benefit from broad diversity, a biased distribution was implemented based 50% on mimicking antibody CDR-H3 diversity and 50% on mimicking evolved affibody sequences in

DOI: 10.1021/acs.biochem.6b01142
Biochemistry 2017, 56, 1656−1671

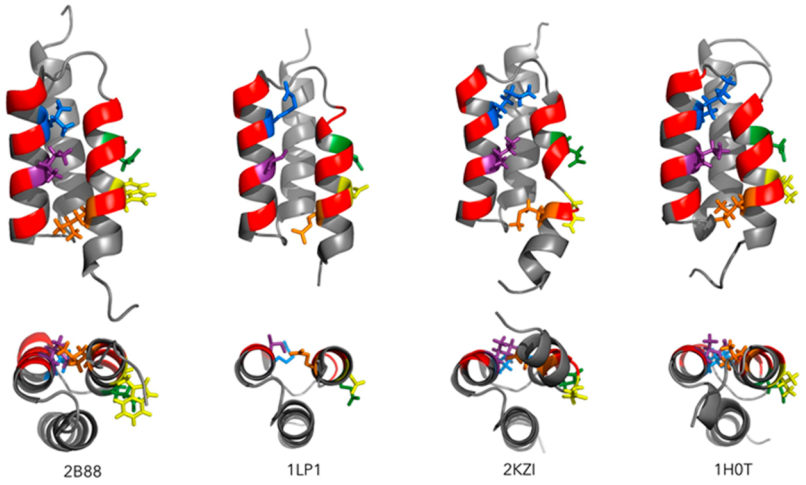


Figure 1. Affibody structure with constrained sites highlighted. Side chains are shown for Q9 (orange), N11 (yellow), E15 (green), R27 (blue), and I31 (violet) in four determined affibody structures.[63,67,75,95] The other 10 classically diversified sites are colored red.

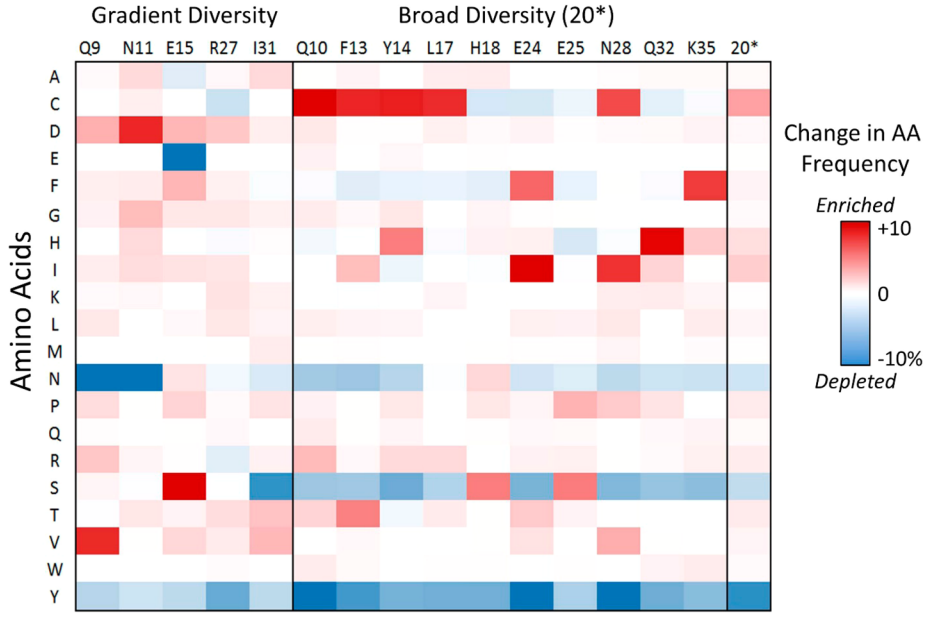


Figure 2. Sitewise amino acid preferences in the context of the first-generation library. The change in amino acid frequency, between the unselected and evolved populations ($f_{evolved} - f_{unselected}$), is shown for each amino acid at each site. The term 20* represents the aggregated frequencies of the 10 sites with broad diversity (20* in Table 1). A total of 6×10^4 unique evolved sequences were used.

the heavily diversified sites (Figure 2, 20* column). There are two exceptions to this balanced design: glycine and arginine. Glycine is frequently observed (12%) in antibody CDR-H3, predominantly for conformational flexibility,[31,53,97] but does not generally support helical structure (Figure 5B) and was not enriched in functional variants from the first-generation library (Figure 2). Thus, glycine content was set on the basis of equally weighting frequency in homologous proteins (2%) and forecasted frequency in enriched binders [7% (Figure S1)], yielding 5% after renormalization. Arginine has been shown to correlate with nonspecific binding when moderately present at binding interfaces.[53,98] Thus, although arginine was modestly enriched in published binders (9% naïve to 12% evolved, though not all studies included counterselections for specificity) and slightly enriched in binders from our first-generation library (2.6–3.5%), the arginine content in the second-generation library was restricted to its frequency in affibody homologues (3%). This broadly diverse distribution, denoted as B* (broad), was designed to facilitate high-affinity, specific binding interactions as well as accommodate stable helices.

Cysteine. The cysteine content in the first-generation library was higher than designed (Figure S2A) and was strongly increased in evolved binders (9–18%; $p < 0.001$) at five broadly diversified sites (10, 13, 14, 17, and 28) and maintained or reduced at all other sites (Figure 6A). Notably, variants with zero or one cysteine were depleted ($p < 0.001$) in evolved binders, whereas variants with at least two cysteines were either maintained or enriched [$p < 0.001$ (Figure 6B)]. These sites of cysteine enrichment are spatially clustered (Figure 6C), and several pairs exhibit substantial epistasis (Figure 6D). The variants having two cysteines predominantly came from the lysozyme binding population with cysteines at sites 10 and 28 across four families. Variants with four cysteines were mostly generated from the rabbit IgG binding population with

DOI: 10.1021/acs.biochem.6b01142
Biochemistry 2017, 56, 1656–1671

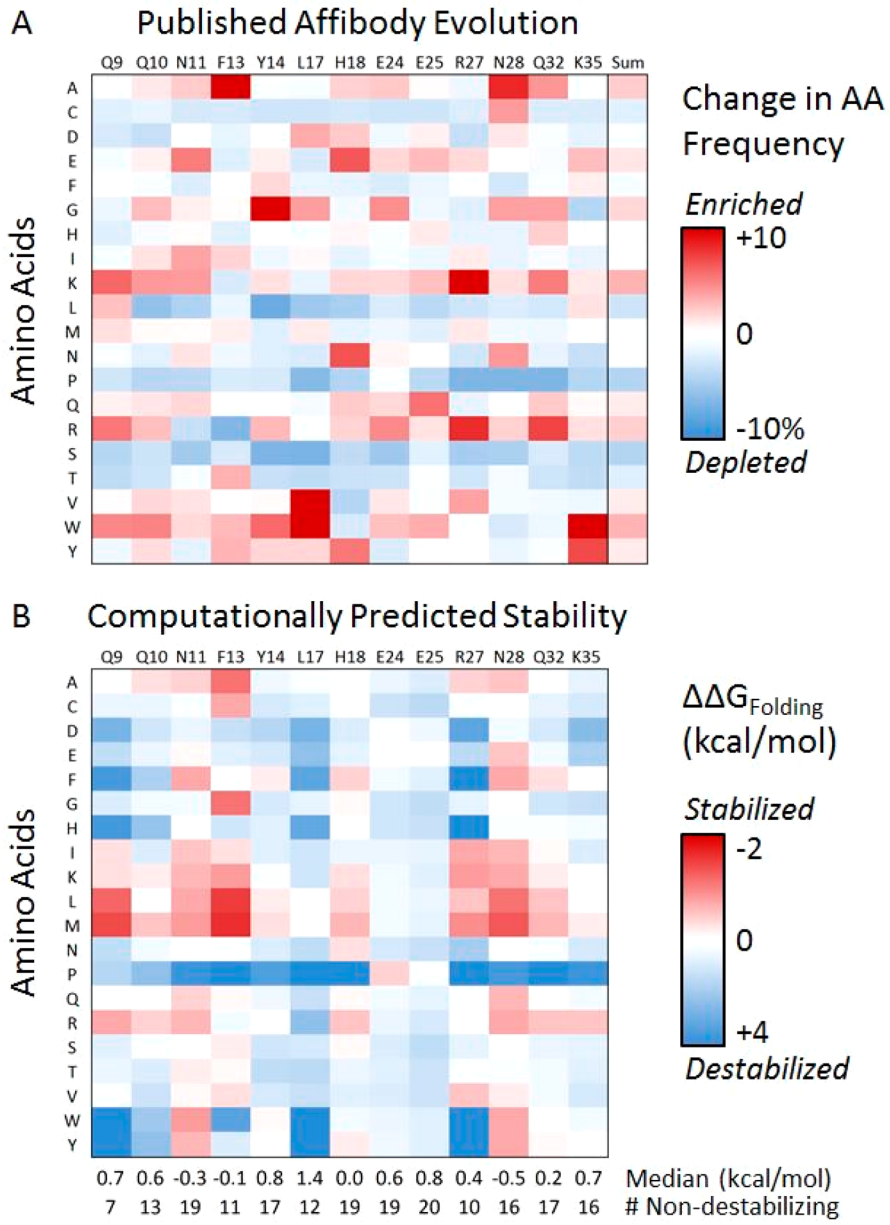


Figure 3. (A) Sitewise amino acid preferences from published affibody evolution. The change in amino acid frequency, between the unselected and evolved populations ($f_{evolved} - f_{unselected}$), is shown for each amino acid at each of the 13 traditionally mutated sites; 345 unique evolved sequences from numerous studies[5−7,9,11−15,18,55−59,61−78,96] were used. (B) Computed destabilization upon mutation in the context of diverse paratopes. The median change in folding free energy ($\Delta\Delta G_f$) upon mutation to the indicated amino acid at the indicated site was computed using FoldX. Saturation scanning (calculation of all mutants) was performed at each site for 312 random library variants with five affibody structures. For each site, the median destabilization and number of tolerated amino acids (mutants with a $\Delta\Delta G_f$ of <1.5 kcal/mol) are also presented.

cysteines at sites 10, 13, 14, and 17 across 57 families (Figure S3A). The importance of potential disulfide formation was evaluated by assessing thermal stability in oxidizing and reducing environments. Four evolved variants, each with unique cysteine locations, exhibited substantially greater thermal stability in the oxidized state (Figure 6E and Figure S3B). Collectively, these observations are consistent with the evolutionary benefit of disulfide bonds.

It is possible that cysteine enrichment resulted from a need for enhanced stabilization in light of the extent of diversification, including newly varied sites 15 and 31. This idea is supported by analysis of the subset of evolved binders with wild-type E15 and I31 in which variants with zero or one cysteine are not depleted in the functional population [$p = 0.16$ (Figure 6B)]. Moreover, in previously published binders from NNK libraries, which conserve E15 and I31, only site 28 exhibited cysteine enrichment (Figure 3A). We aimed to evaluate the ability to evolve affibody ligands with limited cysteines, both to require intramolecular stability in the absence of disulfide bonds and because cysteine-containing variants may hinder production in *Escherichia coli*, increase the likelihood of aggregation or oligomer formation during purification, and complicate the use of thiol chemistries in downstream applications. Thus, in the second-generation library experiments, an additional broad distribution, B_{LC}*, that has a lower cysteine frequency (0.5%) was tested.

Broadened Paratope. Sites 6, 15, 31, and 36 are not traditionally diversified, but natural sequence diversity in

DOI: 10.1021/acs.biochem.6b01142
Biochemistry 2017, 56, 1656−1671

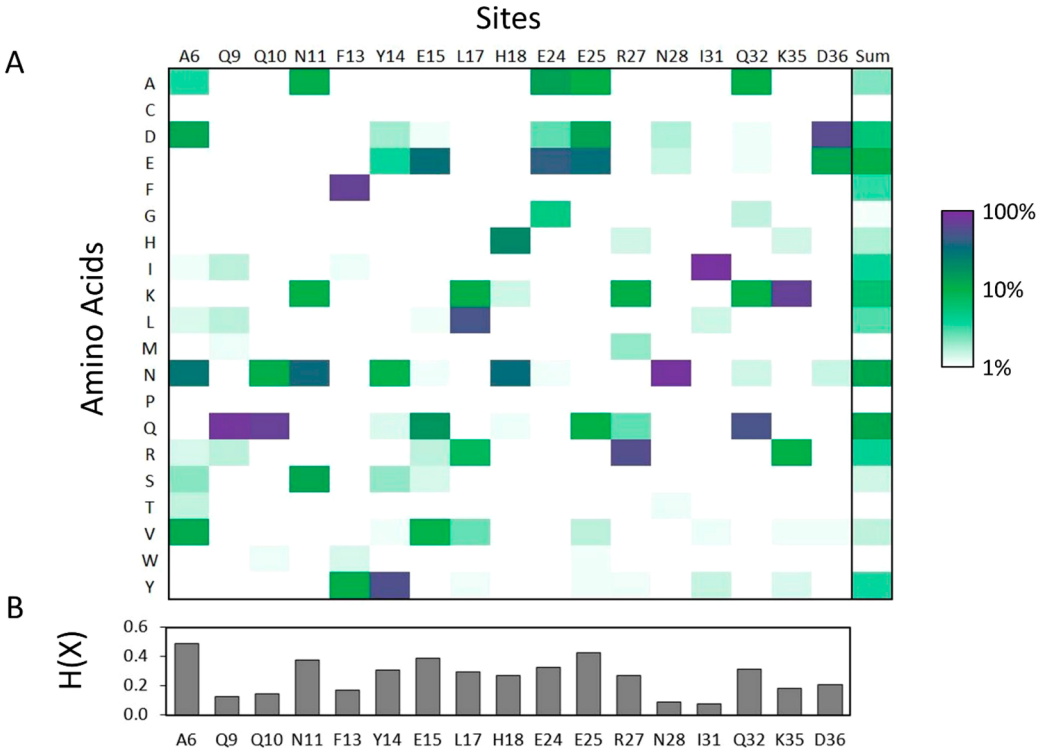


Figure 4. Sitewise amino acid preferences from affibody homologues. (A) Amino acid frequencies of 1484 (119 unique) proteins from Pfam[81] family B, PF02216, which are homologous to the affibody sequence, for each amino acid at each site. (B) Shannon entropy, $H(X) = \sum -f_{aa} \log_{20} f_{aa}$, of each site.

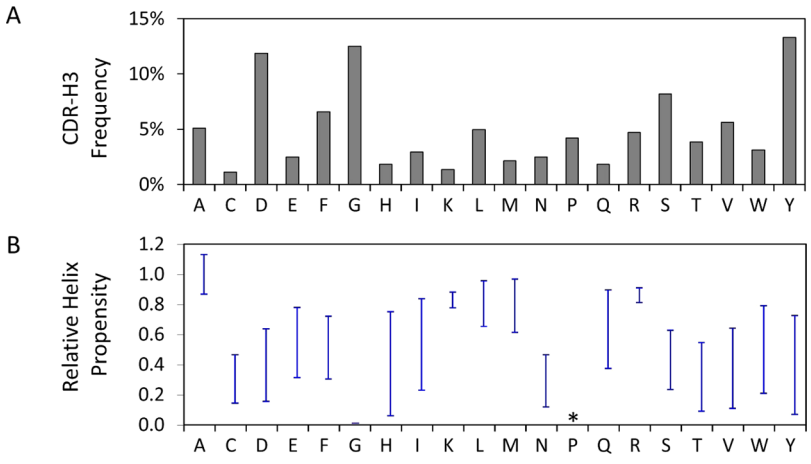


Figure 5. Aggregate amino acid preferences from natural proteins. (A) Amino acid frequencies observed throughout the third complementarity-determining region of antibody heavy chains (CDR-H3; Kabat sites 95−102). (B) Relative helical propensity of each amino acid, based on the observed destabilization of helical secondary structure when used for substitution, calculated as the aggregate of several previous studies.[82−87] The ranges indicate one standard deviation above and below the mean for each residue. The helix propensity of proline (asterisk), calculated to be −2.7, is outside the presented range.

affibody homologues (Figure 4A), computational mutational tolerance (Figure 3B), solvent accessibility (Table 1), and amino acid frequency in evolved affibody sequences (Figure 2) inform mild diversification at these sites. The natural diversity at site 6 is high [Shannon entropy of 0.49 (Figure 4)]. Mutations to E and A in the context of a HER2 binder were not destabilizing.[8] The site is exposed to solvent and could be expected to interact with the target in many cases; in fact, the HER2 ligand mutants impacted affinity. Thus, mild diversity (NSYT) was tested. Site 15 is naturally diverse (Shannon entropy of 0.39) and tolerant of mutation (15 residues) but is pointed away from the paratope and only 38% accessible to solvent. Despite 55% E on natural homologues, biased E was strongly depleted in binding populations (from 38 to 18%) from the initial library. Wild-type Q (30% in homologues) is conserved in published libraries. S is present naturally (1%) and enriched in binders (from 5 to 15%). V (10% naturally and enriched by 3−5% in binders) was also considered. Thus, QSEV diversity was allowed. I31 is conserved in previous libraries and naturally (96% I; Shannon entropy of 0.08) and is predicted to be poorly tolerant of mutation (only eight tolerant residues), yet it is in the center of the planned paratope and

DOI: 10.1021/acs.biochem.6b01142
Biochemistry 2017, 56, 1656−1671

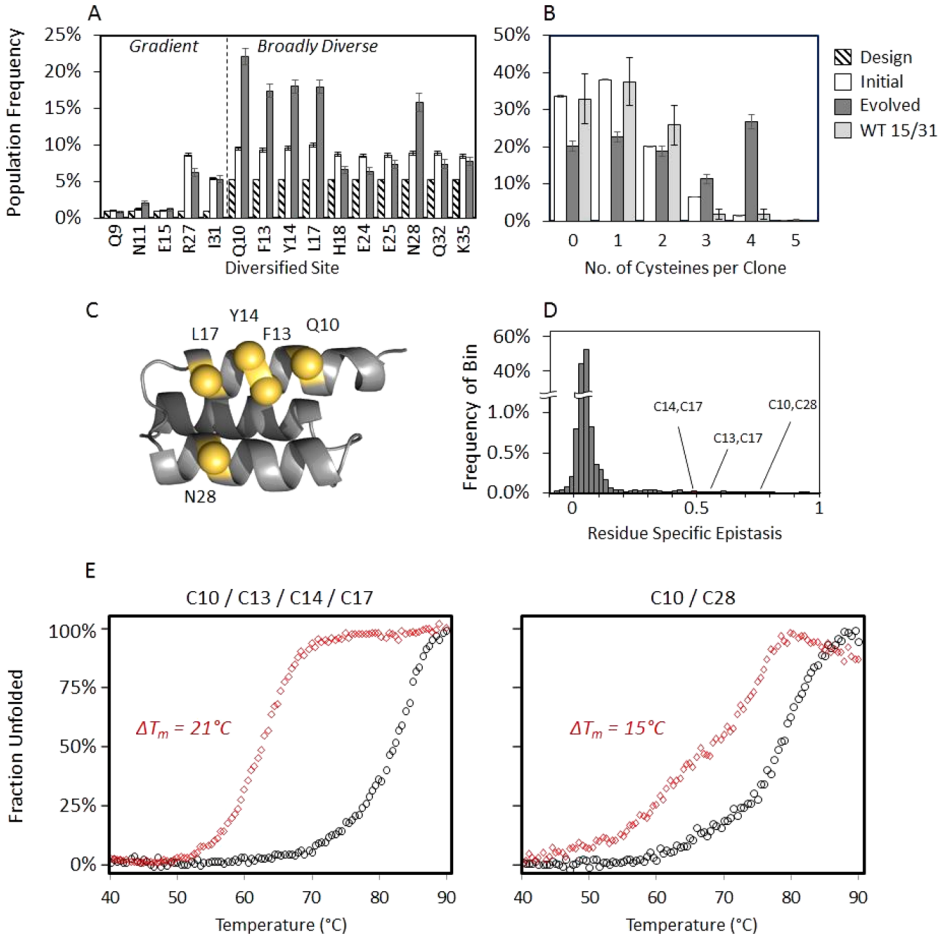


Figure 6. (A) Cysteine content in the first-generation initial library (solid, white bars) that was higher than intended (striped, light bars) and was strongly increased within evolved variants (solid, dark bars) at five broadly diversified sites. (B) Propensity of cysteine(s) within single variants from the first-generation initial and evolved populations as well as a subset of evolved variants in which wild-type residues are observed at positions E15 and I31 (light gray bars). (C) Cysteine rich positions observed within the first-generation evolved variants labeled and colored gold. (D) Residue-specific epistasis[74] was quantified for every observed combination of amino acid and position. The three most prominent cysteine pairs within the first-generation evolved variants are indicated. (E) Change in thermal denaturation midpoint (ΔT_m) between reducing (red) and nonreducing (black) conditions for two high-affinity variants with the most prevalent cysteine pairings (indicated above each plot).

tolerated diversity in the first-generation library analyses. I was maintained at high levels (from 33 to 32%), but most other amino acids were also tolerated with the most significant depletions being S (from 18 to 10%) and Y (from 11 to 8%). A 30:70 mixture of I and the B* codon was used. D36 is conserved in previous libraries and naturally (80% D, 18% E, Shannon entropy of 0.20) yet is solvent accessible (66%), could be expected to contact the target in some cases, and is computationally predicted to be tolerant to mutation. Diversity could be beneficial both inter- and intramolecularly, although such diversity should be minimal to limit detriment. DN diversity was used. Collectively, these mild diversifications were evaluated within the second-generation library.

Selections and Evolution from the Second-Generation Library. We aimed to evaluate the discovery and evolutionary efficacy of sitewise designs. Thus, we performed a competition between the second-generation library design, GS, and the traditional NNK library with broad, near-uniform diversity. We also included the modified second-generation library, GS_{LC}, which has reduced cysteine content as well as modifications in the genetically coupled amino acid preferences (B_{LC}*, details below). The second-generation libraries (Table S1 and Figure S1) were synthesized with custom degenerate oligonucleotides assembled by overlap extension PCR and introduced into the yeast display system by homologous recombination. A total of 1×10^9 variants were achieved for each of the three library designs. High-throughput sequence analysis of each library revealed the expected distributions on an amino acid basis [median absolute deviation from design, $|f_{observed} - f_{design}| = 0.5\%$ (Figure S2B–D)]. To evaluate the generalizable discovery/evolutionary efficacy, the pooled libraries were used to identify binders to a broad panel of seven new targets: death receptor 5, transferrin, cytochrome *c*, glucose-6-phosphate dehydrogenase, CD276, MET, and a GPCR. Specific binders were discovered and made to evolve from the combinatorial library using yeast display with magnetic and flow cytometry selections (Figure S5B). Affibody variants discovered and evolved from the second-generation library exhibited affinities from 2 ± 2 to 82 ± 23 nM (Figure S6A) with high target specificity (Figure S8). Deep sequencing of the evolved variants yielded 6×10^4 unique protein sequences in 5×10^3 diverse families, where families were dictated by having at least 80% homology throughout the library positions.

The study aimed to identify site-specific amino acid preferences consistent with evolutionary efficacy for a broad

DOI: 10.1021/acs.biochem.6b01142
Biochemistry 2017, 56, 1656–1671

array of epitopes. Thus, multiple targets, each with numerous potential epitopes, were used for binder discovery and evolution. While target-specific—and, more importantly, epitope-specific—preferences may be present, the diversity of evolved sequences (6×10^4 unique sequences in 5×10^3 families) mitigates the impact of such biases on the broad analysis.

Library Efficacy Comparison. The library origin of the evolved binders was determined by probabilistic sequence analysis after clustering into families on the basis of 80% similarity (Materials and Methods). Variants from the GS library were enriched, whereas variants from the NNK library were depleted relative to the initial library [$p < 0.001$ (Figure 7)], which demonstrates the improved evolutionary efficiency

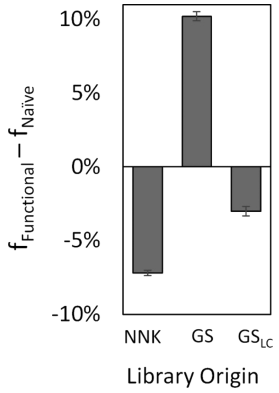


Figure 7. Enrichment of second-generation libraries. Sequences were analyzed for the likelihood of originating from either the NNK, GS, or GS_{LC} library designs. The change in prevalence (frequency in functional population − frequency in naïve library) between the initial library and binding populations is shown. Significant enrichment of the GS library was observed at the expense of depletion of both NNK and GS_{LC} ($p < 0.001$ for GS vs each).

of the sitewise amino acid preferences in the GS library design. This evolutionary benefit of the GS design is even more striking (Figure S7) in a population that was further enriched for even stronger affinity binding [affinities from 0.9 ± 0.1 to 5 ± 2 nM (Figure S6)]. These results summarize the collective advantage of sitewise constraint in evolutionary discovery. Additional analyses and experiments, which follow, were performed to further elucidate the molecular aspects of this advantage.

GS was also superior to GS_{LC}, which differed only in the codon design at the nine broadly diversified sites. Thus, the evolutionary preference for the B* codon design over the B_{LC}* codon was examined to elucidate relative efficacies of each amino acid. As noted, the motivation to create B_{LC}* was to reduce cysteine content (from 2.5 to 0.5%) to isolate disulfide-free and thiol-free binders. Unlike the first-generation library, in the context of the additional amino acid diversity constraint in the second-generation library, the cysteine content is depleted during evolution (Figure 8). In particular for the GS library, the contents of variants with two or more cysteines are dramatically reduced (59-fold for two-cysteine variants and 51-fold for three-cysteine variants). Cysteine depletion is observed during evolution of binders from the NNK library but to a reduced extent (2- and 28-fold for two- and three-cysteine variants, respectively) relative to the sitewise constrained library.

Via genetic code coupling, cysteine reduction also results in reduction of F, S, L, Y, W, G, and R; the aim for the antibody-

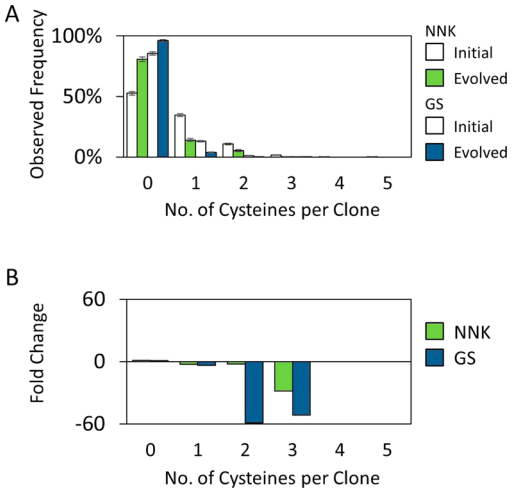


Figure 8. From the second-generation libraries, both the NNK and GS designs yielded fewer cysteines in the evolved populations compared to the initial libraries. (A) Absolute and (B) relative changes in frequency are shown. These modest levels were even slightly depleted among evolved variants.

and first generation affibody-inspired amino acid distribution also results in increases for D, E, H, K, N, and Q (Figure 9A). Of the amino acids with a higher frequency in the initial B* codon sites versus B_{LC}*, W and L are enriched and F and Y are maintained at high levels upon evolution of the binder (Figure 9BC). Conversely, S, G, R, and C are depleted. This is consistent with the evolutionary benefit of these aromatics and hydrophobics leading to the superiority of B* despite the evolutionary inefficiency of S, G, R, and C. Notably, the results support the aforementioned rationalized reduction in the levels of G and R relative to antibody CDR-H3. Of the amino acids with a lower frequency in the initial B* codon sites versus B_{LC}*, levels of H, K, N, and Q are reduced or kept low. Conversely, E is enriched and D is maintained at a high level. These results suggest that depleted H, K, N, and Q also contribute to B* superiority versus B_{LC}* whereas depleted E hinders B*.

Sitewise Amino Acid Frequencies. Sitewise amino acid frequencies (Figure 10) provide valuable insight for elucidating the evolutionary benefit of the constrained GS design and guiding further refinement. P, Q, K, and C are consistently depleted at broadly diversified sites in evolved binders. S is depleted at all sites but site 18. The aforementioned benefit of W is especially observed at sites 14 and 17, adjacent sites on the "top" of helix 1, and is also mildly enriched at other sites. Two other residues that improve B* relative to B_{LC}*, F and L (and homologue I), are generally enriched in helix 2.

At site 9, enrichment of W (from 16% to 31%) and L (from 12% to 21%), as well as maintenance of a high level of hydrophobic M (5%), in evolved binders demonstrates the value in constraining diversity to increase the initial frequencies of these residues within the GS library, yet depletion of the remainder of the constrained subset suggests further constraint on WLM could be advantageous to both increase the frequency of these beneficial residues and decrease the frequency of detrimental options, especially cationic R (from 17% to 5%) and K (from 7% to 2%). At site 11, which is relatively exposed (51% SASA), the strong hydrophilics D and N are enriched whereas the intermediate hydrophilics S, Y, T, and A are depleted. Thus, the GS design benefits from D and N bias but is hindered by bias to S, Y, T, and A. At site 13, aromatics (F

DOI: 10.1021/acs.biochem.6b01142
Biochemistry 2017, 56, 1656−1671

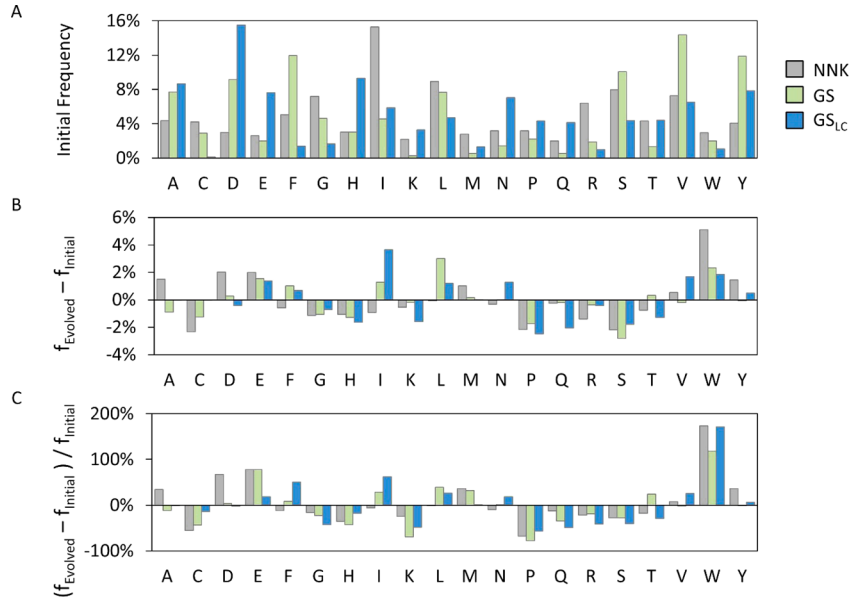


Figure 9. Nine sites were offered at least 50% GS/GS$_{LC}$ diversity (Q10, Y14, L17, H18, E25, N28, I31, Q32, and K35). A comparison between the initial and evolved amino acid frequencies at these sites is shown. (A) Individual frequencies for each of the three sublibraries. (B) Absolute and (C) relative changes in frequency.

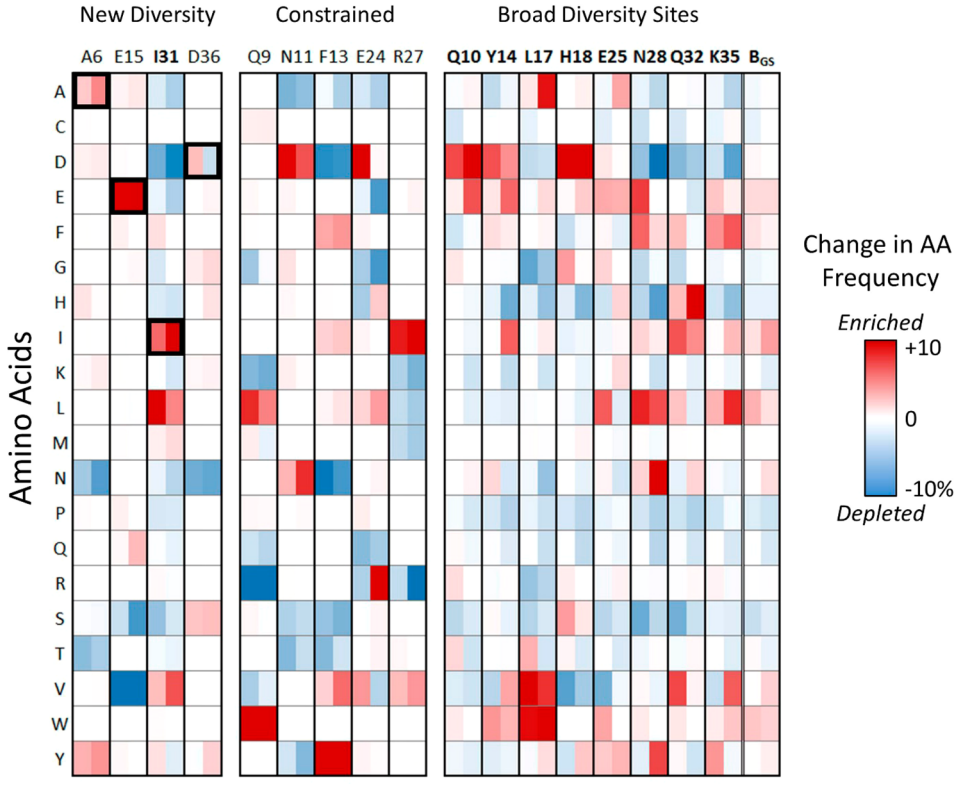


Figure 10. Changes in frequency between the initial and evolved populations are shown for the GS (left side of each column) and GS$_{LC}$ (right) design campaigns. Bold sites indicate the nine broadly diversified positions. These nine sites are averaged in the B$_{GS}$ column on the right.

and Y) and hydrophobics [F, I, and V from already elevated initial frequencies, and L from rare mutagenic PCR (from 0.1% to 0.7%)] were enriched in evolved binders whereas hydrophilics (D, N, S, and T) and the small A were depleted. Further hydrophobic constraint of this modestly buried site (27% SASA) to FLIV would likely provide additional benefit. Site 24 (88% SASA) generally tolerates its relatively broad diversity. The reasonably buried (27% SASA) site 27 benefits from bias to hydrophobic I and V, which are strongly enriched in evolved binders, but is hindered by bias to R, K, M, and L, which are depleted upon evolution.

Strong wild-type enrichment is observed at several of the newly diversified sites (Figures 10 and 11A). At site 15, wild-type E enrichment is countered by depletion of S and V while wild-type homologue Q is slightly enriched. At site 31, in addition to wild-type I, homologue L and similar hydrophobe V

DOI: 10.1021/acs.biochem.6b01142
Biochemistry 2017, 56, 1656−1671

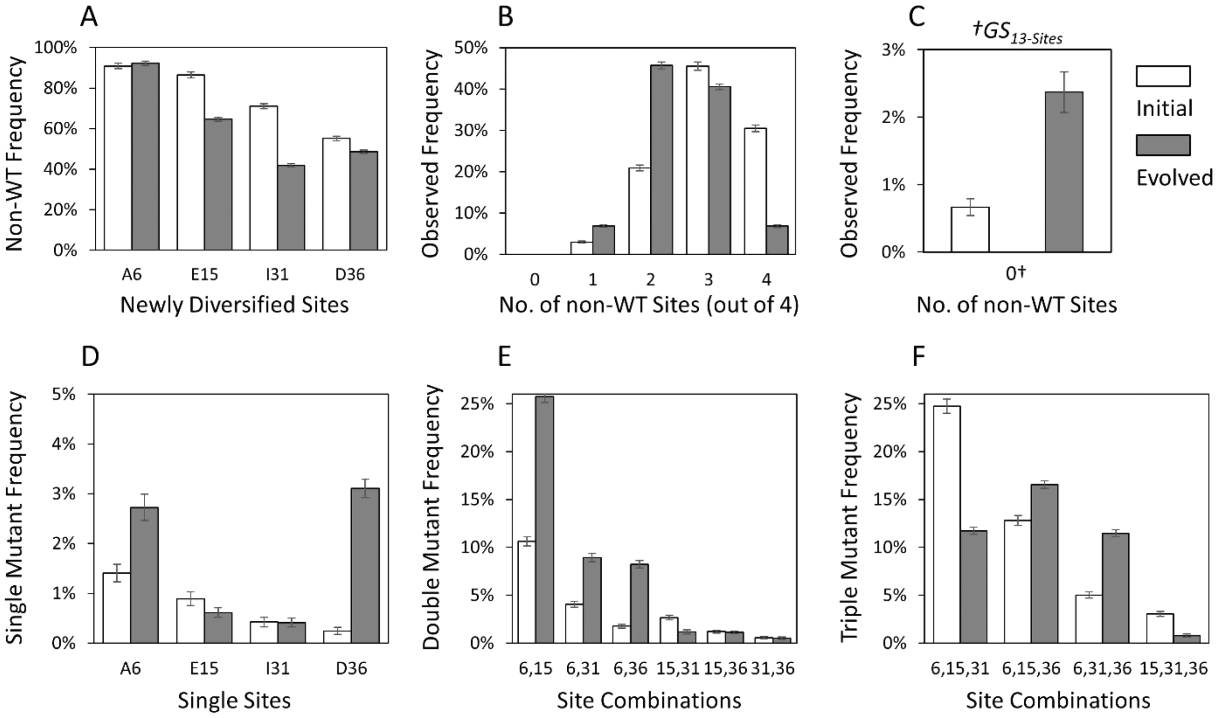


Figure 11. Among the four newly diversified sites, the tendency for wild-type conservation or diversification is shown. (A) Frequency of variants having a residue other than that of the wild type at each site. (B) Distribution of variants from either the initial or evolved populations having exactly zero, one, two, three, or four non-wild-type amino acids. (C) Frequency of clones having WT amino acids at all four newly diversified sites. Frequencies of (D) single, (E) double, and (F) triple mutants exclusively at the position(s) listed on the *x*-axis. The sequences analyzed in panels A, B, and D–F all meet the requirement of originating from the GS library based on probabilistic calculations at all 17 diversified sites. However, the sequences contributing to panel C meet the requirement of originating from the GS library based on probabilistic calculations at only the 13 traditionally mutated sites (†$GS_{13\text{-Sites}}$).

are enriched while numerous amino acids are depleted. Site 36 reveals a more modest shift in its wild type, D, along with significant depletion of the alternative mutant offering, N. At site 6, the Z-domain wild-type N is depleted, as well as T; S is maintained, and Y is enriched. Notably, alanine, conserved in the first-generation library because of its benefit in the optimized affibody backbone[8] and able to appear in the second generation via mutagenic PCR and homologous recombination with a linearized vector encoding A, is enriched. As a result of these wild-type preferences, variants with mutation at three or four of these newly diversified sites are depleted (Figure 11B). Interestingly, though wild-type conservation is modestly preferred at site 36, mutation is strongly enriched in evolved binders if it is the only mutation within this set of newly diversified sites (Figure 11D) or it occurs in tandem with mutation at site 6 (Figure 11E), but co-mutation of site 36 with site 15 or 31 is depleted in binders. Dual mutation of sites 15 and 31, with or without further mutation at site 6 or 15, is strongly depleted. Overall, in the evaluation of the newly created diversity at sites 6, 15, 31, and 36, mild diversity is tolerable, but further constraint from the second-generation design would improve evolutionary efficiency. The ability of the GS library to outperform NNK despite the detrimental overdiversification of these sites indicates significant evolutionary value to the other modifications: B* codon biased diversity rather than NNK at broadly diversified sites and constrained diversity at sites 9, 11, 13, 24, and 27. Analysis of the traditionally diversified 13 sites in functional variants reveals 13-fold enrichment of variants of constrained design (Figure 12).

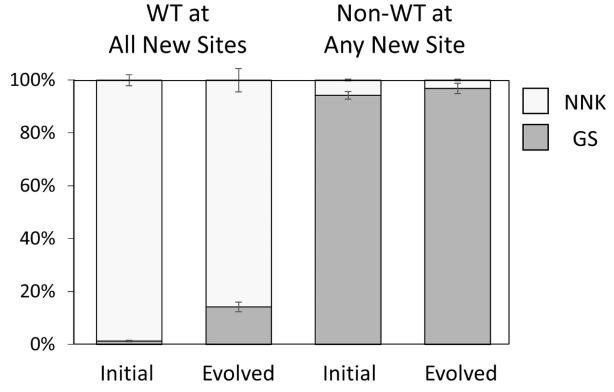


Figure 12. Sequences were placed into one of two bins: one in which variants were entirely wild-type (WT) at all four newly diversified positions (A6, E15, I31, and D36) and a second in which were collected the variants that contained a non-WT amino acid at any of the four new sites. Within each bin, sequences were analyzed at the 13 classically diversified positions for the likelihood of having originated from either the NNK or GS library design.

The NNK library provides a naïve design benchmark, which demonstrated the overall evolutionary advantage of the GS library (Figure 7), and can also be used assess the merit of sitewise amino acid preferences. The changes in sitewise amino acid frequencies from the initial NNK library to binders that evolved from the NNK library (i.e., observed experimental evolution) were compared to the changes designed into the constrained sites of the GS library (i.e., predicted to improve evolution) (Figure 13A). Experimental values correlate with

DOI: 10.1021/acs.biochem.6b01142
Biochemistry 2017, 56, 1656–1671

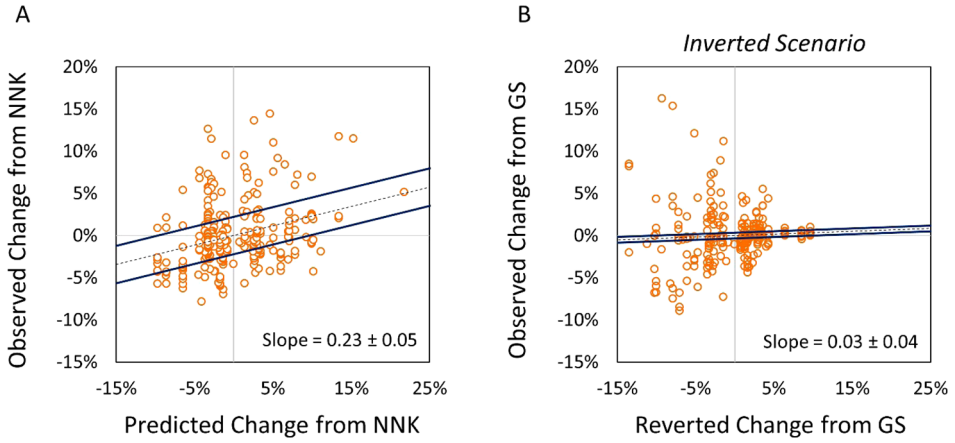


Figure 13. (A) Observed frequency change that reflects the shift in amino acid frequency that was observed within variants having originated from the NNK library design. The predicted frequency change shows the absolute difference in amino acid frequency between the GS and NNK design. Observed change from NNK: $f_{AA\ i,\ site\ j\ (NNK\ evolved)} - f_{AA\ i,\ site\ j\ (NNK\ initial)}$. Predicted change from NNK: $f_{AA\ i,\ site\ j\ (GS\ design)} - f_{AA\ i,\ site\ j\ (NNK\ initial)}$. The 95% prediction interval is shown. (B) Similar to panel A, but for evaluation of the predicted and observed mutations away from the GS design. Observed change from GS: $f_{AA\ i,\ site\ j\ (GS\ evolved)} - f_{AA\ i,\ site\ j\ (GS\ initial)}$. Reverted change from GS: $f_{AA\ i,\ site\ j\ (NNK\ design)} - f_{AA\ i,\ site\ j\ (GS\ initial)}$.

predicted design (slope = 0.23 ± 0.05; $p < 0.001$). For comparison, evaluation of evolution away from the GS library indicates negligible correlation [slope = 0.03 ± 0.04; $p = 0.38$ (Figure 13B)]. This further supports the evolutionary merit of the GS design relative to naïve NNK. Evaluation of experimental versus design correlation at each site reveals comparable correlation at most sites with an especially strong correlation at site 17, driven by the predicted enrichment of Y and D as well as depletion of R, and a lack of correlation at sites 9, 24, and 28 (Figure S4). Overall, the sitewise biases of the GS library design have proven to be superior to those of the broad, near-uniform NNK library for binder evolution.

Stability. Constrained diversity is designed, in a sitewise manner, to elevate the frequency of evolutionarily beneficial amino acids while reducing the frequency of detrimental residues to search more fruitful regions of sequence space. One expected mechanism of this hypothesis is that select detrimental residues destabilize the scaffold, thereby precluding potentially effective binding paratopes because of their entropic cost[99,100] or enthalpic destabilization beyond a foldable limit.[22,24] It has been shown that more stable scaffolds allow improved evolution.[22,24,101] A related, but distinct, hypothesis is that a reduced level of destabilization upon mutation improves evolvability. That is, a combinatorial library that contains variants that are less destabilized will contain a larger fraction of folded variants as well as a smaller entropic penalty upon binding (Figure 14A). To partially address this hypothesis, thermal stabilities of several evolved binders were measured by thermal denaturation and circular dichroism spectroscopy (Figure S5A). Random binding variants that evolved from the GS library exhibit thermal stabilities (T_m = 62 ± 4 °C) higher than those of variants from NNK libraries either in this study (T_m = 42 ± 12 °C; $p = 0.02$) or in the literature (49 ± 8 °C; $p < 0.001$) (Figure 14B). Notably, the NNK-based binders from this study are not more stable than NNK-based binders in the literature. Thus, the eukaryotic expression machinery of yeast surface display (compared to phage display selections from the literature) is not responsible for stability enhancement. Rather, the amino acid constraint in the GS library design accounts for this stabilization.

The improved performance of the GS library was achieved by biasing amino acid diversity, including eliminating select amino acid options, within the 13 sites traditionally diversified while also expanding diversity by varying four previously conserved sites. The constraint reduced possible sequence space 80-fold and also biased the search of the possible space by the preferential occurrence of select amino acids. This biased diversity was 13 ± 3-fold more effective than the uniformly applied NNK diversity (Figure 12). The introduced diversity at sites 6, 15, 31, and 36 increased sequence space 800-fold; variants with zero to two mutations, particularly at sites 6 and 36, were found to be evolutionarily effective, whereas triple and quadruple mutants were less functional (Figure 11). Thus, while diversity at these sites is functional, they would benefit from further bias. Overall, the GS library has a 10-fold greater potential sequence space (Figure 14C) but searches this space in a manner more biased than that of the NNK library. The average uniformity of diversity, as measured by Shannon entropy, of the NNK library is 0.97 versus 0.74 for the GS library (Figure S9). Overall, the result of the sitewise bias in the GS library is a higher frequency of binder discovery (Figure 7 and Figure S7) and more stable binders (Figure 14B).

CONCLUSION

In this study, amino acid frequency distributions for each site in a combinatorial affibody library were designed from a multitude of inputs: high-throughput binder evolution, computed mutant stabilities, target accessibility, helical propensity, chemical complementarity, and frequency in natural homologues. We evaluated the efficiency of the gradient sitewise (GS) design by direct competition of multiple libraries in the context of binder discovery against a collection of protein targets. Site-specific amino acid bias, as well as introduced diversity at four additional positions, in the GS library allowed evolution that was more efficient than that produced by the traditional approach of homogeneous diversity across all diversified sites in the NNK library. Analysis indicates that GS's evolutionary benefit resulted from the sitewise constraint and in spite of the broadened diversity at sites 15 and 31. Amino acid preferences, both overall and at select sites, were revealed. The first-generation library, designed to explore sitewise amino acid preference from a broad range of paratopes, exhibited diversity much broader than optimal, as assessed by sitewise frequencies

DOI: 10.1021/acs.biochem.6b01142
Biochemistry 2017, 56, 1656−1671

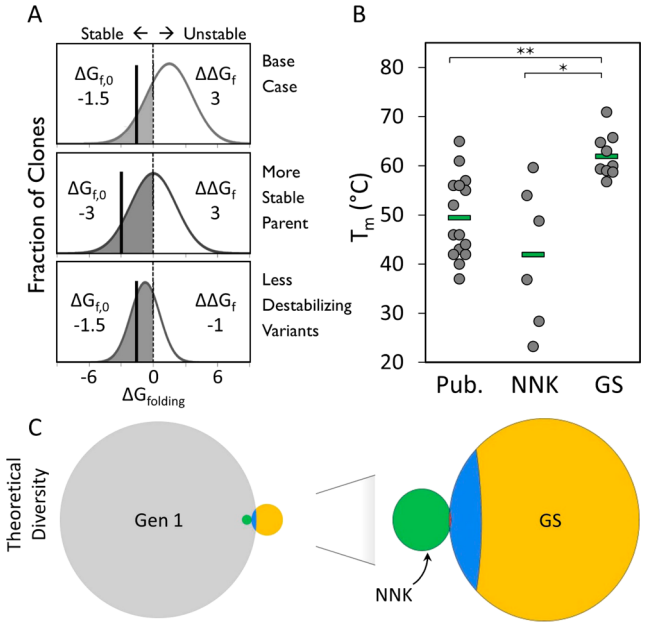


Figure 14. (A) Hypothetical stability distributions of mutant populations derived from a single starting point (i.e., parental clone). A parental clone (vertical solid line labeled $\Delta G_{f,0}$) is randomly mutated, yielding a population with distributed stability. Properly folding proteins are located to the left of the dashed line where the free energy of folding is negative (shaded region). A base case is shown in the top panel. The middle panel presents a more stable parent ($\Delta G_{f,0} = -3$) with an equivalent broad distribution of destabilization like the base case ($\Delta\Delta G_f = 3$). A third example is shown in the bottom panel where the parental clone has stability equivalent to that of the base case ($\Delta G_{f,0} = -3$); however, upon mutation, a much lower extent of destabilization is observed throughout the population. This illustrates situations in which, relative to the base case, an increased fraction in functional variants can be attained using either a more stable starting point or a reduced level of destabilization upon mutation. (B) Clonal stability of several high-affinity binders analyzed using circular dichroism. Measurements of randomly chosen clones having originated from either the gradient sitewise (GS) or NNK design are shown alongside stability measurements previously reported in the literature. Average stabilities (green bars) were found to be 49 ± 8, 42 ± 12, and 62 ± 4 °C for published, NNK, and GS variants, respectively. (C) Relative sequence space afforded by each of the libraries discussed in this work: first generation (Gen 1, gray), traditional design (NNK, green), and second generation (GS, gold). Sequence diversity coverage shared between Gen 1 and GS is colored blue. Sequence space existing within both NNK and GS is colored red between the green and blue regions. Relative to NNK (13 library sites), the theoretical sequence diversity values of GS (17 sites) and Gen 1 (15 sites) are larger by 10- and 400-fold, respectively.

in evolved repertoires. In fact, these "overdiversified" paratopes frequently required cysteine pairs to achieve functionality. Conversely, reduction of the levels of undesirable amino acids at select sites in the second-generation library greatly reduced the presence of cysteine pairs in evolved clones. Moreover, the sitewise bias yielded variants (15 °C higher midpoint of thermal destabilization) with a stability higher than that of NNK-based variants. Evolutionary efficiency benefits from increased frequency of not only amino acids that drive intermolecular interactions but also intramolecularly tolerated mutants that weaken destabilization. Overall, the library design approach favoring constrained amino acid diversities that take into account complementarity, amino acid frequencies in previously discovered binders, diversity in natural homologues, and solvent-exposed surface area produces binding ligands more efficiently than the unconstrained NNK library design in the affibody scaffold.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.biochem.6b01142.

Wild-type affibody sequence and library design summary (Table S1), comparison of broad diversity codon design (Figure S1), design versus observed initial libraries (Figure S2), first-generation evolved binder campaign cysteine content (Figure S3), predictive analysis of GS versus NNK library designs (Figure S4), natural sequence frequencies of affibody homologues, pfam B (PF02216) (Table S2), second-generation library site-wise design summary and oligo list (Table S3), biophysical characterization of representative samples (Figure S5), affinity titration of high- and ultra-high-stringency variants (Figure S6), library of origin comparison for high- and ultra-high-stringency variants (Figure S7), and specificity analysis for high-stringency variants (Figure S8), and uniformity of diversity (Shannon entropy) calculations (Figure S9) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Phone: 612-624-7102. E-mail: hackel@umn.edu.

ORCID

Benjamin J. Hackel: 0000-0003-3561-9463

Funding

This work was supported by National Institutes of Health Grant R21 E019518 to B.J.H.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We appreciate assistance from Baradan Panta and Nicholas Heise in assembling the database of published evolved affibodies, Justin Klesmith for valuable suggestions about the manuscript, the University of Minnesota Genomics Center for assistance with Illumina sequencing, the Masonic Cancer Center Flow Cytometry Shared Resource, and the University of Minnesota Supercomputing Institute for computational resources during sequence analysis.

■ ABBREVIATIONS

CDR-H3, third antibody heavy chain complementarity-determining region; CD276, cluster of differentiation 276; DTT, dithiothreitol; MET or HGFR, hepatocyte growth factor receptor; GPCR, G-protein-coupled receptor; IgG, immunoglobulin G; PBSA, phosphate-buffered saline with 0.1% (w/v) bovine serum albumin; PCR, polymerase chain reaction; WT, wild type.

■ REFERENCES

(1) Banta, S., Dooley, K., and Shur, O. (2013) Replacing antibodies: engineering new binding proteins. *Annu. Rev. Biomed. Eng. 15*, 93−113.
(2) Beerli, R. R., and Rader, C. (2010) Mining human antibody repertoires. *MAbs 2*, 365−378.

DOI: 10.1021/acs.biochem.6b01142
Biochemistry 2017, 56, 1656−1671

(3) Alder, M. N. (2005) Diversity and Function of Adaptive Immune Receptors in a Jawless Vertebrate. *Science 310*, 1970−1973.
(4) Binz, H. K., Amstutz, P., and Plückthun, A. (2005) Engineering novel binding proteins from nonimmunoglobulin domains. *Nat. Biotechnol. 23*, 1257−1268.
(5) Nord, K., Gunneriusson, E., Ringdahl, J., Ståhl, S., Uhlén, M., and Nygren, P.-Å. (1996) Binding proteins selected from combinatorial libraries of an α-helical bacterial receptor domain. *Nat. Biotechnol. 15*, 772−777.
(6) Löfblom, J., Feldwisch, J., Tolmachev, V., Carlsson, J., Ståhl, S., and Frejd, F. Y. (2010) Affibody molecules: engineered proteins for therapeutic, diagnostic and biotechnological applications. *FEBS Lett. 584*, 2670−2680.
(7) Lendel, C., Dincbas-Renqvist, V., Flores, A., Wahlberg, E., Dogan, J., Nygren, P.-Å., and Härd, T. (2004) Biophysical characterization of Z SPA-1 -A phage-display selected binder to protein A. *Protein Sci. 13*, 2078−2088.
(8) Feldwisch, J., Tolmachev, V., Lendel, C., Herne, N., Sjöberg, A., Larsson, B., Rosik, D., Lindqvist, E., Fant, G., Höidén-Guthenberg, I., Galli, J., Jonasson, P., and Abrahmsén, L. (2010) Design of an optimized scaffold for affibody molecules. *J. Mol. Biol. 398*, 232−247.
(9) Lundberg, E., Brismar, H., and Gräslund, T. (2009) Selection and characterization of Affibody ligands to the transcription factor c-Jun. *Biotechnol. Appl. Biochem. 52*, 17−27.
(10) Nygren, P. Å. (2008) Alternative binding proteins: Affibody binding proteins developed from a small three-helix bundle scaffold. *FEBS J. 275*, 2668−2676.
(11) Lindborg, M., Cortez, E., Höidén-Guthenberg, I., Gunneriusson, E., von Hage, E., Syud, F., Morrison, M., Abrahmsén, L., Herne, N., Pietras, K., and Frejd, F. Y. (2011) Engineered high-affinity affibody molecules targeting platelet-derived growth factor receptor β in vivo. *J. Mol. Biol. 407*, 298−315.
(12) Kronqvist, N., Malm, M., Göstring, L., Gunneriusson, E., Nilsson, M., Höidén-Guthenberg, I., Gedda, L., Frejd, F. Y., Ståhl, S., and Löfblom, J. (2011) Combining phage and staphylococcal surface display for generation of ErbB3-specific Affibody molecules. *Protein Eng., Des. Sel. 24*, 385−396.
(13) Gunneriusson, E., Nord, K., Uhlén, M., and Nygren, P. (1999) Affinity maturation of a Taq DNA polymerase specific affibody by helix shuffling. *Protein Eng., Des. Sel. 12*, 873−878.
(14) Eklund, M., Axelsson, L., Uhlén, M., and Nygren, P.-Å. (2002) Anti-idiotypic protein domains selected from protein A-based affibody libraries. *Proteins: Struct., Funct., Genet. 48*, 454−462.
(15) Grimm, S., Lundberg, E., Yu, F., Shibasaki, S., Vernet, E., Skogs, M., Nygren, P.-Å., and Gräslund, T. (2010) Selection and characterisation of affibody molecules inhibiting the interaction between Ras and Raf in vitro. *New Biotechnol. 27*, 766−773.
(16) Grimm, S., Salahshour, S., and Nygren, P.-Å. (2012) Monitored whole gene in vitro evolution of an anti-hRaf-1 affibody molecule towards increased binding affinity. *New Biotechnol. 29*, 534−542.
(17) Wernérus, H., Samuelson, P., and Ståhl, S. (2003) Fluorescence-activated cell sorting of specific affibody-displaying staphylococci. *Appl. Environ. Microbiol. 69*, 5328−5335.
(18) Kronqvist, N., Löfblom, J., Jonsson, A., Wernérus, H., and Ståhl, S. (2008) A novel affinity protein selection system based on staphylococcal cell surface display and flow cytometry. *Protein Eng., Des. Sel. 21*, 247−255.
(19) Lindberg, H., Johansson, A., Härd, T., Ståhl, S., and Löfblom, J. (2013) Staphylococcal display for combinatorial protein engineering of a head-to-tail affibody dimer binding the Alzheimer amyloid-β peptide. *Biotechnol. J. 8*, 139−145.
(20) Case, B. A., and Hackel, B. J. (2016) Synthetic and natural consensus design for engineering charge within an affibody targeting epidermal growth factor receptor. *Biotechnol. Bioeng. 113*, 1628−1638.
(21) Chen, J., Sawyer, N., and Regan, L. (2013) Protein-protein interactions: general trends in the relationship between binding affinity and interfacial buried surface area. *Protein Sci. 22*, 510−5.
(22) Bloom, J. D., Labthavikul, S. T., Otey, C. R., and Arnold, F. H. (2006) Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A. 103*, 5869−5874.
(23) Dellus-Gur, E., Toth-Petroczy, A., Elias, M., and Tawfik, D. S. (2013) What Makes a Protein Fold Amenable to Functional Innovation? Fold Polarity and Stability Trade-offs. *J. Mol. Biol. 425*, 2609−2621.
(24) Tokuriki, N., and Tawfik, D. S. (2009) Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol. 19*, 596−604.
(25) Tokuriki, N., Stricher, F., Serrano, L., and Tawfik, D. S. (2008) How protein stability and new functions trade off. *PLoS Comput. Biol. 4*, e1000002.
(26) Nagatani, R. A., Gonzalez, A., Shoichet, B. K., Brinen, L. S., and Babbitt, P. C. (2007) Stability for Function Trade-Offs in the Enolase Superfamily "Catalytic Module. *Biochemistry 46*, 6688−6695.
(27) Mukaiyama, A., Haruki, M., Ota, M., Koga, Y., Takano, K., and Kanaya, S. (2006) A hyperthermophilic protein acquires function at the cost of stability. *Biochemistry 45*, 12673−12679.
(28) Hackel, B. J., Ackerman, M. E., Howland, S. W., and Wittrup, K. D. (2010) Stability and CDR Composition Biases Enrich Binder Functionality Landscapes. *J. Mol. Biol. 401*, 84−96.
(29) Woldring, D. R., Holec, P. V., Zhou, H., and Hackel, B. J. (2015) High-Throughput Ligand Discovery Reveals a Sitewise Gradient of Diversity in Broadly Evolved Hydrophilic Fibronectin Domains. *PLoS One 10*, e0138956.
(30) Zemlin, M., Klinger, M., Link, J., Zemlin, C., Bauer, K., Engler, J. A., Schroeder, H. W., and Kirkham, P. M. (2003) Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. *J. Mol. Biol. 334*, 733−749.
(31) Fellouse, F. A., Esaki, K., Birtalan, S., Raptis, D., Cancasci, V. J., Koide, A., Jhurani, P., Vasser, M., Wiesmann, C., Kossiakoff, A. A., Koide, S., and Sidhu, S. S. (2007) High-throughput generation of synthetic antibodies from highly functional minimalist phage-displayed libraries. *J. Mol. Biol. 373*, 924−940.
(32) Binz, H. K., Amstutz, P., Kohl, A., Stumpp, M. T., Briand, C., Forrer, P., Grütter, M. G., and Plückthun, A. (2004) High-affinity binders selected from designed ankyrin repeat protein libraries. *Nat. Biotechnol. 22*, 575−582.
(33) Seeger, M. A., Zbinden, R., Flütsch, A., Gutte, P. G. M., Engeler, S., Roschitzki-Voser, H., and Grütter, M. G. (2013) Design, construction, and characterization of a second-generation DARPin library with reduced hydrophobicity. *Protein Sci. 22*, 1239−1257.
(34) Schilling, J., Schöppe, J., and Plückthun, A. (2014) From DARPins to LoopDARPins: novel LoopDARPin design allows the selection of low picomolar binders in a single round of ribosome display. *J. Mol. Biol. 426*, 691−721.
(35) Koide, A., Wojcik, J., Gilbreth, R. N., Hoey, R. J., and Koide, S. (2012) Teaching an Old Scaffold New Tricks: Monobodies Constructed Using Alternative Surfaces of the FN3 Scaffold. *J. Mol. Biol. 415*, 393−405.
(36) Diem, M. D., Hyun, L., Yi, F., Hippensteel, R., Kuhar, E., Lowenstein, C., Swift, E. J., O'Neil, K. T., and Jacobs, S. a. (2014) Selection of high-affinity Centyrin FN3 domains from a simple library diversified at a combination of strand and loop positions. *Protein Eng., Des. Sel. 27*, 419−29.
(37) Fellouse, F. a, Wiesmann, C., and Sidhu, S. S. (2004) Synthetic antibodies from a four-amino-acid code: a dominant role for tyrosine in antigen recognition. *Proc. Natl. Acad. Sci. U. S. A. 101*, 12467−12472.
(38) Fellouse, F. a, Li, B., Compaan, D. M., Peden, A. a, Hymowitz, S. G., and Sidhu, S. S. (2005) Molecular recognition by a binary code. *J. Mol. Biol. 348*, 1153−62.
(39) Koide, S., and Sidhu, S. S. (2009) The importance of being tyrosine: lessons in molecular recognition from minimalist synthetic binding proteins. *ACS Chem. Biol. 4*, 325−334.
(40) Wojcik, J., Hantschel, O., Grebien, F., Kaupe, I., Bennett, K. L., Barkinge, J., Jones, R. B., Koide, A., Superti-Furga, G., and Koide, S.

DOI: 10.1021/acs.biochem.6b01142
Biochemistry 2017, 56, 1656−1671

(2010) A potent and highly specific FN3 monobody inhibitor of the Abl SH2 domain. *Nat. Struct. Mol. Biol. 17*, 519−527.
(41) Lee, C. V., Liang, W.-C., Dennis, M. S., Eigenbrot, C., Sidhu, S. S., and Fuh, G. (2004) High-affinity human antibodies from phage-displayed synthetic Fab libraries with a single framework scaffold. *J. Mol. Biol. 340*, 1073−1093.
(42) Araya, C. L., and Fowler, D. M. (2011) Deep mutational scanning: Assessing protein function on a massive scale. *Trends Biotechnol. 29*, 435−442.
(43) Fowler, D. M., Stephany, J. J., and Fields, S. (2014) Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat. Protoc. 9*, 2267−2284.
(44) Fowler, D. M., and Fields, S. (2014) Deep mutational scanning: a new style of protein science. *Nat. Methods 11*, 801−807.
(45) Wrenbeck, E. E., Faber, M. S., and Whitehead, T. A. (2017) Deep sequencing methods for protein engineering and design. *Curr. Opin. Struct. Biol. 45*, 36−44.
(46) Rockah-Shmuel, L., Tóth-Petróczy, Á., and Tawfik, D. S. (2015) Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations. *PLoS Comput. Biol. 11*, e1004421.
(47) Smith, C. A., and Kortemme, T. (2011) Predicting the Tolerated Sequences for Proteins and Protein Interfaces Using RosettaBackrub Flexible Backbone Design. *PLoS One 6*, e20451.
(48) Au, L., and Green, D. F. (2016) Direct Calculation of Protein Fitness Landscapes through Computational Protein Design. *Biophys. J. 110*, 75−84.
(49) Guerois, R., Nielsen, J. E., and Serrano, L. (2002) Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J. Mol. Biol. 320*, 369−387.
(50) Magliery, T. J. (2015) Protein stability: Computation, sequence statistics, and new experimental methods. *Curr. Opin. Struct. Biol. 33*, 161−168.
(51) Brender, J. R., and Zhang, Y. (2015) Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles. *PLoS Comput. Biol. 11*, e1004494.
(52) Xiong, P., Zhang, C., Zheng, W., and Zhang, Y. (2016) BindProfX: Assessing mutation-induced binding affinity change by protein interface profiles with pseudo counts. *J. Mol. Biol. 429*, 426.
(53) Birtalan, S., Zhang, Y., Fellouse, F. A., Shao, L., Schaefer, G., and Sidhu, S. S. (2008) The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. *J. Mol. Biol. 377*, 1518−1528.
(54) Fellouse, F. A., Barthelemy, P. A., Kelley, R. F., and Sidhu, S. S. (2006) Tyrosine plays a dominant functional role in the paratope of a synthetic antibody derived from a four amino acid code. *J. Mol. Biol. 357*, 100−114.
(55) Engfeldt, T., Renberg, B., Brumer, H., Nygren, P.-Å., and Eriksson Karlström, A. (2005) Chemical synthesis of triple-labelled three-helix bundle binding proteins for specific fluorescent detection of unlabelled protein. *ChemBioChem 6*, 1043−1050.
(56) Friedman, M., Nordberg, E., Höidén-Guthenberg, I., Brismar, H., Adams, G. P., Nilsson, F. Y., Carlsson, J., and Ståhl, S. (2007) Phage display selection of Affibody molecules with specific binding to the extracellular domain of the epidermal growth factor receptor. *Protein Eng., Des. Sel. 20*, 189−199.
(57) Friedman, M., Orlova, A., Johansson, E., Eriksson, T. L. J., Höidén-Guthenberg, I., Tolmachev, V., Nilsson, F. Y., and Ståhl, S. (2008) Directed evolution to low nanomolar affinity of a tumor-targeting epidermal growth factor receptor-binding affibody molecule. *J. Mol. Biol. 376*, 1388−1402.
(58) Grimm, S., Yu, F., and Nygren, P.-Å. (2011) Ribosome display selection of a murine IgG1 Fab binding affibody molecule allowing species selective recovery of monoclonal antibodies. *Mol. Biotechnol. 48*, 263−76.
(59) Grimm, S. (2011) Ribosome display for selection and evolution of affibody molecules. Ph.D. Thesis, Royal Institute ot Technology, Stockholm.
(60) Grönwall, C., Jonsson, A., Lindström, S., Gunneriusson, E., Ståhl, S., and Herne, N. (2007) Selection and characterization of Affibody ligands binding to Alzheimer amyloid β peptides. *J. Biotechnol. 128*, 162−183.
(61) Grönwall, C., Snelders, E., Palm, A. J., Eriksson, F., Herne, N., and Ståhl, S. (2008) Generation of Affibody® ligands binding interleukin-2 receptor α/CD25. *Biotechnol. Appl. Biochem. 50*, 97.
(62) Hansson, M., Ringdahl, J., Robert, a, Power, U., Goetsch, L., Ngoc Nguyen, T., Uhlén, M., Ståhl, S., and Nygren, P. a. (1999) An in vitro selected binding protein (affibody) shows conformation-dependent recognition of the respiratory syncytial virus (RSV) G protein. *Immunotechnology 4*, 237−52.
(63) Högbom, M., Eklund, M., Nygren, P.-Å., and Nordlund, P. (2003) Structural basis for recognition by an in vitro evolved affibody. *Proc. Natl. Acad. Sci. U. S. A. 100*, 3191−3196.
(64) Jonsson, A. (2009) Development of molecular recognition by rational and combinatorial engineering. Ph.D. Thesis, Royal Institute of Technology, Stockholm.
(65) Jonsson, A., Wållberg, H., Herne, N., Ståhl, S., and Frejd, F. Y. (2009) Generation of tumour-necrosis-factor-alpha-specific affibody molecules capable of blocking receptor binding in vitro. *Biotechnol. Appl. Biochem. 54*, 93−103.
(66) Lendel, C. (2005) Molecular principles of protein stability and protein-protein interactions. Ph.D. Thesis, Royal Institute of Technology, Stockholm.
(67) Lendel, C., Dogan, J., and Härd, T. (2006) Structural Basis for Molecular Recognition in an Affibody:Affibody Complex. *J. Mol. Biol. 359*, 1293−1304.
(68) Li, J., Lundberg, E., Vernet, E., Larsson, B., Hoiden-Guthenberg, I., and Graslund, T. (2010) Selection of affibody molecules to the ligand-binding site of the insulin-like growth factor-1 receptor. *Biotechnol. Appl. Biochem. 55*, 99−109.
(69) Löfdahl, P., and Nygren, P. (2010) Affinity maturation of a TNFα-binding Affibody molecule by Darwinian survival selection. *Biotechnol. Appl. Biochem. 55*, 111−120.
(70) Löfdahl, P.-Å., Nord, O., Janzon, L., and Nygren, P.-Å. (2009) Selection of TNF-α binding affibody molecules using a β-lactamase protein fragment complementation assay. *New Biotechnol. 26*, 251−259.
(71) Nord, K., Nord, O., Uhlén, M., Kelley, B., Ljungqvist, C., and Nygren, P. A. (2001) Recombinant human factor VIII-specific affinity ligands selected from phage-displayed combinatorial libraries of protein A. *Eur. J. Biochem. 268*, 4269−4277.
(72) Orlova, A., Magnusson, M., Eriksson, T. L. J., Nilsson, M., Larsson, B., Höiden-Guthenberg, I., Widström, C., Carlsson, J., Tolmachev, V., Ståhl, S., and Nilsson, F. Y. (2006) Tumor imaging using a picomolar affinity HER2 binding Affibody molecule. *Cancer Res. 66*, 4339−4348.
(73) Rönnmark, J., Grönlund, H., Uhlén, M., and Nygren, P.-Å. (2002) Human immunoglobulin A (IgA)-specific ligands from combinatorial engineering of protein A. *Eur. J. Biochem. 269*, 2647−2655.
(74) Sandström, K., Xu, Z., Forsberg, G., and Nygren, P.-A. (2003) Inhibition of the CD28-CD80 co-stimulation signal by a CD28-binding affibody ligand developed by combinatorial protein engineering. *Protein Eng., Des. Sel. 16*, 691−697.
(75) Wahlberg, E., Lendel, C., Helgstrand, M., Allard, P., Dincbas-Renqvist, V., Hedqvist, A., Berglund, H., Nygren, P.-Å., and Härd, T. (2003) An affibody in complex with a target protein: structure and coupled folding. *Proc. Natl. Acad. Sci. U. S. A. 100*, 3185−3190.
(76) Wållberg, H., Löfdahl, P.-Å., Tschapalda, K., Uhlén, M., Tolmachev, V., Nygren, P.-Å., and Ståhl, S. (2011) Affinity recovery of eight HER2-binding affibody variants using an anti-idiotypic affibody molecule as capture ligand. *Protein Expression Purif. 76*, 127−135.
(77) Wikman, M., Steffen, A.-C., Gunneriusson, E., Tolmachev, V., Adams, G. P., Carlsson, J., and Ståhl, S. (2004) Selection and characterization of HER2/neu-binding affibody ligands. *Protein Eng., Des. Sel. 17*, 455−462.

DOI: 10.1021/acs.biochem.6b01142
Biochemistry 2017, 56, 1656−1671

(78) Wikman, M., Rowcliffe, E., Friedman, M., Henning, P., Lindholm, L., Olofsson, S., and Ståhl, S. (2006) Selection and characterization of an HIV-1 gp120-binding affibody ligand. *Biotechnol. Appl. Biochem. 45*, 93.
(79) Fraczkiewicz, R., and Braun, W. (1998) Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J. Comput. Chem. 19*, 319−333.
(80) Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005) The FoldX web server: an online force field. *Nucleic Acids Res. 33*, W382−W388.
(81) Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., and Punta, M. (2014) Pfam: The protein families database. *Nucleic Acids Res. 42*, D222.
(82) Myers, J. K., Pace, C. N., and Scholtz, J. M. (1997) A direct comparison of helix propensity in proteins and peptides. *Proc. Natl. Acad. Sci. U. S. A. 94*, 2833−2837.
(83) Blaber, M., Zhang, X.-J., and Matthews, B. W. (1993) Structural basis of amino acid α-helix propensity. *Science (Washington, DC, U. S.) 260*, 1637−1640.
(84) Horovitz, a. (1996) Double-mutant cycles: a powerful tool for analyzing protein structure and function. *Folding Des. 1*, R121−6.
(85) Lyu, P., Liff, M., Marky, L., and Kallenbach, N. (1990) Side chain contributions to the stability of alpha-helical structure in peptides. *Science (Washington, DC, U. S.) 250*, 669−673.
(86) Williams, R. W., Chang, A., Juretić, D., and Loughran, S. (1987) Secondary structure predictions and medium range interactions. *Biochim. Biophys. Acta, Protein Struct. Mol. Enzymol. 916*, 200−204.
(87) O'Neil, K. T., and DeGrado, W. F. (1990) A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science 250*, 646−51.
(88) Benatuil, L., Perez, J. M., Belk, J., and Hsieh, C.-M. (2010) An improved yeast transformation method for the generation of very large human antibody libraries. *Protein Eng., Des. Sel. 23*, 155−159.
(89) Boder, E. T., and Wittrup, K. D. (1997) Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol. 15*, 553−557.
(90) Ackerman, M., Levary, D., Tobon, G., Hackel, B., Orcutt, K. D., and Wittrup, K. D. (2009) Highly avid magnetic bead capture: an efficient selection method for de novo protein engineering utilizing yeast surface display. *Biotechnol. Prog. 25*, 774−783.
(91) Chao, G., Lau, W. L., Hackel, B. J., Sazinsky, S. L., Lippow, S. M., and Wittrup, K. D. (2006) Isolating and engineering human antibodies using yeast surface display. *Nat. Protoc. 1*, 755−768.
(92) Hackel, B. J., Kapila, A., and Wittrup, K. D. (2008) Picomolar affinity fibronectin domains engineered utilizing loop length diversity, recursive mutagenesis, and loop shuffling. *J. Mol. Biol. 381*, 1238−1252.
(93) Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G., and Neufeld, J. D. (2012) PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinf. 13*, 31.
(94) Woldring, D. R., Holec, P. V., and Hackel, B. J. (2016) ScaffoldSeq: Software for characterization of directed evolution populations. *Proteins: Struct., Funct., Genet. 84*, 869−874.
(95) Eigenbrot, C., Ultsch, M., Dubnovitsky, A., Abrahmsén, L., and Härd, T. (2010) Structural basis for high-affinity HER2 receptor binding by an engineered protein. *Proc. Natl. Acad. Sci. U. S. A. 107*, 15039−15044.
(96) Grönwall, C., Jonsson, A., Lindström, S., Gunneriusson, E., Ståhl, S., and Herne, N. (2007) Selection and characterization of Affibody ligands binding to Alzheimer amyloid beta peptides. *J. Biotechnol. 128*, 162−183.
(97) Mian, I. S., Bradwell, A. R., and Olson, A. J. (1991) Structure, function and properties of antibody binding sites. *J. Mol. Biol. 217*, 133−151.
(98) Birtalan, S., Fisher, R. D., and Sidhu, S. S. (2010) The functional capacity of the natural amino acids for molecular recognition. *Mol. BioSyst. 6*, 1186−1194.
(99) Searle, M. S., and Williams, D. H. (1992) The cost of conformational order: entropy changes in molecular associations. *J. Am. Chem. Soc. 114*, 10690−10697.
(100) Cole, C., and Warwicker, J. (2002) Side-chain conformational entropy at protein-protein interfaces. *Protein Sci. 11*, 2860−2870.
(101) Bloom, J. D., Wilke, C. O., Arnold, F. H., and Adami, C. (2004) Stability and the evolvability of function in a model protein. *Biophys. J. 86*, 2758−64.

DOI: 10.1021/acs.biochem.6b01142
Biochemistry 2017, 56, 1656−1671