Protein Science (1995), *4*:1108–1117. Cambridge University Press. Printed in the USA.

A phage display system for studying the sequence determinants of protein folding

HONGDI GU,[1] QIAN YI,[1] SUSAN T. BRAY,[1] DAVID S. RIDDLE,[1]
ANDREW K. SHIAU,[2] AND DAVID BAKER[1]

[1] Department of Biochemistry, SJ-70, University of Washington, Seattle, Washington 98195
[2] Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, California 94143

(RECEIVED March 27, 1995; ACCEPTED April 10, 1995)

Abstract

We have developed a phage display system that provides a means to select variants of the IgG binding domain of peptostreptococcal protein L that fold from large combinatorial libraries. The premise underlying the selection scheme is that binding of protein L to IgG requires that the protein be properly folded. Using a combination of molecular biological and biophysical methods, we show that this assumption is valid. First, the phage selection procedure strongly selects against a point mutation in protein L that disrupts folding but is not in the IgG binding interface. Second, variants recovered from a library in which the first third of protein L was randomized are properly folded. The degree of sequence variation in the selected population is striking: the variants have as many as nine substitutions in the 14 residues that were mutagenized. The approach provides a selection for "foldedness" that is potentially applicable to any small binding protein.

Keywords: combinatorial mutagenesis; phage display; protein folding; protein L

The complexity of the protein folding problem argues for the detailed study of the simplest possible experimental models. The difficulty of the problem derives in part from the very large number of interactions involved, and thus there is a strong incentive to focus on the smallest proteins. On the other hand, the minimum number of residues necessary for formation of a globular protein domain is set by the requirement for a buried hydrophobic core and is on the order of 50 amino acids. There are numerous relatively stable proteins of lesser size, but stability in these cases is almost invariably achieved through the incorporation of disulfide bonds or the binding of metals or other cofactors.

To understand how amino acid sequences determine protein three-dimensional structures, it is imperative to vary the sequence and examine the changes in both structure and folding pathway. Combinatorial mutagenesis of multiple sites can be a powerful approach to obtaining general insights into protein folding (Reidhaar-Olson & Sauer, 1988; Gregoret & Sauer, 1993). Such an approach requires a means to select out rare sequences that fold from a large highly mutagenized population.

Most proteins in the 50–70-amino acid size range are binding proteins rather than enzymes. The binding sites are generally formed by residues adjacent in the three-dimensional structure but not entirely contiguous in the amino acid sequence, and thus the binding activities are likely to require that the proteins be properly folded. This opens the door to a powerful and general approach to identifying the residues and interactions critical for specifying the folds of such small binding proteins: the residues directly involved in binding can be held constant and the remaining "scaffold" residues randomized or simplified using doped oligonucleotide-mediated mutagenesis. Affinity selection of variants that retain the binding activity then allows rapid recovery of viable scaffold sequences. Determination of the sequences of viable variants and further characterization is greatly facilitated if the randomized proteins are tethered to the randomized DNA sequences that encode them so the DNA is recovered with the protein.

The phage display technology (Scott & Smith, 1990; Smith & Scott, 1993) is ideally suited to this strategy. Phage display methods take advantage of the fact that two of the filamentous phage coat proteins have surface-exposed N-terminal domains that tolerate the insertion of foreign peptides and proteins. The great virtue of phage display systems is that phage displaying peptides or proteins able to bind a particular ligand can be rapidly isolated from vast libraries of phage expressing different variants on their surface. Because the genetic material encoding the displayed peptide or protein is contained within the phage particle, the nucleic acid sequences encoding the active variants can easily be determined after affinity selection of the phage from the library.

Reprint requests to: David Baker, Department of Biochemistry, SJ-70, University of Washington, Seattle, Washington 98195; e-mail: baker@ben.bchem.washington.edu.

We have chosen as a model system one of the shortest sequences known to adopt a stable native conformation without the complication of disulfide bonds: the 62-residue B1 immunoglobulin G (IgG) binding domain of the peptostreptococcal protein L (Kastern et al., 1992; protein L refers specifically to the IgG binding domain throughout the text). High-resolution NMR studies provide detailed structural information on the protein L native state (Wikstrom et al., 1993, 1994; see Fig. 1). Protein L is extremely stable and contains no prolines (which can complicate folding kinetics) (Kastern et al., 1992; Wikstrom et al., 1993, 1994). Mutant proteins that fold can readily be screened or recovered by exploiting the IgG binding property (Nilson et al., 1992). Finally, studies of two structurally related proteins having little or no sequence similarity to protein L—ubiquitin (Kraulis, 1991) and the IgG binding domain of protein G (Gronenborn et al., 1991; Achari et al., 1992)—provide interesting points of comparison.

Here we describe a phage display selection procedure that allows the retrieval of rare protein L variants that retain the ability to fold from large combinatorial libraries. We demonstrate that mutations that block folding are strongly selected against, and conversely, that variants that are recovered in the selection are properly folded.

Results

Both the M13 gene VIII and gene III proteins have been used successfully as fusion partners to display a wide variety of proteins and peptides. The fusion proteins are created by fusing a DNA fragment encoding the foreign peptide or protein to a coat protein gene either directly in a phage genome or in a plasmid having a phage origin of replication (Bass et al., 1990; Lowman et al., 1991). In the latter case, infection of bacteria bearing the plasmid with M13 helper phage leads to the packaging of the plasmid into phagemid particles, a subset of which display the fusion protein on their surfaces. We have chosen to fuse protein L to the gene VIII protein and to insert the fusion protein into a phagemid vector rather than into intact phage for several reasons. First, because there are more copies of the gene VIII protein (2,700 copies per phage) than the gene III protein (5 copies per phage), the number of fusion proteins per phage can be varied over a broad range without compromising infectivity (Markland et al., 1991). The number of copies of the fusion protein on the phage (and hence the stringency of the binding selection) can be controlled by titrating the expression level of the fusion protein prior to infection with helper phage. Second, the use of the phagemid system eliminates some of the problems associated with cloning in phage vectors (e.g., insert instability) and facilitates routine manipulations. Third, the phagemid system can also serve as a protein expression system, making possible an extremely useful secondary screen of colonies using nitrocellulose filter lifts (see Fig. 3).

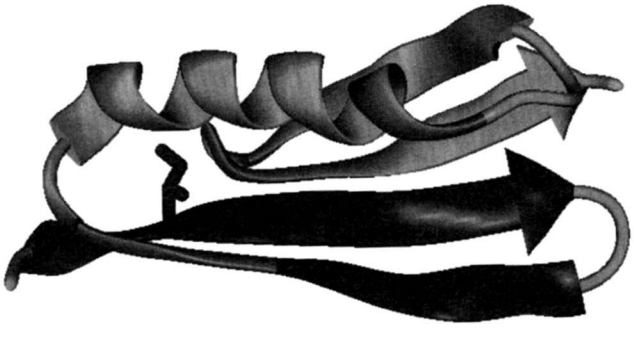


Fig. 1. Ribbon diagram of the protein L IgG binding domain. Shown are residues 1–62 of the IgG binding domain; these correspond to residues 94–155 of the intact protein L. Mutagenized residues are darkly shaded and the I4 side chain is displayed (the diagram was produced using the ribbonjr option of MIDASPLUS). The coordinates of the protein L IgG binding domain were generously provided by Drs. Lars Bjorck and Mats Wikstrom.

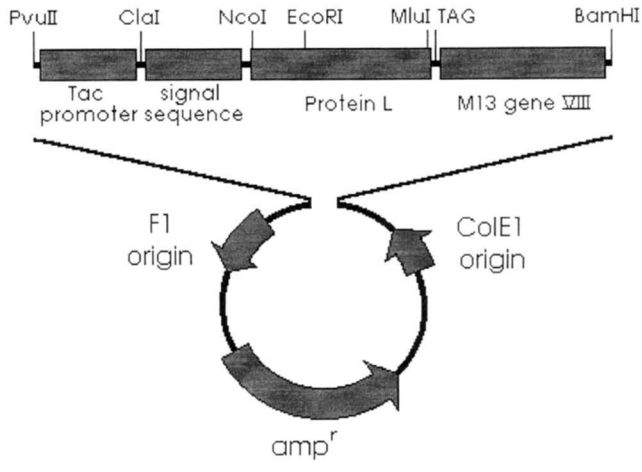


Fig. 2. Schematic of the protein L phage display construct. Vector portion is essentially identical to pBluescript II KS+ without the lacZ and polylinker containing portion (bases 500–977).

The salient features of the protein L phage display construct are shown in Figure 2. The Tac (*trp/lac*) promoter is followed by a ribosome binding site, the ompA signal sequence, the 62-residue protein L IgG binding domain, a suppressible stop codon (TAG), and the M13 geneVIII protein. The parent vector, a pBluescriptII derivative, contains a phage origin of replication and is readily packaged into phagemid particles (Short & Fernandez, 1988). The protein L phagemid construct directs the expression of a protein L–geneVIII fusion protein upon induction of supE cells with IPTG (data not shown).

An affinity matrix for protein L phagemid particles was prepared by coating streptavidin-coated magnetic beads with human IgG (protein L binds to the κ light chain). Incubation of protein L displaying phagemid particles with IgG-coated beads followed by extensive washing and elution at low pH resulted in recovery of approximately 0.1% of the input phagemids (Table 1, column 2). In contrast, control (pBluescript) phagemid particles not displaying protein L were recovered only at very low levels (<0.000004%) (Table 1, column 1).

As false positives are inevitably recovered in any affinity selection scheme, the ability to screen for true positives at the single colony level is extremely useful in distinguishing interesting mutants from experimental background. The colony lift assay described in the Materials and methods provides such a secondary screen. Nitrocellulose lifts of colonies expressing wild-type protein L react specifically with alkaline phosphatase-conjugated

Table 1. *Selectivity of biopanning procedure*[a]

	1	2	3	4		5	
	Bluescript	Protein L	I4D	Protein L	Bluescript	Protein L	I4D
In	5×10^8	5×10^8	5×10^8	10^5	10^{10}	10^5	10^{10}
Out	<20	5×10^5	<20	190	530	440	300
% Recovery	<0.000004	0.1	<0.000004	0.2	0.000005	0.4	0.000003

[a] The indicated amounts of protein L, Bluescript, or I4D phagemid particles were added either separately (columns 1–3) or in combination (columns 4–5) to IgG-coated magnetic beads and subjected to the biopanning procedure described in the Materials and methods. The results of the experiment described in column 5 are shown in more detail in Figure 4.

IgG, whereas cells transformed with pBluescript give essentially no background (Fig. 3).

To test the efficacy of biopanning, protein L-containing phagemid particles were mixed with a 100,000-fold excess of control (pBluescript) phagemid particles and subjected to the biopanning procedure. Dilutions of the input mixture and the low pH eluate were used to infect *Escherichia coli*. Colonies receiving the fusion protein construct were distinguished from colonies receiving pBluescript using the colony lift assay. A single panning step leads to ~100,000-fold enrichment for the protein L-containing phagemid particles (Table 1, column 4).

Our approach is based on the premise that binding of protein L to IgG requires that the protein be properly folded. To test this assumption, a protein L variant was constructed using PCR mutagenesis in which an isoleucine residue (I4) that forms part of the hydrophobic core of the protein (Fig. 1; 157 $Å^2$ are buried) was replaced by an aspartate. Although this residue is not part of the IgG binding interface (M. Wikstrom, pers. comm.) and thus should not affect binding directly, the placement of a negative charge in the hydrophobic core is very likely to significantly destabilize the protein and thus disrupt binding indirectly if proper folding is required for the binding activity.

Importantly, phagemids containing the I4D mutant of protein L were recovered from the biopanning procedure at only very low background levels (Table 1, column 3). Furthermore, the I4D mutant gave no signal in the colony blot assay (Fig. 3). Thus, the mutation blocks binding of IgG to protein L both on phage and on nitrocellulose, presumably by interfering with the proper folding of the protein (see Fig. 7).

To test the ability of the biopanning procedure to recover protein L variants that fold amidst a large excess of nonfolding mutants, 10^5 phage displaying wild-type protein L were mixed with 10^{10} phage displaying the I4D mutant and then subjected to the standard biopanning method. The rather striking results are shown in Figure 4 (see also Table 1, column 5). The selection procedure results in a 100,000-fold enrichment of wild-type phage over the I4D phage. These results are particularly impressive given that the proteins differ by only one amino acid.

A database of widely divergent sequences that adopt the protein L fold could be extremely useful for unraveling sequence–structure relationships. As illustrated in Figure 4, our methodology is capable of recovering a very small fraction of sequences that fold from a large excess of sequences that do not. The generation of a high diversity library requires a high mutagenesis rate.

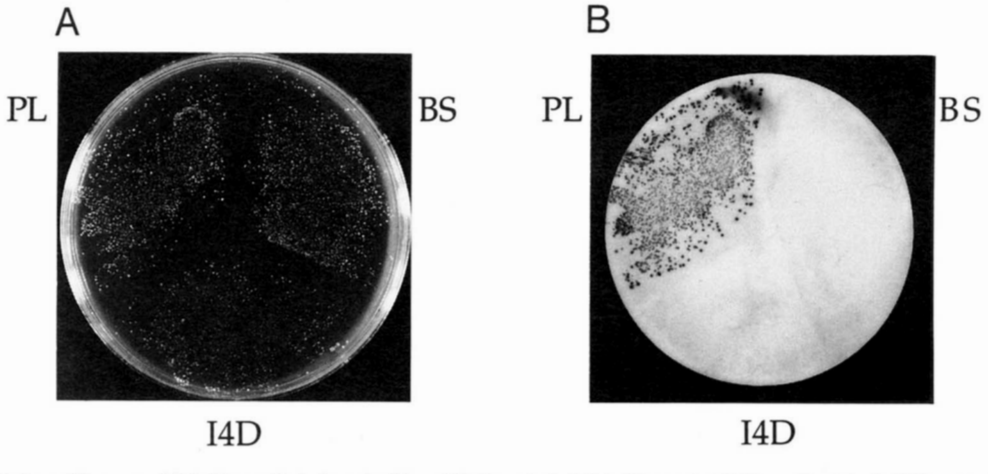


Fig. 3. Colony lift assay. XL1-Blue cells infected with ~200 Protein L (PL), Bluescript (BS), or Protein L I4D mutant phagemid particles were spread on an LB plate containing carbenicillin as indicated. After 12 h, colonies were transferred to nitrocellulose filters and treated as described in the Materials and methods. **A:** LB plate prior to colony transfer. **B:** Nitrocellulose lift after development with BCIP/NBT. The protein L-expressing colonies give rise to strong blue signals, whereas neither the Bluescript nor I4D colonies produce a detectable signal.

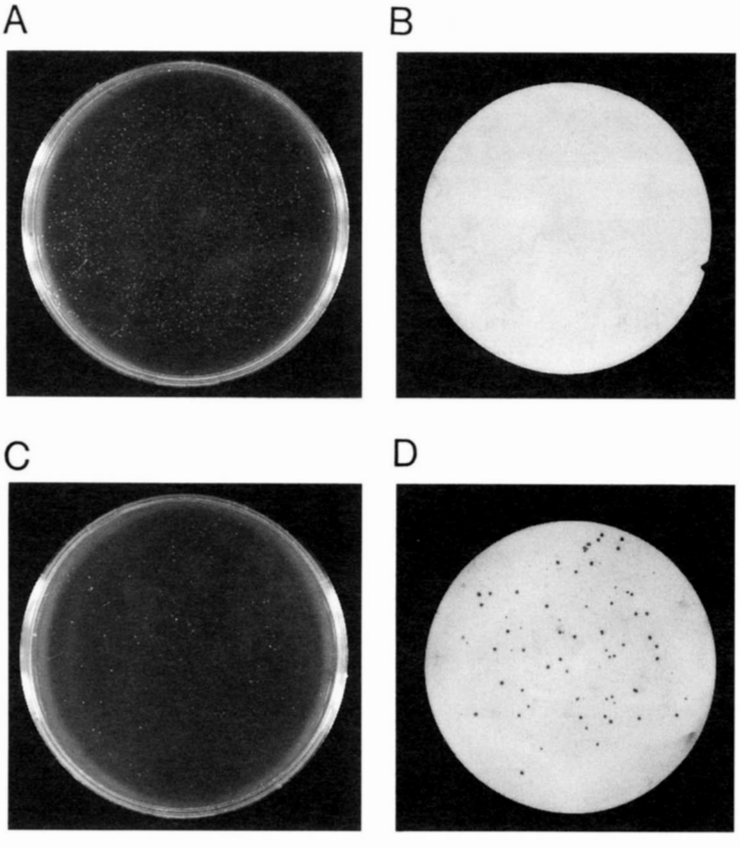


Fig. 4. Biopanning procedure can retrieve protein L sequences that fold from a 100,000-fold excess of nonfolding sequences. 10^{10} I4D phagemid particles were combined with 10^5 wild-type protein L phagemid particles and the mixture was subjected to the biopanning procedure described in the Materials and methods. ~1,000 phagemid particles from the original mixture (**A**, **B**) and ~100 phage particles eluted from the IgG-coated beads (**C**, **D**) were transfected into XL1-Blue cells and spread on LB-carbenicillin plates. Original plates with colonies (A, C) and nitrocellulose colony lifts prepared from the plates (B, D) are shown.

However, if too many residues are changed simultaneously, there will be very few folding sequences in the mutagenized population, and thus database size will be compromised. Clearly, there is a trade-off between database size and diversity.

How many residues can be changed simultaneously? There is little in the literature that allows prediction of the fraction of sequences that fold as a function of the mutagenesis rate. Evolution provides examples of very distantly related sequences that adopt the same fold, making it clear that few residues are absolutely essential to a fold. Similar conclusions have been reached in experimental studies (Heinz et al., 1992; Lim et al., 1992; Baldwin et al., 1993; Munson et al., 1994). However, it is not clear from these studies what volume of sequence space will give rise to a given fold; that is, are these alternative sequences that adopt the same fold extremely rare solutions or are they in fact fairly common. Perhaps the most relevant experimental result is the demonstration that the four-helix bundle fold apparently can tolerate a wide variety of extremely different sequences as long as the pattern of hydrophobic and hydrophilic residues is kept constant (Kamteker et al., 1993). If these results are general, it should be possible to identify very different sequences that adopt the protein L fold.

To investigate the power of the approach, we mutated 14 residues in the first two β-strands of protein L (Fig. 1; see Materials and methods). A phage library in which approximately 10^6 protein L variants were displayed was generated as described in the Materials and methods. The nucleotide frequencies in the unselected phage population were checked both by sequencing individual clones and by directly sequencing DNA prepared from ~1 μg of library DNA. The mutagenized positions had the expected distribution of the four bases (data not shown). Phage displaying protein L variants that retained the IgG binding activity were recovered by biopanning. Colony lift analysis showed that, although very few of the mutants in the original library bound IgG, essentially all mutants recovered in the biopanning procedure bound IgG (Fig. 5). Different colonies gave rise to signals of differing intensities (Fig. 5B,C); these were classified into three categories (Fig. 6A, column 3). Classifications were verified by streaking colonies on a fresh plate and repeating the nitrocellulose lift procedure (Fig. 5C).

Sequence analysis of 43 of the variants recovered in the panning procedure revealed widely different amino acid sequences (Fig. 6). The wild-type sequence was never recovered, as expected given the high level of mutagenesis. The mutants had up to nine residue substitutions.

Due to incomplete randomization of the mutagenized codons and the structure of the genetic code, different amino acid residues occurred with different frequencies even in the unselected phage population. With the exception of residues I4 and L8, the variation observed at each position in the subpopulation selected by the biopanning procedure was not substantially less than that in the unselected population. The conservation at positions 4 and 8 is likely due to the important role they play in the small hydrophobic core of the protein (157 and 152 Å^2 of buried surface, respectively). The degenerate codons at positions 8 and 10 encode L, F, C, or W. At position 8 only L and F were observed in the selected population, whereas at position 10, all four residues were observed (W relatively infrequently, however). The

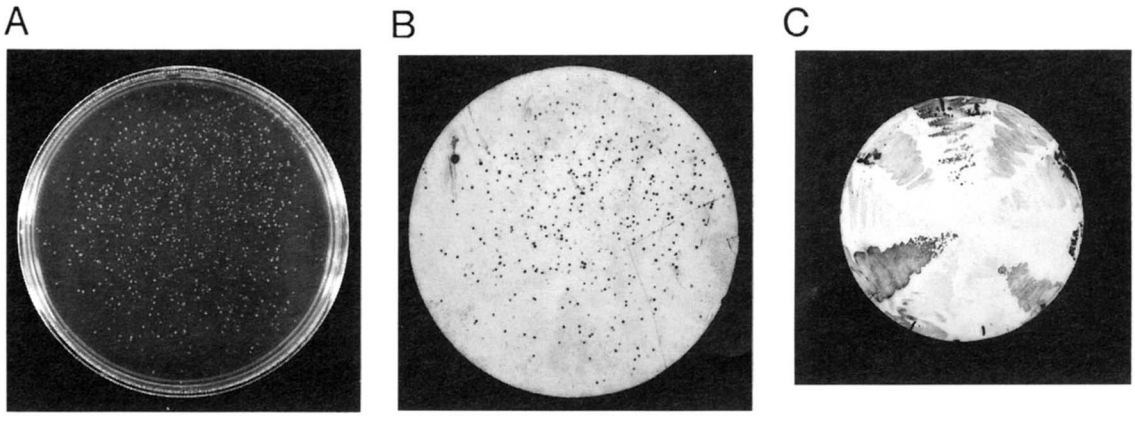


Fig. 5. Colony lift of phage library after biopanning. The phage library with the first third of the IgG binding domain randomized was subjected to the biopanning procedure as described in the text, and the eluted phage were infected into *E. coli*. **A:** Plate with transfected colonies. **B:** Nitrocellulose lift of plate shown in A. **C:** Individual colonies were streaked onto a fresh plate and subjected to the nitrocellulose lift procedure. Different mutants produced signals of varying intensity that could reflect differences in binding affinities or levels of protein accumulation.

lack of variation in E1, A11, N12, and G13 and the limited variation observed at S14 are consequences of the mutagenesis strategy and do not reflect structural constraints.

From colony lift analysis of large numbers of colonies resulting from transfection with the unselected phage population, we estimate that roughly 1 in 200 of the variants in the original library bind IgG. This selective pressure is not immediately evident in the wide range of IgG binding sequences shown in Figure 6A. That the selection for IgG binding does impose constraints, however, can be readily seen in a comparison of the frequencies of occurrence of different numbers of sequence changes in the selected and unselected phage populations (Fig. 6B). The distribution of sequence changes in 22 unselected phage variants (Fig. 6B, dashed line) was similar to that calculated directly from the nucleotide doping strategy (Fig. 6B, dotted line) but significantly shifted in the direction of more sequence changes compared to that of the 43 IgG binding variants (Fig. 6B, solid line). Comparison of the fraction of sequences that fold in the selected and unselected population provides a preliminary answer to the question raised earlier about the number of residues that can be changed simultaneously: the fraction of sequences that fold falls off dramatically as the number of changes increases from one to four, and then declines more slowly from four to seven (Fig. 6B, circles).

Are all of the IgG binding variants indeed properly folded? The expression, purification, and biophysical characterization of all of the mutants would be a formidable task. Instead, we chose to analyze six of the mutants having relatively large numbers (5–9) of substitutions and significantly different sequences. Every position that was mutagenized had at least one substitution in one of the six mutants, including mutations to proline in the first β-strand and to cysteine and threonine in the protein core. We note that the 62-residue IgG binding domain studied here is truncated with respect to the protein studied earlier (see Materials and methods), but the structure adopted by the 62-amino acid protein appears to be identical to that of the corresponding portion of the 76-residue protein based on similarity of 2D NMR spectra (data not shown).

These six variants, the wild-type protein, and the I4D mutant discussed above were subcloned into a T7-based expression vector with six N-terminal histidine residues to facilitate protein purification. All of the proteins were expressed at high levels in *E. coli*. I4D and mutant 20 were solubilized from inclusion bodies using 6 M guanidinium-HCl; all of the other proteins were soluble. Affinity chromatography on a zinc column yielded essentially pure protein as judged by SDS-PAGE. Although a number of the mutants contain cysteine residues, the mass spectra showed only monomeric species in each case.

CD spectroscopy was used to compare the secondary structure content of the variant proteins to that of wild-type protein L. CD spectra of the six IgG binding mutants were qualitatively similar to that of the wild-type protein and very different from that of the I4D mutant (Fig. 7). Interestingly, the IgG binding mutant with the most altered spectrum, mutant 17, was also the only variant in the set to give a weak signal in the colony lift assay (Fig. 6A). Thus, the strength of the colony lift signal (Fig. 5B,C) may reflect relatively subtle structural perturbations.

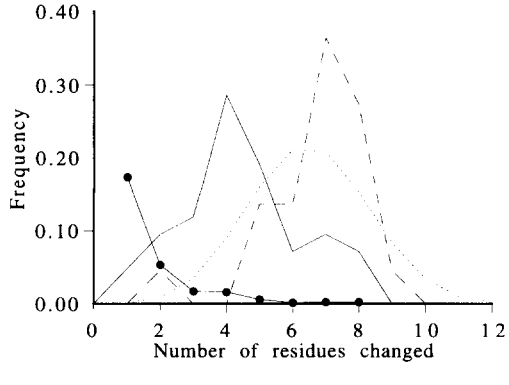
One-dimensional NMR spectra of the mutant proteins showed the sharp line shapes typical of folded proteins rather than the broad peaks characteristic of molten globule conformations (Fig. 8). The amide proton resonances of wild-type protein L and the mutants were well dispersed over 3 ppm, in comparison with the spectra of "random coil" polypeptide chains, which correspond closely to the sum of the resonances in the constituent amino acid residues (Wüthrich, 1986). This indicates that different backbone protons in the mutants as well as the wild-type protein experience distinct microenvironments that lead to conformation-dependent chemical shift dispersion. In addition, a number of well-resolved methyl resonances were visible in the aliphatic regions (1 to −0.3 ppm) shown in Figure 8. These upfield shifts are probably due to the local ring-current effect near aromatic residues. In summary, the 1D proton NMR spectra suggest that the mutants with multiple sequence changes have relatively well-defined tertiary structures.

Taken together with the fact that the binding site is intact, these data strongly suggest that the mutants adopt the native

A

ID	Sequence	Binding Activity
WT	EVTIKANLIFANGSTQTA	3
1	K	3
2	F N	2
3	G H	1
4	K C	1
5	H K	2
6	F G	3
7	VI A	2
8	T S C	2
9	G V F	1
10	F L T	3
11	I F T	2
12	IF K	3
13	H S C	3
14	T FVL	3
15	L G D A	1
16	E C RI	3
17	PKFQ	1
18	E L F A	2
19	T F I N	1
20	E T F C K	3
21	S T F C S	3
22	G F C SK	2
23	E T F C K	2
24	P CYF A	1
25	GS FLC	1
26	ES YF A	2
27	I VIT V	2
28	TIV F H	2
29	E VE F K R	3
30	E QFTC S	2
31	E FNC H T	2
32	KS F C IK	2
33	EI R K V V	1
34	E VT NC R	1
35	E Q VL CKR	2
36	EI E F C SG	1
37	QT C RHSK	2
38	K S FSC RK	1
39	A Q FVC I K	3
40	EL FSL C L E	3
41	P V FRW C H E	2
42	E NVH NL CSK	2
43	IP T FTC RAV	2

B



protein L fold. Importantly, the I4D mutant, which, unlike the above set of variants does not come through the selection (Table 1), is largely unfolded at room temperature (Fig. 7). Indeed, for all of the mutants we have characterized thus far, there is a perfect correlation between folding and recovery in the selection.

The stability of the wild-type and mutant proteins was compared by monitoring the ellipticity at 214 nm at increasing temperatures (Fig. 9). The wild-type protein has a T_m of 70 °C. In contrast, most of the mutants began to denature just above room temperature (Fig. 9; Table 2). Although molten globule conformations typically exhibit monotonic thermal denaturation profiles, five out of the six mutants had cooperative melting transitions with appreciable van't Hoff enthalpies of denaturation (Table 2).

The fact that many of the mutants begin to denature just above room temperature (Fig. 7) illustrates an important point: like any selection scheme, the phage selection only yields what is selected for, in this case the ability to fold at room temperature. There may also be constraints associated with assembly onto phagemid particles at 37 °C.

Conclusions

We have described a powerful selection for retrieving rare sequences that fold from large combinatorial libraries. The strategy is based on the assumption that the IgG binding activity of protein L requires proper folding. The following data support this premise. The I4D mutation disrupts both the folding of protein L and the binding of protein L phage to IgG-coated beads. Conversely, the six IgG binding variants that were studied in detail have CD spectra nearly superimposable on that of the wild-type protein, and one-dimensional NMR spectra characteristic of native proteins.

Although all of the sequences that come through the selection behave similarly at 25 °C, there are marked differences when the conditions are changed. Five of the six mutants analyzed in detail melt at significantly lower temperatures than the wild-type protein. Several of the mutant proteins have quite unusual properties. For example, mutant 40 undergoes only partial denaturation at high temperature: even at 90 °C there appears to be substantial secondary content. More detailed biophysical characterization of the mutant proteins is currently underway. Importantly for future studies, there appear to be few constraints associated with the production and purification of the mutant proteins; even the I4D mutant, which was completely denatured at room temperature, was expressed and purified with relatively high yield.

A related phage display system has been recently described by O'Neil et al. (1995) for the structurally related B1 domain of pro-

Fig. 6. Viable mutant sequences. **A:** Sequence changes in the mutants are given in the second column; the third column gives the strength of the signal (1 weak–3 strong) in the colony lift assay (see Fig. 5C). Only the mutagenized region (residues 1–18) is shown. **B:** Distribution of amino acid substitutions in the selected and unselected mutant populations. Solid line, selected population (43 sequences); dashed line, unselected sequences (22 sequences); dotted line, distribution expected given the doping strategy. The fraction of sequences that fold for a given number of substitutions ([fraction of sequences in selected population with n changes × size of selected population]/[fraction of sequences in unselected population with n changes × size of unselected population]) is shown in solid dots.

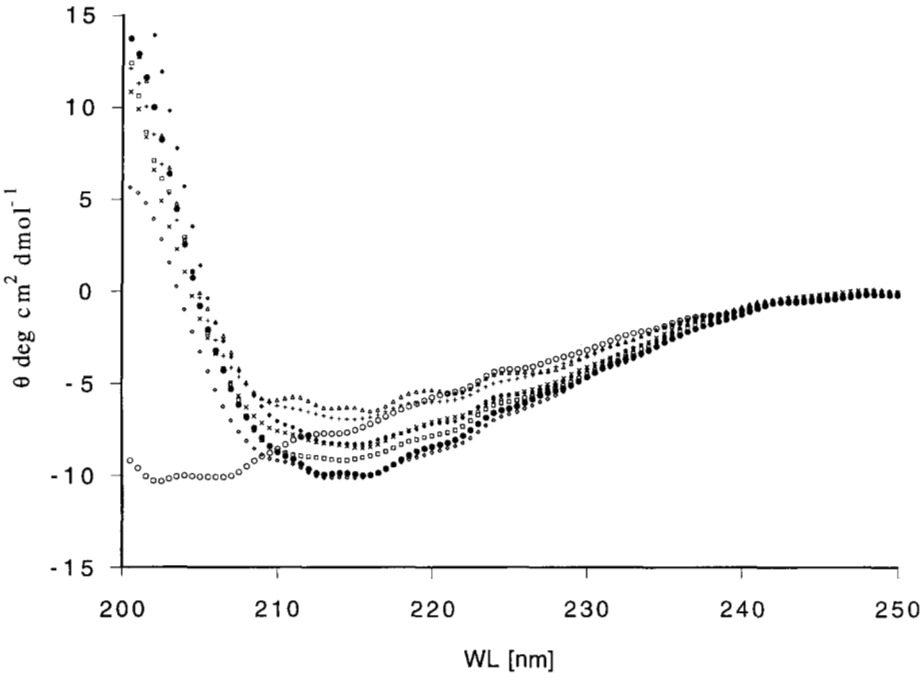


Fig. 7. CD spectra for: ●, protein L; □, mutant 40; ◆, mutant 43; +, mutant 20; ◇, mutant 17; △, mutant 30; X, mutant 37; and ○, I4D mutant at 22 °C. Data were collected from 200 to 250 nm at a scan speed of 100 nm per min; each spectrum is an average over 12 scans. Concentration of protein samples was 40 μM.

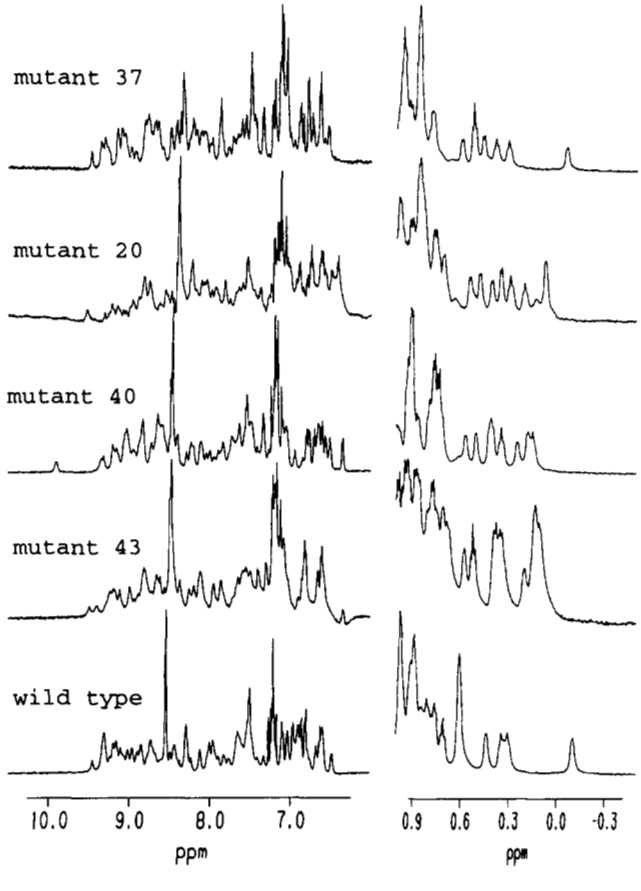


Fig. 8. Downfield (11.0 to 6.0 ppm) and upfield (1.0 to −0.5 ppm) regions of 1D proton NMR spectra of wild-type protein L and mutants in 90% H_2O/10% D_2O 50 mM potassium phosphate buffer at pH 6.0, 22 °C. Residual H_2O resonance (4.76 ppm) was used as a chemical shift reference.

tein G. Instead of using a phagemid system as described here, protein G was fused directly to the endogenous gene III protein of M13. Such recombinant phage had compromised levels of infectivity and thus were rapidly selected against during phage propagation. Mutations that partially destabilized the structure of protein G also reduced the perturbation to the phage life cycle. Thus, propagation of phage followed by selection for IgG binding provided a selection for destabilized mutants of protein G that retain the ability to fold. The absence of effects on phage production or infectivity in our experiments is likely due to the use of a phagemid system: the phagemid particles contain mainly wild-type coat proteins and only a very small number of the fusion proteins.

Although the mutagenesis experiment described here was carried out to test the selection scheme and is far too limited to extract general conclusions about the sequence determinants of the protein L fold, analysis of the results yields some interesting conclusions. First, the two positions that exhibited the greatest selectivity were I4 and L8. Both are highly buried positions. Interestingly, A6 and F10, which are also nearly completely buried, were considerably less well conserved. Second, residues in the binding site were not particularly more conserved (M. Wikstrom, pers. comm.). It may be that the residues in the second strand contact IgG mainly through backbone hydrogen bonds, and thus that the interactions are somewhat insensitive to the type of side chains at these positions. Third, the fraction of mutants that fold is an extremely high 1/200. Why is this so much higher than that extrapolated from single point mutations? The phage selection is probably more forgiving than most in vivo selections, which typically require retention of enzymatic activity or multiple protein-protein or DNA interaction surfaces. Here, the selection is simply for a scaffold that places residues at the binding interface at approximately the right position.

We are currently generating libraries in which all residues not involved in binding are heavily randomized. Application of the

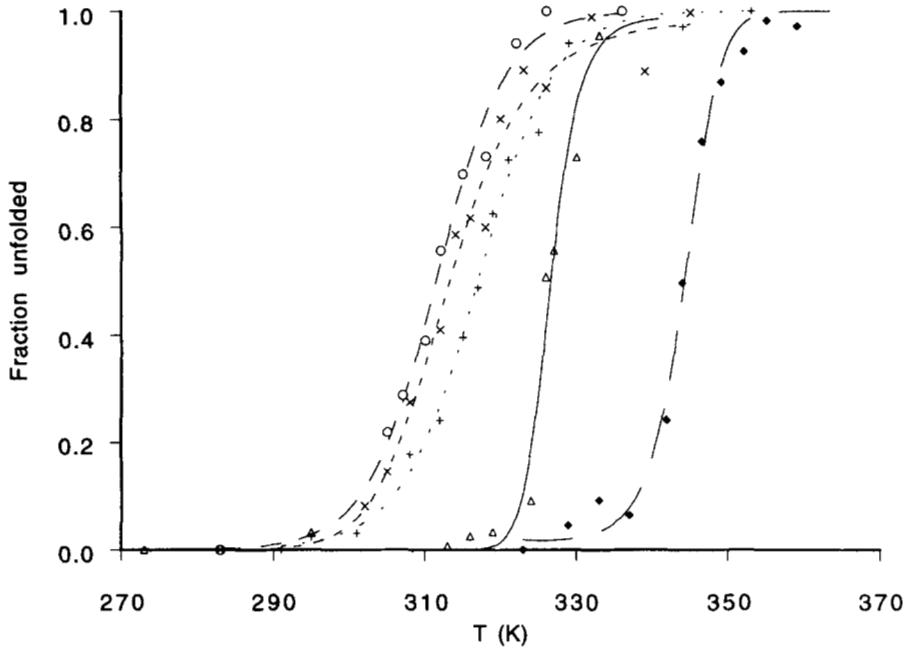


Fig. 9. Thermal denaturation profiles of: ◇, protein L; ○, mutant 30; +, mutant 20; X, mutant 17; and △, mutant 37. Data points were collected by monitoring ellipticity at 214 nm. Sample conditions were as described in Figure 7 legend.

combined phage selection and colony screen described in this paper should make possible the generation of an intensive database of information on the sequence determinants of the protein L fold. Characterization of the thermodynamics and kinetics of these mutant proteins should yield insights into the sequence determinants of the kinetics and thermodynamics of protein folding. Of particular interest will be the range of folding times in this "laboratory-generated" set of proteins not exposed to the constraints of natural selection.

Materials and methods

Media, chemicals, and enzymes

Antibiotics, media, and solutions used in molecular biological procedures were prepared as described (Sambrook et al., 1989). SDS, *N,N'*-methylene-bisacrylamide, and acrylamide were from Bio-Rad (Hercules, California). Electrophoresis grade agarose was from FMC BioProducts (Rockland, Maine). Restriction enzymes and T4 DNA ligase were from Life Technologies, Inc. (Gaithersburg, Maryland) and New England Biolabs (Beverly, Massachusetts). Deoxyribonuclease (DNase) and bovine serum albumin (BSA) were from Sigma.

Table 2. *Melting temperatures and van't Hoff enthalpies of denaturation of wild-type and mutant IgG binding domains*[a]

	T_m (°C)	ΔH_{vH} (kJ/mol)
Protein L	70	285
Mutant 17	40	167
Mutant 20	44	161
Mutant 30	38	182
Mutant 37	54	301
Mutant 43	39	154

[a] Thermodynamic parameters were extracted from plots of ln K versus $1/T$, which were linear for all of the proteins. Because we do not yet have accurate estimates of ΔC_p, we cannot extract folding free energies from the thermal denaturation data (the curve fits in Fig. 9 are for illustrative purposes only).

Plasmid construction

The protein L phage display construct contains a Tac (*trp/lac*) promoter followed by a ribosome binding site, the ompA signal sequence, the 62-residue B1 IgG binding domain of protein L, a suppressible stop codon (TAG), a short linker (MHGGNA), and the M13 geneVIII protein. The plasmid was constructed by digesting pSVK3 (Pharmacia) with *Pvu* II and *Bam*H I and inserting a synthetic polylinker (*Xho* I, *Cla* I, *Nco* I, *Mlu* I, *Bam*H I). The TAC promoter, the ompA signal sequence, the protein L IgG binding domain (residues 94–155; renumbered 1–62 for convenience), and M13 geneVIII were then inserted using standard methods (Fig. 2). Previous structural studies of the IgG binding domain have focused on residues 78–155, but in these studies residues 78–93 were found to be disordered. Because for biophysical analysis it is desirable to work with the shortest sequence possible, the N-terminal 15-residue disordered segment was omitted in the phage display construct. The DNA fragment encoding the IgG binding domain was constructed from double-stranded synthetic oligonucleotide cassettes. Restriction sites were included in the synthetic gene through translationally silent base changes (Fig. 2). All constructs were verified by DNA sequencing. Throughout the text, the IgG binding domain (residues 94–155) is referred to as protein L.

Random mutagenesis of protein L

Two oligonucleotides with sequences 5'-CAG GCG GCC ATG GAA gta aca atc aaa gct aac T*tt* atc T*tt* GCA AAT GGG T-3' and 5'-GT TCC TTT GAA TTC tgc agt ttg tgt ggA CCC ATT TGC-3' were purchased from Keystone Laboratories (Menlo Park, California). Lowercase letters indicate mixtures of 77.5%

of the indicated base and 7.5% of each of the other three bases. A mixture of 50% T and 50% G was used in the italicized positions. The 10 bases at the 3′ end of both oligonucleotides are complementary and thus each can prime second strand synthesis for the other. The oligonucleotides were purified by denaturing polyacrylamide gel electrophoresis, annealed, and the second strand was synthesized using a Klenow fragment (Reidhaar-Olson et al., 1991). The resulting double-stranded DNA was digested with *Nco* I and *Eco*R I, gel purified, and cloned into a similarly digested protein L phage display construct. For construction of the phage library, ten 40-μL portions of competent XL1-blue cells (Stratagene) were each electroporated with 5 μL of ligation product and combined in a total of 10 mL SOC (Sambrook et al., 1989). After 1 h at 37 °C, an aliquot was removed to estimate the size of the library and the rest of the culture was further diluted with 10 mL 2xYT (Sambrook et al., 1989) with 25 μg/mL carbenicillin and 10 μg/mL tetracycline, incubated an additional hour, and then supplemented with 100 mL of 2xYT with 50 μg/mL carbenicillin and 10 μg/mL tetracycline. After an additional hour, 10^{10} pfu M13K07 helper phage (Bio-Rad) were added, the culture was incubated for 1 h, and kanamycin was added to 10 μg/mL. The culture was incubated overnight, cells were removed by centrifugation, and the phage-containing supernatant was stored at −20 °C.

Phagemid DNA for DNA sequencing was prepared using standard methods except 0.5% glucose was added during growth. This greatly increased phagemid titers, presumably by suppressing basal expression of the somewhat toxic geneVIII fusion protein. DNA sequencing was carried out using standard Sequenase sequencing protocols from United States Biochemical.

Panning procedure

Preparation of affinity matrix

Streptavidin-coated magnetic beads (Dynabeads M-280, Dynal, Inc., Lake Success, New York) were coated with human IgG in a two-step process. First, the beads were incubated for 20 min at room temperature with 0.5 mg/mL biotinylated protein G (Pierce, Rockford, Illinois). Following several washes in Tween-TBS (0.5% Tween 20 [Sigma] in TBS, pH 7.0) with 0.1% BSA, the beads were incubated for 1 h at room temperature with 0.1 mg/mL human IgG (Pierce) and then washed five times with 1 mL Tween-TBS. Little success was achieved with direct coupling of biotinylated IgG to streptavidin-coated beads; the biotinylation reaction may interfere with the protein L binding site on the κ light chain.

Biopanning

Phage were incubated with rocking with 30 μg IgG-coated beads for 1 h at room temperature. The beads were then washed seven times with 1 mL Tween-TBS. Phage bound to the beads after washing were eluted with 200 μL 0.1 M glycine, 1 mg/mL BSA, pH 2.4. The eluate was neutralized by addition of Tris base. The eluted phage were transfected into XL1 Blue cells and plated on LB agar with carbenicillin (50 μg/mL).

Colony lift

Nitrocellulose filters (0.5-mm pore size; Schleicher and Schell, Keene, New Hampshire) were overlaid on LB plates containing freshly grown colonies. Filters were placed colony side up on LB agar with IPTG (1.3 μg/mL) for 2 h at 37 °C to induce expression of the fusion protein. After induction, the colonies were washed from the filters with deionized water, and the filters were blocked for 10 min in 3% BSA solution. Blocked filters were first incubated 1 h at room temperature in 5 μg/mL human whole molecule IgG (Pierce), then washed with Tween-TBS. Filters were then incubated 1 h at room temperature in 0.3 μg/mL goat anti-human IgG conjugated to alkaline phosphatase (Bio-Rad) and washed with Tween-TBS. The lifts were equilibrated in 0.1 M Na_2CO_3, 0.1 M $NaHCO_3$, pH 9.8, then developed in this same high-pH solution containing 1.5 mg/mL BCIP, 3.0 mg/mL NBT.

Protein expression and purification

To overexpress and purify proteins for biophysical analysis, protein L and selected mutants were cloned into a modified T7 expression vector based on pET14b (Novagen, Madison, Wisconsin) in which a T7 promoter is followed by six histidine codons, an *Nco* I site, an *Mlu* I site, and a TAA stop codon. Protein L and the mutants were cloned into the modified vector between the *Nco* I and *Mlu* I sites.

For protein expression, 1 L of LB medium was inoculated with 10 mL of an overnight culture of transformed BL21(DE3)pLysS *E. coli* (Novagen). The culture was grown at 37 °C to mid-log phase and induced with 0.5 mM IPTG. After 5 h of induction, cells were harvested by centrifugation and immediately resuspended in 40 mL of binding buffer (5 mM imidazole, 0.5 M NaCl, 20 mM Tris-HCl, pH 7.9) with protease inhibitors (benzamidine-HCl 2.5 mM, PMSF 1.0 mM) and DNase. The resuspended cells were frozen in dry ice and thawed at room temperature. The resulting cell extract was centrifuged at 12,000 × *g* for 30 min. The wild-type protein and several of the mutants were in the soluble fraction; in these cases, the extracts were loaded directly on the zinc column (see below). Mutant proteins that were primarily in the pellet fraction were loaded onto the zinc column after solubilization with 6 M guanidinium-HCl.

All proteins were purified on a metal-chelating column using chelating sepharose fast flow resin (Pharmacia Biotech) via the six N-terminal histidine residues as described by the manufacturer. The protein fractions eluted from the metal column were pooled, dialyzed extensively against H_2O, and lyophilized. The molecular weight and purity of all purified proteins were verified by SDS-PAGE and mass spectrometry. Protein concentration was determined using an extinction coefficient of 5,200 M^{-1} cm^{-1}, calculated using the method of Gill and von Hippel (1989).

CD analysis

CD measurements were acquired on a Jasco 720 spectropolarimeter using a 0.1-cm water-jacketed cuvette. Protein samples were prepared in 50 mM sodium phosphate buffer, pH 6.5.

NMR spectroscopy

All spectra were recorded on Bruker DMX 500MHz spectrometer at 22 °C with water suppression using a 3-9-19 pulse sequence with gradients (Sklenar et al., 1993). The protein concentration was approximately 1 mM in 90% H_2O/10% D_2O

50 mM potassium phosphate buffer at pH 6.0. For each spectrum, 8,192 complex points were collected within a 7,008.2-Hz spectrum width. The residual H_2O resonance was used as a chemical shift reference (4.76 ppm). Data were processed using the Felix 2.30 software (Biosym, San Diego, California).

Acknowledgments

We thank D. Kim, F. Edfelt, D. Krylov, J. Young, E. Johnson, and J. Cartwright for their valuable contributions to this work. We thank L. Bjorck, M. Wikstrom, W. DeGrado, and K. O'Neil for communicating results prior to publication; L. Bjorck for generous contributions of reagents and advice; and C. Manoil for *E. coli* strains and advice. This work was initiated in the laboratory of D. Agard. We thank R. Klevit for generous access to her NMR spectrometer; K. Walsh for use of his mass spectrometer; and R. Klevit, B. Jones, D. Teller, J. Foote, and members of the Baker laboratory for comments on the manuscript. This work was supported in part by grants from the NSF (STC cooperative agreement BIR-9214821) and the NIH (GM51888), and by young investigator awards to D.B. from the NSF and the Packard Foundation.

References

Achari A, Hale SP, Howard AJ, Clore GM, Gronenborn AM, Hardman KD, Whitlow M. 1992. 1.67 Å X-ray structure of the B2 immunoglobulin-binding domain of streptococcal protein G and comparison to the NMR structure of the B1 domain. *Biochemistry 31*:10449–10457.

Baldwin EP, Hajiseyedjavadi O, Baase WA, Matthews BW. 1993. The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme. *Science 262*:1715–1717.

Bass S, Greene R, Wells J. 1990. Hormone phage: An enrichment method for variant proteins with altered binding properties. *Proteins Struct Funct Genet 8*:309–314.

Gill SC, von Hippel PH. 1989. Calculation of protein extinction coefficients from amino acid sequence data. *Anal Biochem 182*:319–326.

Gregoret LM, Sauer RT. 1993. Additivity of mutant effects assessed by binomial mutagenesis. *Proc Natl Acad Sci USA 90*:4246–4250.

Gronenborn AM, Filpula DR, Essig NZ, Achari A, Whitlow M, Wingfield PT, Clore GM. 1991. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science 253*:657–661.

Heinz DW, Baase WA, Matthews BW. 1992. Folding and function of a T4 lysozyme containing 10 consecutive alanines illustrate the redundancy of information in an amino acid sequence. *Proc Natl Acad Sci USA 89*:3751–3755.

Kamteker S, Schiffer JM, Xiong H, Babik JM, Hecht MH. 1993. Protein design by binary patterning of polar and nonpolar amino acids. *Science 262*:1680–1685.

Kastern W, Sjobring U, Bjorck L. 1992. Structure of peptostreptococcal protein L and identification of a repeated immunoglobulin light chain-binding domain. *J Biol Chem 267*:12820–12825.

Kraulis PJ. 1991. Similarity of protein G and ubiquitin. *Science 254*:581–582.

Lim WA, Farruggio DC, Sauer RT. 1992. Structural and energetic consequences of disruptive mutations in a protein core. *Biochemistry 31*: 4324–4333.

Lowman HB, Bass SH, Simpson S, Wells JA. 1991. Selecting high affinity binding proteins by monovalent phage display. *Biochemistry 30*:10832–10838.

Markland W, Roberts BL, Saxena MJ, Guterman SK, Ladner RC. 1991. Design, construction and function of a multicopy display vector using fusions to the major coat protein of bacteriophage M13. *Gene 109*:13–19.

Munson M, O'Brien R, Sturtevant JM, Regan L. 1994. Redesigning the hydrophobic core of a four-helix-bundle protein. *Protein Sci 3*:2015–2022.

Nilson BHK, Solomon A, Bjorck L, Akerstrom B. 1992. Protein L binds to the κ light chain variable domain. *J Biol Chem 267*:2234–2239.

O'Neil KT, Hoess RH, Raleigh DP, DeGrado WF. 1995. Thermodynamic genetics of the folding of the B1 immunoglobulin-binding domain from streptococcal protein G. *Proteins Struct Funct Genet 21*:11–21.

Reidhaar-Olson JF, Bowie JU, Breyer RM, Hu JC, Knight KL, Lim WA, Mossing MC, Parsell DA, Shoemaker KR, Sauer RT. 1991. Random mutagenesis of proteins sequences using oligonucleotide cassettes. *Methods Enzymol 208*:564–586.

Reidhaar-Olson JF, Sauer RT. 1988. Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. *Science 241*:53–57.

Sambrook J, Fritsch EF, Maniatis T. 1989. *Molecular cloning: A laboratory manual, 2nd edition*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.

Scott JK, Smith GP. 1990. Searching for peptide ligands using an epitope library. *Science 249*:386–390.

Short JM, Fernandez JM. 1988. pBluescript: Gene mapping vectors. *Nucleic Acids Res 16*:7583–7600.

Sklenar V, Piotto M, Leppik R, Saudek V. 1993. Gradient-tailored suppression for H1-N15 HSQC experiments optimized to retain full sensitivity. *J Magn Reson A 102*:241–245.

Smith GP, Scott JK. 1993. Libraries of peptides and phage displayed on filamentous phage. *Methods Enzymol 217*:228–257.

Wikstrom M, Drakenberg T, Forsen S, Sjobring U, Bjorck L. 1994. Three-dimensional solution structure of an immunoglobulin light chain-binding domain of protein L. Comparison with the IgG-binding domains of protein G. *Biochemistry 33*:14011–14017.

Wikstrom M, Sjobring U, Kastern W, Bjorck L, Drakenberg T, Forsen S. 1993. Proton nuclear magnetic resonance sequential assignments and secondary structure of an immunoglobulin light chain-binding domain of protein L. *Biochemistry 32*:3381–3386.

Wüthrich K. 1986. *NMR of proteins and nucleic acids*. New York: John Wiley & Sons.