SEE COMMENTARY

A minimal sequence code for switching protein structure and function

Patrick A. Alexander, Yanan He, Yihong Chen, John Orban, and Philip N. Bryan[1]

Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, MD 20850

Edited by Adriaan Bax, National Institutes of Health, Bethesda, MD, and approved October 12, 2009 (received for review June 9, 2009)

We present here a structural and mechanistic description of how a protein changes its fold and function, mutation by mutation. Our approach was to create 2 proteins that (*i*) are stably folded into 2 different folds, (*ii*) have 2 different functions, and (*iii*) are very similar in sequence. In this simplified sequence space we explore the mutational path from one fold to another. We show that an IgG-binding, 4β+α fold can be transformed into an albumin-binding, 3-α fold via a mutational pathway in which neither function nor native structure is completely lost. The stabilities of all mutants along the pathway are evaluated, key high-resolution structures are determined by NMR, and an explanation of the switching mechanism is provided. We show that the conformational switch from 4β+α to 3-α structure can occur via a single amino acid substitution. On one side of the switch point, the 4β+α fold is >90% populated (pH 7.2, 20 °C). A single mutation switches the conformation to the 3-α fold, which is >90% populated (pH 7.2, 20 °C). We further show that a bifunctional protein exists at the switch point with affinity for both IgG and albumin.

evolution | NMR | protein design | protein folding

Protein molecules are capable of self-organizing into 3D topologies that create biologic functions. The fundamental principles of how the sequence of amino acids in a protein determines its structure remain poorly understood, however, despite its central importance to biology. The primary approach to the "folding problem" has been to determine a detailed structural and energetic description of the equilibrium between the native state and the random population of disordered, unfolded states. It is well known that the equilibrium between folded and unfolded can be radically shifted in either direction with a few mutations. There is accumulating evidence, however, that a few mutations sometimes can dramatically shift the equilibrium into new tertiary (and/or quaternary) structures (1, 2). Understanding the capacity of a protein to acquire a completely different structure as a result of minor mutagenic perturbation is central to understanding both protein folding in general and more specifically how new protein structures and functions evolve. Most natural proteins populate only the native state significantly, with $\Delta G_{unfolding} \geq 5$ kcal/mol. It is also generally assumed that many mutations are required to shift the equilibrium such that $\Delta G_{unfolding}$ for some alternative state is ≥ 5 kcal/mol. This assumption underpins most bioinformatics methods, in fact. Most mutations in a protein that increase its propensity toward an alternative fold destabilize the original fold. Thus, it seems intuitive that a pathway of single amino acid substitutions would result in a long series of mutants that would be unfolded before enough folding information accumulates to significantly populate an alternative fold. Both natural and engineered examples demonstrate, however, that the sequence space separating 2 proteins with different structures can be quite small (3–5). To understand this seemingly paradoxical situation, one needs to methodically examine the sequence space separating 2 stable folds. In concept this is simple. One begins with 2 stable proteins of similar size but different folds and mutates one to be more like the other until a switch in structure occurs. In practice this approach is not trivial, however. Any mutation in a protein will change the context of other amino acids. This is the essence of the folding problem. Our approach, therefore, was to create a simplified sequence space in which the mutational path from one fold to another can be explored and shifts in the equilibrium between the 2 folded states (and unfolded states) can be measured as a function of mutation.

Previously we and others have studied the structure, folding, and stability of 2 binding domains of *Streptococcus* protein G (6). Protein G contains 2 types of domains that bind to serum proteins in blood: the G_A domain of 45 structured amino acids that bind to human serum albumin (HSA) (7, 8), and the G_B domain of 56 structured amino acids that bind to the constant (Fc) region of IgG (9, 10). The natural versions of G_A and G_B domains share no significant sequence homology and have different folds, 3-α and 4β+α, respectively. From these studies we have been able to create high-identity versions of G_A and G_B, which have wild-type stabilities and binding function but which are 77% identical. These proteins are denoted G_A77 and G_B77. G_A77 binds to HSA with a K_d = 100 nM and has a $\Delta G_{unfolding}$ of 5 kcal/mol (20 °C, 0.1 M KPO_4, pH 7.2) (11). Amino acids 1–8 and 54–56 are disordered in G_A77. The remaining 45 aa are well ordered in a 3-α helix bundle (12). G_B77 binds to the constant (Fc) region of IgG with a K_d = 100 nM and has a $\Delta G_{unfolding}$ of 5 kcal/mol (20 °C, 0.1 M KPO_4, pH 7.2) (13, 14). All 56 aa of G_B77 are well ordered in a 4-stranded β-sheet with an α-helix connecting strands 2 and 3 (12). The proteins were engineered such that the IgG and HSA binding epitopes are encoded in both proteins. The IgG-binding epitope is functional in the 4β+α fold and latent in the 3-α fold, whereas the albumin-binding epitope is functional in the 3-α fold and latent in the 4β+α fold. This results in an experimental system in which unmasking the latent function is linked to a switch in conformation (Fig. 1). This work is described in refs. 3, 4, 11, 12, and 15. The fact that G_A77 and G_B77 are so close in mutational space greatly simplifies subsequent searches of the sequence space that separates them. The context problem is not eliminated but is reduced to a practicable level. This allowed a systematic exploration of the sequence space separating these 2 functional folds.

Results

The positions of nonidentity between the G_A77 and G_B77 proteins are shown in Fig. 1. Our approach was to explore binary sequence space (choice of either the G_A or G_B amino acid at positions of nonidentity). Obviously making 13 sequential substitutions in any order for the corresponding amino acid at a position of nonidentity will result in a conformational switch. Finding the path with the fewest unstructured intermediates required a systematic approach, however. We first analyzed all 13 single-site mutants in G_A77 and G_B77. We were able to produce and purify folded proteins for approximately half of these mutants. Proteins were purified using an affinity-cleavage tag system that we developed (16), essentially

Author contributions: P.A.A., Y.H., J.O., and P.N.B. designed research; P.A.A., Y.H., Y.C., J.O., and P.N.B. performed research; P.A.A., Y.H., J.O., and P.N.B. analyzed data; and P.N.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

See Commentary on page 21011.

[1]To whom correspondence should be addressed. E-mail: bryan@umbi.umd.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0906408106/DCSupplemental.

BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Downloaded from https://www.pnas.org by 71.188.61.216 on May 2, 2022 from IP address 71.188.61.216.

Petitioner Merck, Ex. 1041, p. 21149

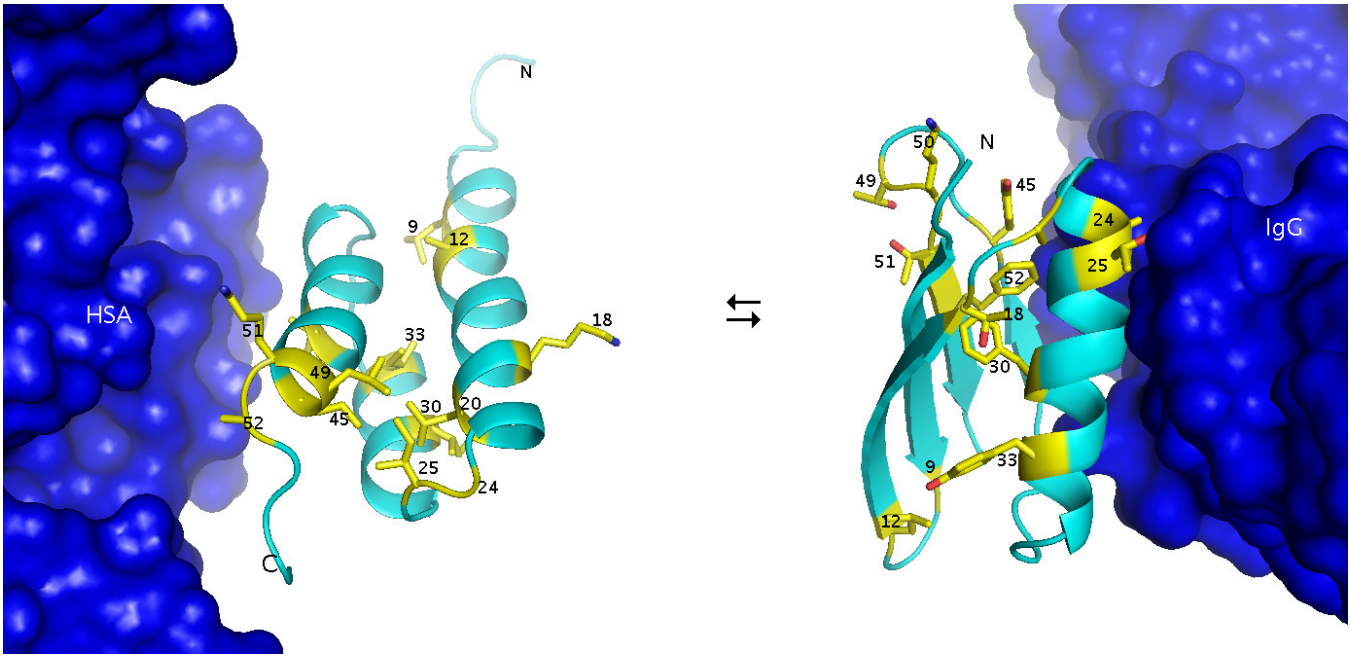

Fig. 1. G_A77 in complex with HSA (from 1TFO.pdb) (36) and G_B77 in complex in IgG (from 1fcc.pdb) (37). The side chains of amino acids at the 13 positions of nonidentity are depicted as yellow sticks.

as described in ref. 3. The system enabled the rapid, standardized purification of mutant proteins, even of low stability. Mutants were characterized by circular dichroism (CD) to assess secondary structure (Fig. S1), thermal denaturation by CD to assess stability (Fig. S2), the ability to bind HSA and IgG to assess function, and by 2D heteronuclear single quantum correlation (HSQC) spectra using NMR to assess tertiary structure. Monomeric state was established using size exclusion chromatography and multiple-angle laser-light scattering. High-resolution structures were determined by standard 3D NMR methods for key proteins. Midpoints of thermal denaturation (T_M) (0.1 M KPO4, pH 7.2, 20 °C) are reported for all folded mutants in Table 1. We found that every position in one fold or the other could be mutated without uncoding the native structure. Key mutants in the pathway to a conformational switch are shown in Fig. 2. A heteromorphic pair of 88% identity (G_A88 and G_B88b) was found in which stability and function remain similar to some naturally occurring IgG and HSA binding domains. Assembling heteromorphic pairs of greater than 88% identity was also possible, although additional mutation causes stability to fall below the threshold observed for most natural proteins. The pair G_A91 and G_B91 have a $\Delta G_{unfolding}$ of >3 kcal/mol at 20 °C. Both proteins show undiminished binding affinity to their respective ligands and are monomeric.

The binary mutational space separating G_A91 and G_B91 comprises only 32 sequences. We constructed, expressed, and purified 17 of these variants to effectively sample the sequence space (Tables 1 and S1). The main observations are as follows. The mutation Y33I has little effect on the stability of the $4\beta+\alpha$ fold (−0.1 kcal/mol), and the mutation L50K has the least effect on the stability of the 3-α fold (−1.0 kcal/mol). Mutation at position 20, 33, or 45 is not tolerated in the 3-α fold in any of the 32 contexts (Table S1). Mutation at position 30, 45, or 50 is not tolerated in the $4\beta+\alpha$ fold in any of the contexts. The only position that cannot be changed in either fold (without unfolding it) was position 45. Eight of the 17 proteins were predominantly folded into one of the native structures: four were 3-α and 4 were $4\beta+\alpha$. Of the 9 "unfolded" proteins, all were purified and can be seen by CD to have significant secondary structure content. The exact nature of this residual structure can probably be determined in the future using 3D NMR techniques.

The variants denoted G_A95 and G_B95 differ only at positions 20, 30, and 45, yet are fully folded, with $\Delta G_{unfolding}$ of ≈3 kcal/mol at 20 °C (Fig. S2). Both proteins show binding affinity to their respective ligands (K_D <1 μM) and are monomeric. High-resolution structures of these proteins were determined to better understand how so few amino acids control the conformational switch. This is discussed in detail below. Statistics for the G_B95 and G_A95 ensembles of 20 structures are shown in Table S2. Protein Data Bank accession codes for G_A95 and G_B95 are 2KDL and 2KDM, respectively.

Mutation of I30F in G_A95 (G_A98) and A20L in G_B95 (G_B98) leads to a heteromorphic pair differing at only 1 amino acid. G_A98 is folded into the 3-α conformation (>90% populated, pH 7.2, 20 °C), and G_B98 is folded into the $4\beta+\alpha$ conformation (>90% populated, pH 7.2, 20 °C). The CD spectra of G_A98 and G_B98 are essentially identical to the spectra of their parent G_A and G_B proteins (Fig. S1). The assigned HSQC spectra of G_A98 and G_B98 are compared in Fig. 3. G_A98 exhibits diminished affinity for HSA (K_D >1 μM) but has acquired affinity for IgG (K_D <1 μM). G_B98 binds tightly to IgG but not HSA. The ability of G_A98 to bind IgG as well as HSA may reveal a hidden propensity to switch into the $4\beta+\alpha$ conformation and unmask the IgG-binding epitope. The population of the $4\beta+\alpha$ conformation is too low to detect in the unliganded state of G_A98, but IgG binding may shift the equilibrium away from 3-α and toward $4\beta+\alpha$. It is also possible that G_A98 binds IgG via a new mechanism. This will be sorted out using NMR to map the binding epitope (17). In either case, exploration of binary sequence space shows that one functional protein can switch into a completely different conformation with a different function via a mutational pathway in which neither function nor native structure is completely lost. The G_A98 and G_B98 proteins are only marginally stable, but restoration of close to wild-type stability and function in either direction can be attained with only 3 additional mutations (e.g., to G_A88 and G_B88b).

Overall Change in Topology. High-resolution structures were determined for G_A95 and G_B95 and provide insight into the structural

Petitioner Merck, Ex. 1041, p. 21150

Downloaded from https://www.pnas.org by 71.188.61.216 on May 2, 2022 from IP address 71.188.61.216.

SEE COMMENTARY

Table 1. Purified mutants

T_M	9	12	18	20	24	25	30	33	45	49	50	51	52	
Mutants with G_A 3-α fold														
77.5°	L	A	K	L	G	I	I	I	L	I	L	K	A	G_A77
65.8°	L	A	K	L	A*	I	I	I	L	I	L	K	A	
67.3°	L	A	K	L	G	T*	I	I	Y	I	L	K	A	
62.5°	L	A	K	L	G	I	F*	I	L	I	L	K	A	
60.2°	L	A	K	L	G	I	I	I	Y*	I	L	K	A	
65.5°	L	A	K	L	G	I	I	I	L	I	K*	K	A	
71.6°	L	A	K	L	G	I	I	I	L	I	L	T*	A	
75.3°	L	A	K	L	G	I	I	I	L	I	L	K	F*	
69.4°	L	A	K	L	G	I	I	I	L	I	L	T*	F*	G_A88
61.5°	L	A	K	L	G	T*	I	I	L	I	L	T*	F*	G_A91
50.0°	L	A	K	L	G	T*	I	I	L	I	K*	T*	F*	G_A95
37.0°	L	A	K	L	G	T*	F*	I	L	I	K*	T*	F*	G_A98
Mutants with G_B 4β+α fold														
62.4°	G	L	T	A	A	T	F	Y	Y	T	K	T	F	G_B77
63.8°	L*	A*	T	A	A	T	F	Y	Y	T	K	T	F	
55.8°	G	L	K*	A	A	T	F	Y	Y	T	K	T	F	
49.9°	G	L	T	L*	A	T	F	Y	Y	T	K	T	F	
58.3°	G	L	T	A	G*	T	F	Y	Y	T	K	T	F	
58.0°	G	L	T	A	A	I*	F	Y	Y	T	K	T	F	
61.9°	G	L	T	A	A	T	F	I*	Y	T	K	T	F	
60.2°	G	L	T	A	A	T	F	Y	Y	I*	K	T	F	
57.5°	L*	A*	T	A	G*	T	F	Y	Y	I*	K	T	F	G_B88b
49.3°	L*	A*	K*	A	G*	T	F	Y	Y	I*	K	T	F	G_B91
48.7°	L*	A*	K*	A	G*	T	F	I*	Y	I*	K	T	F	G_B95
37.0°	L*	A*	K*	L*	G*	T	F	Y	Y	I*	K	T	F	
35.0°	L*	A*	K*	L*	G*	T	F	I*	Y	I*	K	T	F	G_B98

*Amino acid substitution.

determinants of the conformational switch. Amino acid substitutions at 20, 30, and 45 shift the equilibrium from >99% 4β+α to >99% 3-α. Amino acids 1–8 are disordered in G_A95 and form the central β1 strand in G_B95 (Fig. 4). Amino acids 9–23 form helix 1 in G_A95 and in G_B95 form the turn between β1 and β2, the β2 strand, and the turn between β2 and the central helix. Amino acids 27–33 are helical in G_A95. In G_B95 the helix is slightly longer: 24–37. Amino acids 39–51 form helix 3 in G_A95 and in G_B95 form the turn between the central helix and β3, the β3 strand, the turn between β3 and β4, and the first part of the β4 strand. Amino acids 52–56 are disordered in G_A95 and form the end of the β4 strand in G_B95. Except for common helical residues 27–33, every amino acid has undergone a change in secondary structure in the switch from 3-α to 4β+α.

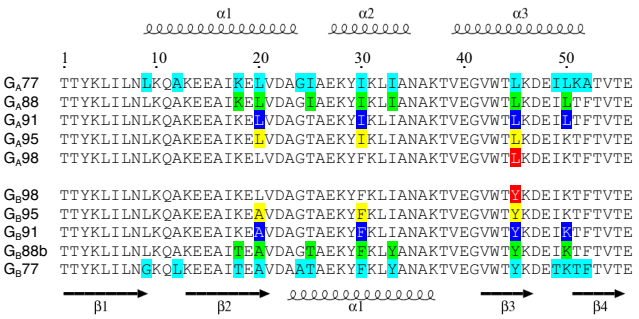

Fig. 2. Sequence alignment for 3-α G_A proteins (*Top*) and 4β+α G_B proteins (*Bottom*) described in the text. Secondary structure regions for G_A95 and G_B95 are shown at the top and bottom of the alignment, respectively. The 13 nonidentities between G_A77 and G_B77 are shown in cyan, the 7 nonidentities between G_A88 and G_B88 in green, the 4 nonidentities between G_A91 and G_B91 in blue, the 3 nonidentities between G_A95 and G_B95 yellow, and the single amino acid difference between G_A98 and G_B98 at residue 45 in red.

Hydrophobic Interactions. The hydrophobic core of G_A95 comprises A12, A16, and L20 from α1; I30, I33, and A36 from α2; and V42, K46, and I49 from α3 (Fig. 5). The aliphatic side chain of K46 contributes to packing of the hydrophobic core with the ammonium group protuding into solvent. A16, L20, I30, I33, and I49 form a tight network of mutually dependent hydrophobic contacts. L45 is 35% buried and actually not highly conserved among natural G_A domains. Its principal contacts are with boundary residues Y29, L32, and the core residue I33. The hydrophobic core of G_B95 comprises Y3, L5, and L7 from β1; A16 and A20 from β2; A26, F30, and A34 from the α-helix; W43 and Y45 from β3; and F52 and V54 from β4. The Y3, L5, F30, and F52 cluster accounts for ≈50% of the hydrophobic core. F30 interacts primarily with Y3 and L5 in β1 and F52 in β4. Y45 is packed against F52 and also forms a very strong H-bond with D47, which raises the pK_a of Y45 to >12.0 (18). The Y45, D47, and F52 interactions within the β3-β4 hairpin help give it the independent stability observed in previous studies of protein G mutants (19). Seven critical hydrophobic residues in the 3-α core (A12, A16, I33, A36, V42, K46, and I49) are preserved in the 4β+α fold of G_B95. Nine critical residues in the 4β+α fold (Y3, L5, L7, A16, A26, A34, W43, F52, and V54) are preserved in the 3-α fold of G_A95. Buried residues in one fold are frequently more solvent exposed in the other fold (Fig. S3).

The Critical Role of N- and C-Termini in Switching. One can consider the protein in 2 parts: amino acids 9–51, which are fully structured in both folds, and the 2 ends (1–8 and 52–56), which are unstructured in 3-α but form β strands in G_B95 (Fig. 6). In the switch from 3-α to 4β+α, hydrophobic interactions in the core of the helix bundle are released and replaced with mutually exclusive hydrophobic interactions between the central helix (A26, F30, and A34) and the central β-strands: β1 (Y3, L5, and L7) and β4 (F52 and V54). Thus there is a competition between formation of the helix bundle and recruitment of the N- and C-terminal residues to form

BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Downloaded from https://www.pnas.org by 71.188.61.216 on May 2, 2022 from IP address 71.188.61.216.

Petitioner Merck, Ex. 1041, p. 21151

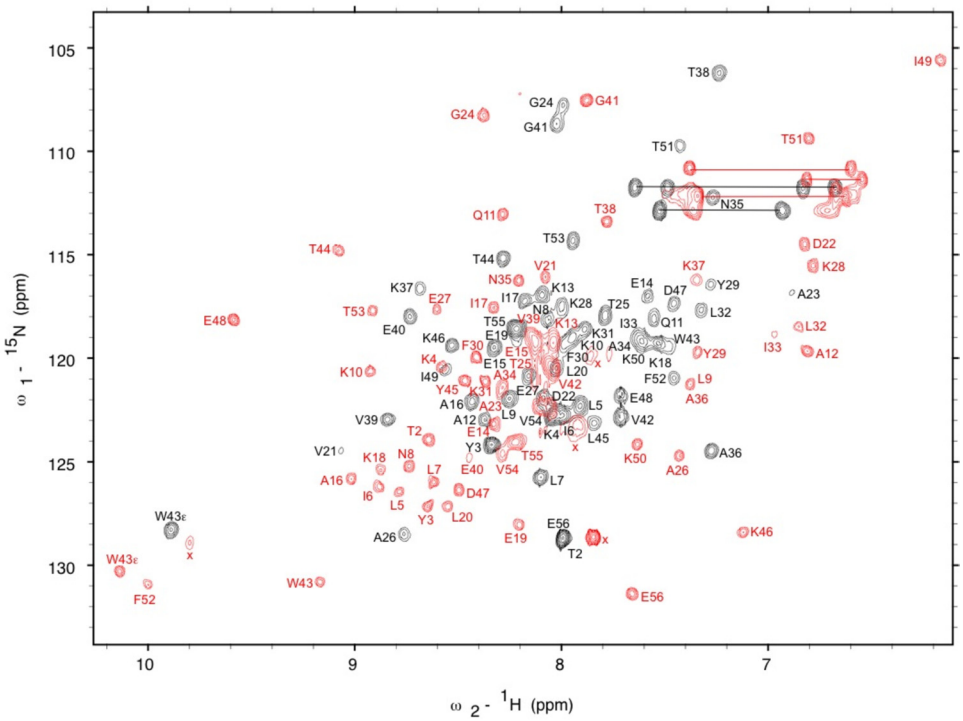

Fig. 3. ^{1}H, ^{15}N HSQC spectra. Main chain amide assignments are shown for G_A98 (black) and G_B98 (red). Of the 56 aa in these 2 proteins, 55 are identical but have different chemical environments reflecting the 2 different folds. Amide proton signals for side chains are connected by the horizontal lines.

a β-sheet with an alternative hydrophobic core. When 20 = L, 30 = I, and 45 = L, their preferred binding partners are in the hydrophobic core of the 3-α fold (e.g., A16, I33, and I49). When 20 = A, 30 = F, and 45 = Y, their preferred binding partners are 3, 5, and 7 in the β1 strand and 52 and 54 in the β4 strand. The binary sequence space between G_A95 and G_B95 is only 8 combinations. In this sequence space, L vs. Y at 45 is critical to which way the balance tips. When 45 = L, variants are unfolded if 20 = A. If 45 = L and 20 = L the protein will remain predominantly 3-α, even when I30 becomes F. Likewise, when 45 = Y, variants are unfolded if 30 = I. If 45 = Y and 30 = F, the protein will remain predominantly 4β+α, even when A20 becomes L. Thus the switch can be distilled down to 1 amino acid.

Thermodynamic Linkage. To fully understand the energetics of the switch it is critical to appreciate that (*i*) the 2 alternative folds are thermodynamically linked, and (*ii*) this linkage develops only as the context of the alternative fold develops. This is best illustrated by the variations at 45 and 52. L45/Y45 residues have ≈35% solvent accessibility in both folds. In G_A95, L45 is located in the α3-helix and has no direct interactions with F52. As noted above, Y45 is located in the β3-strand of G_B95 and is packed against F52. Neither L45Y nor A52F mutations are catastrophic in the 3-α fold (Table 1). In G_A77, the L45Y mutation costs 1.5 kcal/mol and the A52F mutation costs only 0.3 kcal/mol in $\Delta G_{unfolding}$. Either the Y45L or the F52A mutation strongly destabilizes the 4β+α fold, however. The ultimate tertiary structure is strongly influenced by the stability of the β3-β4 hairpin structure. When Y45 and F52

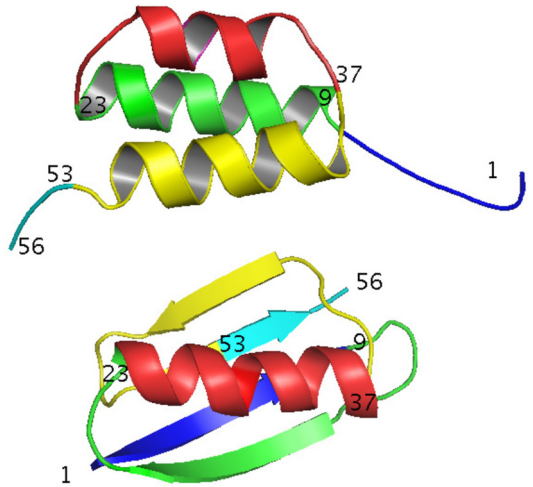

Fig. 4. Cartoon depiction of backbone topology of G_A95 and G_B95. Residues 1–8 are blue, 9–23 are green, 24–37 are red, 38–52 are yellow, and 53–56 are cyan.

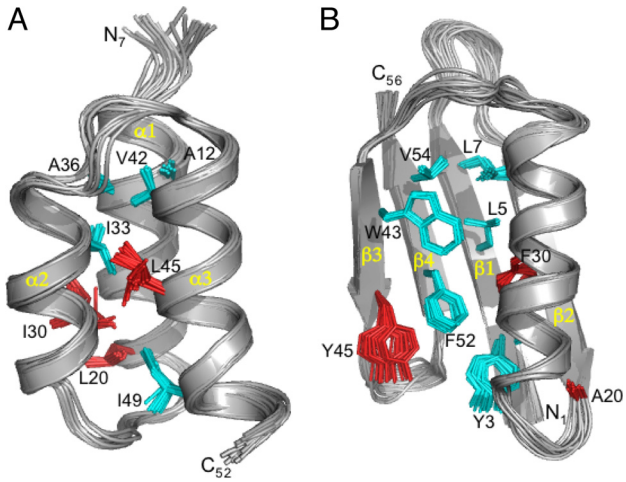

Fig. 5. Ensemble of 20 NMR structures for (*A*) G_A95 (residues 7–52) and (*B*) G_B95 (residues 1–56). Backbone RMSDs are ≈0.5 Å for both structures. Full structure statistics are in Table S2. Ordered core residues are shown in cyan, whereas the 3 amino acid differences between G_A95 and G_B95 are highlighted in red. Note that L20 is ordered in the hydrophobic core of G_A95, whereas I30 and L45 are more disordered and solvent accessible. Conversely, F30 and Y45 contribute significantly to stabilization of the core in G_B95.

Petitioner Merck, Ex. 1041, p. 21152

Downloaded from https://www.pnas.org by 71.188.61.216 on May 2, 2022 from IP address 71.188.61.216.

SEE COMMENTARY

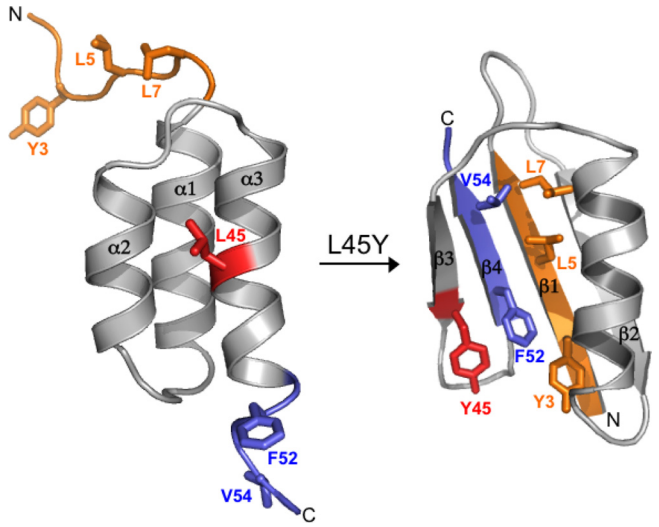

Fig. 6. Switching mechanism. Alternative conformations of the N-terminal (orange) and C-terminal (blue) amino acids in the 3-α and 4$\beta+\alpha$ folds. The critical switch amino acid occurs at position 45 (red). Also depicted are the hydrophobic packing of the N- and C-terminal amino acids in the core of the 4$\beta+\alpha$ fold.

occur together, the β3-β4 hairpin is favored. In terms of logical operators this is an AND gate. Other combinations at 45 and 52 favor 3-α. A switch can result from the additive effects of destabilization of the 3-α fold by Y45 (-1.5 kcal/mol) and stabilization of the 4$\beta+\alpha$ fold when Y45 and F52 interact ($+2$ kcal/mol).

The default propensity of the isolated 9–51 sequence in G_B98 is 3-α. The effects of short N- and C-terminal sequences can override the independent propensity. It has been demonstrated previously with short peptides that local propensities can be overcome by the larger protein context (20, 21), but here the propensity of an independently stable, 45-aa domain is overcome by 2 short sequences on its ends. The importance of the ends is not manifest until the context of the 4$\beta+\alpha$ fold develops, however. For example, the N- (1–8) and C- (52–56) termini of G_B were substituted early in the engineering of G_A. The changes are neutral to the stability of 3-α until the switch point is approached in mutational space. However, it should not be generally assumed that unstructured N- and C-termini are neutral to stability. In earlier work, we found that replacing 8 unstructured amino acids on the N terminus of another 3-α protein (staphylococcal protein A) with 8 aa from G_B resulted in a loss in stability of 3 kcal/mol (4).

The importance of context is also revealed in the Y vs. L choice at position 45. In G_A98, the L45Y mutation will switch conformation to 4$\beta+\alpha$. In G_A91 and G_A95, where insufficient context is present to support a full switch to 4$\beta+\alpha$, L45Y causes unfolding. The destabilization of 3-α may result from stabilization of a "nonnative" (β3-β4 hairpin) structure. This is consistent with a general phenomenon in protein folding. The changes in stability upon mutation result from interactions in both the native state and the unfolded state (e.g., ref. 22). In this case the "unfolded" state of 3-α probably includes the β3-β4 hairpin. In general, effects of mutation on the unfolded state are difficult to predict because specific structural elements in the unfolded state are unknown.

Bypassing the Unfolded State. The fact that sequential mutations can switch the conformational preference of a protein without populating the unfolded state seems at first paradoxical. Once dual propensities develop, however, a critical point is reached, and mutations tend to populate the alternative state rather than the unfolded state. This is an extreme case of another well-established phenomenon in folding. Denaturation with heat or a chaotropic agent such as urea or Gu-HCl generally produces a highly disordered unfolded state characterized by exposed hydrophobic groups and high conformational entropy (23). When mutation is the denaturant (instead of heat or chaotropic solvent), the "unfolded" state is often a limited ensemble of conformations characterized by high secondary structure content, although usually not a well-defined tertiary structure. The persistence of order in mutants of low stability is consistent with basic principles: (*i*) proteins collapse in water into compact structures; (*ii*) compactness drives acquisition of secondary structure; and (*iii*) final structural propensities are a subtle mix of many compensating weak forces. The important general point is that many destabilizing mutations will create propensities for new folds because certain nonnative states are strongly preferred over others. New propensity for a specific tertiary structure generally may not be discernible in a molten population, but even weak propensity may lead to short mutational pathways to stable, alternative topologies. Unfolded mutants such as L45Y in G_A91 and G_A95 populate neither the 3-α nor 4$\beta+\alpha$ folds detectibly but may contain certain structural elements of each. For example, β3-β4 hairpin structure may be forming but not β1-β2. This type of hybrid propensity may lead to short mutational paths to yet a third type of fold, were we to expand choices of amino acids and positions.

Discussion

We demonstrate that fold and function can be switched through a mutational pathway in which function is never lost and in which the unfolded state is never significantly populated. The switch from an IgG-binding, predominantly 4$\beta+\alpha$ fold into a bifunctional, predominantly 3-α fold can occur via a single amino acid substitution. The switch was achieved despite constraints common to protein engineering efforts: (*i*) defined end points for fold and function; and (*ii*) examination of a restricted (i.e., binary) sequence space. Nature does not have predetermined destinations in fold/function space or operate in restricted mutational space. Even so, numerous engineering efforts have demonstrated the potential of proteins to switch folds. This topic was reviewed in 2006 (2). There are also natural examples of fold switching. Classic examples of major conformational changes include influenza virus hemagglutinin upon interaction with the host cell (24), serpins upon proteolysis of a loop (25), and prion proteins that change from a mainly α-helical benign form (PrP^C) to an infectious state (PrP^{Sc}) with increased β-sheet content (26). Another recent example is the chemokine lymphotactin, which has 2 populated conformational states under physiologic conditions: the canonical chemokine fold (3-stranded β-sheet and C-terminal α-helix) in equilibrium with an all β-sheet dimer (27). Each conformation has a different binding function. In addition, natural Cro transcription factors exist with 40% identity but different structures for the 25 aa at their C-termini (28).

Most examples of natural and engineered fold switching involve a contextual change in the protein monomer. The most common contextual change is a change in quaternary structure. For example, mutation of several core residues in protein G can result in switching from a stable monomer to a stable tetramer with many of the native secondary structural elements (29). In the prion protein example, the driving force for the conformational switch from monomer to aggregate is extensive interaction of the edges of outer β-strands in the aggregated state. Metal binding has also been demonstrated to mediate major changes in secondary, tertiary, and quaternary structure in computationally designed proteins (30–32). Although our results show a switch from one monomeric fold into another, the energetic situation is similar in some ways to conformation switching driven by ligand binding. In our case the "binding" of the N- and C-termini to form the central strands of the β-sheet in the 4$\beta+\alpha$ fold is the primary energetic driving force for the switch.

Our work also shows that latent binding function can be linked to alternative folding propensity. In fact, our selection for high-

BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Downloaded from https://www.pnas.org by 71.188.61.216 on May 2, 2022 from IP address 71.188.61.216.

Petitioner Merck, Ex. 1041, p. 21153

identity alternative folds relied on migration from IgG-binding to albumin-binding function (3). Latent functionality likely exists in nature but is difficult to discern because an alternative function is revealed only in an unknown alternative fold. We would suggest that when latent function exists in an accessible alternative fold, functional migration is likely. We have demonstrated in our artificial system that there is no automatic penalty associated with evolving alternative propensity or latent function. We suggest that nature will explore sequence space when there is no penalty for doing so—that is, nature will follow any functional path. Small changes in amino acid sequence can cause a protein to veer toward propensity for another fold and on occasion switch folds, resulting in functional migration. The fact that fold and function can switch through short mutational pathways suggests that functional annotation based on sequence similarity should be regarded circumspectly. Even so, similar sequence does generally imply similar structure. If a fold switch occurs in nature, sequences rapidly diverge thereafter. Lymphotactin and the Cro transcription factors seem to be 2 natural proteins that have been "caught in the act," however (27, 28).

Fold switching seems to be more probable between some folds than others (33). Certain structural aspects of G_A and G_B make them amenable to switching. Many core residues in each fold are largely solvent exposed in the other fold, allowing significant folding information for both folds to coexist in a single protein. The degree to which hydrophobic core positions overlap in any 2 arbitrary folds of similar size will vary widely. This means that there may be some predictability as to what folds can come close in sequence space. Given the limited number of fold families, this observation may provide a coarse guide to how proteins migrate through fold space in the process of evolution.

Materials and Methods

Mutagenesis. Site-directed mutants were made using QuikChange (Stratagene) mutagenesis kits according to the manufacturer's instructions.

Protein Expression and Purification. G_A and G_B variants were cloned into the vector pG58 (16), which encodes an engineered subtilisin prosequence as an N-terminal fusion domain, and the resulting fusion proteins were purified using an affinity-cleavage tag system that we developed (16), essentially as described in ref. 3. A commercial version of the purification system is available through Bio-Rad Laboratories (Profinity eXact Purification System). Minimal medium (14) was used for ^{15}N and ^{13}C labeling. Soluble cell extract of prodomain (eXact tag) fusion protein was injected on a 5-mL S189 column at 5 mL/min to allow binding and then washed with 10 column volumes of 100 mM KPO_4, pH 7.2 to remove impurities (16). To cleave and elute the purified target protein, 6 mL of 10 mM sodium azide, 100 mM KPO_4, pH 7.2 was injected at 0.5 mL/min. The purified protein was then concentrated to 0.2 to 0.3 mM, as required for NMR analysis. All proteins are monomeric under these conditions. This was demonstrated by using gel filtration with an in-line multiple angle laser light scattering detector (miniDAWN TREOS; Wyatt Technology).

Binding to IgG and HSA. IgG and HSA were immobilized on GE HT1 columns containing NHS-activated agarose resin according to the manufacturer's instructions. Binding of mutants was carried out in 0.1 M KPO_4, pH 7.2 by injecting 1 mL of a 1-mg/mL solution at 0.5 mL/min. Washing was with 15 mL of 0.1M KPO_4, pH 7.2 at 1 mL/min. Elution was with 6 mL of 0.1 M phosphoric acid.

NMR Spectroscopy. ^{15}N- and $^{13}C/^{15}N$-labeled samples were prepared at 0.2–0.3-mM concentrations in 100 mM KPO_4 buffer (pH 7.2) containing 10 mM sodium azide and 5–10% D_2O. NMR spectra were collected on a Bruker AVANCE 600 MHz spectrometer fitted with a *z* axis gradient $^1H/^{13}C/^{15}N$ triple resonance cryoprobe. Spectra were recorded at 15 °C for G_A95 and 20 °C for G_B95. NmrPipe (34) was used for data processing, and analysis was done with Sparky (35). Standard triple resonance experiments were used to obtain NMR assignments. Main chain assignments were made from HNCACB, CBCA(CO)NH, HBHA(CO)NH, and HNCO spectra, whereas side chains were assigned with 3D (H)C(CO)NH-TOCSY, 3D H(CCO)NH-TOCSY, 2D CBHD, and 2D CBHE experiments. Interproton distance information was derived from 3D ^{15}N NOESY and aliphatic and aromatic 3D ^{13}C NOESY spectra with mixing times of 100 and 150 ms. For additional information see *SI Text*.

ACKNOWLEDGMENTS. We thank David Shortle and George Rose for helpful discussion; Yang Shen, David Baker, and Ad Bax for making results of computational analysis available before publication; and Jane Ladner for laser light scattering analysis. This work was supported by National Institutes of Health Grant GM62154.

1. Davidson AR (2008) A folding space odyssey. *Proc Natl Acad Sci USA* 105:2759–2760.
2. Ambroggio XI, Kuhlman B (2006) Design of protein conformational switches. *Curr Opin Struct Biol* 16:525–530.
3. Alexander PA, He Y, Chen Y, Orban J, Bryan PN (2007) The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proc Natl Acad Sci USA* 104:11963–11968.
4. Alexander PA, Rozak DA, Orban J, Bryan PN (2005) Directed evolution of highly homologous proteins with different folds by phage display: Implications for the protein folding code. *Biochemistry* 44:14045–14054.
5. Dalal S, Regan L (2000) Understanding the sequence determinants of conformational switching using protein design. *Protein Sci* 9:1651–1659.
6. Fahnestock SR, Alexander P, Nagle J, Filpula D (1986) Gene for an immunoglobulin-binding protein from a Group G Streptococcus. *J Bacteriol* 167:870–880.
7. Falkenberg C, Bjork L, Åkerstrom B (1992) Localization of the binding site for streptococcal Protein G on human serum albumin. Identification of a 55-kilodalton protein G binding albumin fragment. *Biochemistry* 31:1451–1457.
8. Frick IM, et al. (1992) Convergent evolution among immunoglobulin G-binding bacterial proteins. *Proc Natl Acad Sci USA* 89:8532–8536.
9. Myhre EB, Kronvall G (1977) Heterogeneity of nonimmune immunoglobulin Fc reactivity among gram-positive cocci: Description of three major types of receptors for human immunoglobulin G. *Infect Immun* 17:475–482.
10. Reis KJ, Ayoub EM, Boyle MDP (1984) Streptococcal Fc receptors. II. Comparison of the reactivity of a receptor from a group C streptococcus with staphylococcal protein A. *J Immunol* 132:3098–3102.
11. Rozak DA, et al. (2006) Using offset recombinant polymerase chain reaction to identify functional determinants in a common family of bacterial albumin binding domains. *Biochemistry* 45:3263–3271.
12. He Y, Chen Y, Alexander P, Bryan PN, Orban J (2008) NMR structures of two designed proteins with high sequence identity but different fold and function. *Proc Natl Acad Sci USA* 105:14412–14417.
13. Orban J, Alexander P, Bryan P (1994) Hydrogen-deuterium exchange in the free and immunoglobulin G-bound Protein G B-domain. *Biochemistry* 33:5702–5710.
14. Alexander P, Fahnestock S, Lee T, Orban J, Bryan P (1992) Thermodynamic analysis of the folding of the Streptococcal Protein G IgG-binding domains B1 and B2: Why small proteins tend to have high denaturation temperatures. *Biochemistry* 31:3597–3603.
15. He Y, et al. (2006) Structure, dynamics, and stability variation in bacterial albumin binding modules: Implications for species specificity. *Biochemistry* 45:10102–10109.
16. Ruan B, Fisher KE, Alexander PA, Doroshko V, Bryan PN (2004) Engineering subtilisin into a fluoride-triggered processing protease useful for one-step protein purification. *Biochemistry* 43:14539–14546.
17. He Y, Chen Y, Rozak DA, Bryan PN, Orban J (2007) An artificially evolved albumin binding module facilitates chemical shift epitope mapping of GA domain interactions with phylogenetically diverse albumins. *Protein Sci* 16:1490–1494.
18. Khare D, et al. (1997) pK_a measurements from nuclear magnetic resonance for the B1 and B2 immunoglobulin G-binding domains of protein G: Comparison with calculated values for nuclear magnetic resonance and x-ray structures. *Biochemistry* 36:3580–3589.
19. Sari N, Alexander P, Bryan PN, Orban J (2000) Structure and dynamics of an acid-denatured protein G mutant. *Biochemistry* 39:965–977.
20. Minor DL, Jr, Kim PS (1996) Context-dependent secondary structure formation of a designed protein sequence. *Nature* 380:730–734.
21. Anderson TA, Cordes MH, Sauer RT (2005) Sequence determinants of a conformational switch in a protein structure. *Proc Natl Acad Sci USA* 102:18344–18349.
22. Flanagan JM, Kataoka M, Fujisawa T, Engelman DM (1993) Mutations can cause large changes in the conformation of a denatured protein. *Biochemistry* 32:10359–10370.
23. Dill KA, Shortle D (1991) Denatured states of proteins. *Annu Rev Biochem* 60:795–825.
24. Skehel JJ, Wiley DC (2000) Receptor binding and membrane fusion in virus entry: The influenza hemagglutinin. *Annu Rev Biochem* 69:531–569.
25. Gooptu B, Lomas DA (2009) Conformational pathology of the serpins: Themes, variations, and therapeutic strategies. *Annu Rev Biochem* 78:147–176.
26. Harrison PM, Chan HS, Prusiner SB, Cohen FE (1999) Thermodynamics of model prions and its implications for the problem of prion protein folding. *J Mol Biol* 286:593–606.
27. Tuinstra RL, et al. (2008) Interconversion between two unrelated protein folds in the lymphotactin native state. *Proc Natl Acad Sci USA* 105:5057–5062.
28. Roessler CG, et al. (2008) Transitive homology-guided structural studies lead to discovery of Cro proteins with 40% sequence identity but different folds. *Proc Natl Acad Sci USA* 105:2343–2348.
29. Kirsten Frank M, Dyda F, Dobrodumov A, Gronenborn AM (2002) Core mutations switch monomeric protein GB1 into an intertwined tetramer. *Nat Struct Biol* 9:877–885.
30. Ambroggio XI, Kuhlman B (2006) Computational design of a single amino acid sequence that can switch between two distinct protein folds. *J Am Chem Soc* 128:1154–1161.
31. Cerasoli E, Sharpe BK, Woolfson DN (2005) ZiCo: A peptide designed to switch folded state upon binding zinc. *J Am Chem Soc* 127:15008–15009.
32. Hori Y, Sugiura Y (2004) Effects of Zn(II) binding and apoprotein structural stability on the conformation change of designed antennafinger proteins. *Biochemistry* 43:3068–3074.
33. Meyerguz L, Kleinberg J, Elber R (2007) The network of sequence flow between protein structures. *Proc Natl Acad Sci USA* 104:11627–11632.
34. Delaglio F, et al. (1995) NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* 6:277–293.
35. Goddard TD, Kneller DG (2004) *SPARKY 3* (University of California, San Francisco).
36. Lejon S, Frick IM, Bjorck L, Wikstrom M, Svensson S (2004) Crystal structure and biological implications of a bacterial albumin binding module in complex with human serum albumin. *J Biol Chem* 279:42924–42928.
37. Derrick JP, Wigley DB (1992) Crystal structure of a streptococcal protein G domain bound to an Fab fragment. *Nature* 359:752–754.

Petitioner Merck, Ex. 1041, p. 21154

Downloaded from https://www.pnas.org by 71.188.61.216 on May 2, 2022 from IP address 71.188.61.216.

Supporting Information

Alexander et al. 10.1073/pnas.0906408106

SI Text

Circular Dichroism. CD measurements were performed in 0.1 M 100 mM KPO_4, pH 7.2 with a Jasco spectropolarimeter (model J-720), using water-jacketed quartz cells with path lengths of 0.1 cm on protein concentrations of 10 μM. The ellipticity results were expressed as mean residue ellipticity, $[\theta]$, deg cm^2 $dmol^{-1}$. Temperature-induced unfolding was performed in the temperature range between 5 °C and 100 °C. Ellipticities at 222 nm were continuously monitored at a scanning rate of 1°/min. The fraction native is determined by subtracting unfolded baseline from the experimental CD signal and then dividing by the total CD difference between 100% folded and not at all folded at that temperature. Reversibility of the denaturations was confirmed by comparing the CD spectra at 25 °C before melting and after heating to 100 °C and cooling to 5 °C. The temperature unfolding profiles measured by far-UV CD for G_A and G_B were converted to an apparent $\Delta G_{unfolding}$ and fit to a theoretical curve calculated using the Gibbs-Helmholtz equation: $\Delta G_{unfolding} = \Delta H_o - T\Delta S_o + \Delta Cp(T - T_o - T\ln T/T_o)$, where $T_o = 298$°K and $\Delta Cp = 0.83$ kcal/°mol for G_B and 0.26 kcal/°mol for G_A (1, 2).

Structure Calculations and Analysis. CNS 1.1 (3) was used for structure calculation, starting with an extended polypeptide chain. Standard simulated annealing and torsion dynamics protocols were used and, prochiral groups were given floating assignments until they could be unambiguously assigned from the structure. Initial NOE restraints were generated automatically from NOEID, an in-house NOE assignment program. Subsequent assignments were obtained in a semiautomated mode using NOEID and intermediate structures to narrow down ambiguous assignments iteratively. Interproton distance restraints were based on peak intensities and categorized as strong (1.8–3.0 Å), medium (1.8–4.0 Å), weak (1.8–5.0 Å), and very weak (2.8–6.0 Å). Backbone dihedral restraints were obtained from chemical shift data using TALOS (4). Hydrogen bond restraints, 1.5–2.5 Å for $r_{HN\text{-}O}$ and 2.3–3.2 Å for $r_{N\text{-}O}$, were used only in the final stages of refinement. Final values for force constants were 1,000 kcal $mol^{-1}Å^{-2}$ for bond lengths, 500 kcal mol^{-1} rad^{-2} for angles and improper torsions, 40 kcal $mol^{-1}Å^{-2}$ for experimental distance restraints, 200 kcal mol^{-1} rad^{-2} for dihedral angle restraints, and 4.0 kcal $mol^{-1}Å^{-4}$ for the van der Waals repulsion term. The final ensemble of 20 structures was chosen using standard criteria including low total energy, no NOE distance violations more than 0.3 Å, no dihedral angle violations greater than 5°, and other measures of structure quality shown in Table S2. Structures were analyzed using PROCHECK-NMR (5), QUANTA (Molecular Simulations), MOLMOL (6), PyMol (7), and GETAREA (8).

1. Alexander P, Fahnestock S, Lee T, Orban J, Bryan P (1992) Thermodynamic analysis of the folding of the Streptococcal Protein G IgG-binding domains B1 and B2; Why small proteins tend to have high denaturation temperatues. *Biochemistry* 31:3597–3603.
2. Rozak DA, Orban J, Bryan PN (2005) G148-GA3: A streptococcal virulence module with atypical thermodynamics of folding optimally binds human serum albumin at physiological temperatures. *Biochim Biophys Acta* 1753:226–233.
3. Brunger AT, et al. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D (Biol Crystallogr) 54:905–921.*
4. Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR* 13:289–302.
5. Laskowski RA, Rullmann JA, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *J Biomol NMR* 8:477–486.
6. Koradi R, Billeter M, Wuthrich K (1996) MOLMOL: A program for display and analysis of macromolecular structures. *J Mol Graphics* 14:51–55.
7. Delano W (2002) *The PyMOL Molecular Graphics System* (DeLano Scientific, San Carlos, CA).
8. Fraczkiewicz R, Braun W (1998) Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J Comp Chem* 19:319–333.

Petitioner Merck, Ex. 1041, p. 21155

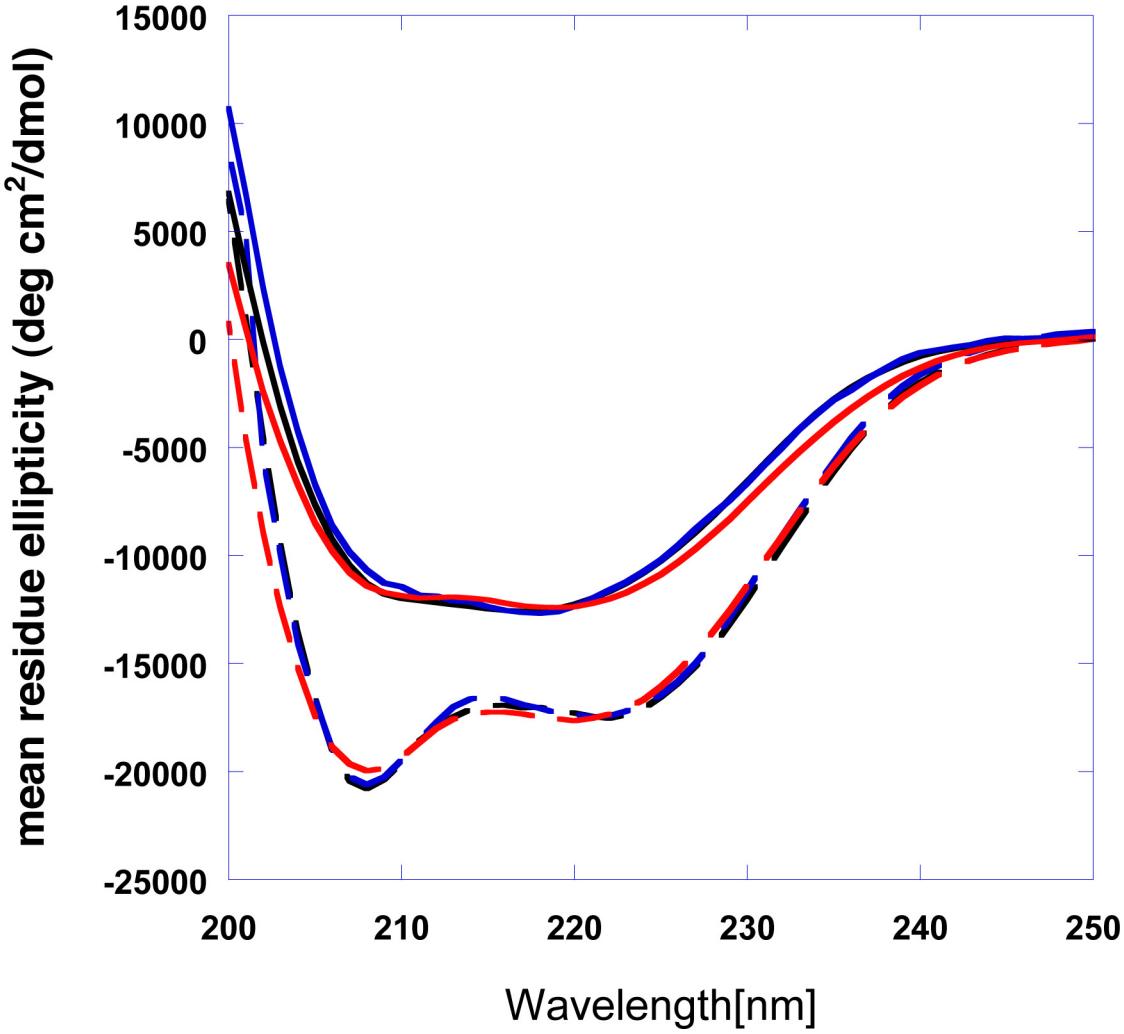

Fig. S1. Analysis of conformation by CD. CD spectra and thermal denaturation curves are shown for G_A (PSD-1; black dashed line), G_A77 (blue dashed line), G_A98 (red dashed line), G_{B1} (black solid line), G_B77 (blue solid line), and G_B98 (red solid line). Mean residue ellipticity (deg cm^2/dmol) is plotted vs. wavelength. Spectra were measured in 0.1 M KPO_4 buffer, pH 7.2 using a 0.1-cm cylindrical cuvette at 25 °C with [protein] = 10 μM.

Petitioner Merck, Ex. 1041, p. 21156

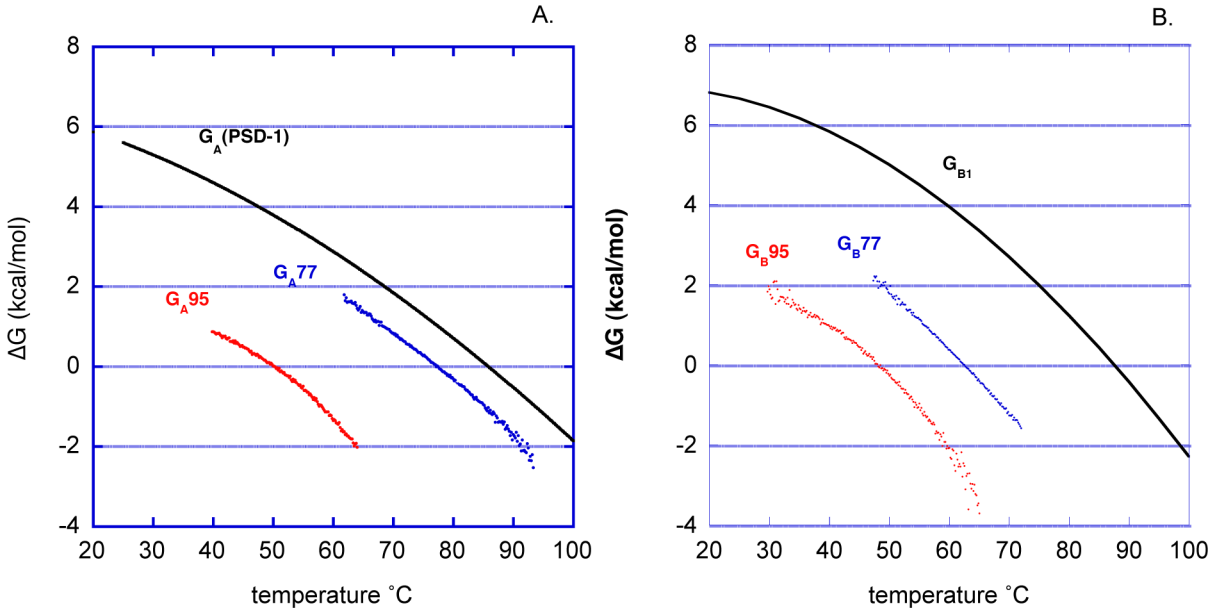

Fig. S2. Stability profiles of natural and mutant G_A and G_B proteins. (*A*) ΔG vs. temperature plots are shown for G_A77 (blue) and G_A95 (red). For reference the stability curve for the parent protein PSD-1 is shown in black. (*B*) ΔG vs. temperature plots are shown for G_B77 (blue) and G_B95 (red). For reference the stability curve for the parent protein G_{B1} is shown in black.

Petitioner Merck, Ex. 1041, p. 21157

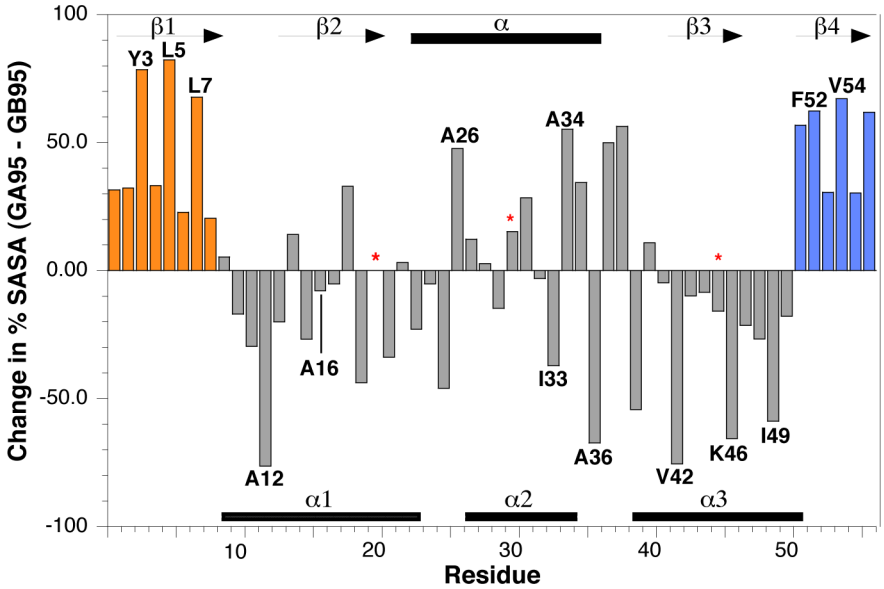

Fig. S3. Analysis of solvent accessible surface area (SASA) in G_A95 and G_B95. The histogram plot shows the change in percentage SASA for each residue ($SASA_{GA95}$–$SASA_{GB95}$). Central core residues in G_B95 are labeled in the top half of the plot, whereas amino acids in the interior of G_A95 are highlighted in the lower half. Thus residues buried in G_B95 tend to be solvent accessible in G_A95 and vice versa, illustrating the nonoverlapping nature of the respective hydrophobic cores. The 3 amino acid differences are shown with red asterisks. Residues in the N- and C-terminal regions are represented in orange and blue, respectively, and correspond to the similarly highlighted structures shown in Fig. 6. The secondary structures for G_B95 and G_A95 are shown at the top and bottom of the plot, respectively.

Petitioner Merck, Ex. 1041, p. 21158

Table S1. Mutants predominantly unfolded at 20 °C

	9	12	18	20	24	25	30	33	45	49	50	51	52
U	L	A	K	L	G	T*	I	Y*	L	I	L	T*	F*
U	L	A	K	A*	G	T*	I	I	L	I	L	T*	F*
U	L	A	K	A*	G	T*	F*	I	L	I	L	T*	F*
U	L	A	K	A*	G	T*	F*	Y*	L	I	K*	T*	F*
U	L*	A*	K*	L*	G*	T	F	Y	Y	I*	L*	T	F
U	L*	A*	K*	L*	G*	T	I*	I*	Y	I*	L*	T	F
U	L*	A*	K*	A	G*	T	I*	Y	Y	I*	K	T	F
U	L*	A*	K*	A	G*	T	F	I*	Y	I*	L*	T	F
U	L*	A*	K*	A	G*	T	F	Y	Y	I*	L*	T	F

Melting curves were measured in 0.1 M KPO_4, pH 7.2. U indicates that a mutant was predominantly unfolded at 20 °C.
*Amino acid substitution.

Petitioner Merck, Ex. 1041, p. 21159

Table S2. Statistics for the G_B95 and G_A95 ensembles of 20 structures

Parameter	G_B95	G_A95
Experimental restraints		
NOE restraints		
All NOEs	1,046	838
Intraresidue	637	554
Sequential ($\|i-j\| = 1$)	196	159
Medium-range ($1 < \|i-j\| \leq 5$)	58	75
Long-range ($\|i-j\| > 5$)	155	50
Hydrogen bond restraints	64	50
Dihedral angle restraints	74	72
Total restraints	1,184	960
RMSDs to the mean structure (Å)		
Overall residues*		
Backbone atoms	0.57 ± 0.12	0.48 ± 0.11
Heavy atoms	1.35 ± 0.18	1.35 ± 0.20
Secondary structures†		
Backbone atoms	0.43 ± 0.09	0.38 ± 0.09
Heavy atoms	1.10 ± 0.14	1.28 ± 0.16
Measures of structure quality		
Ramachandran distribution		
Most favored (%)	78.3 ± 3.5	77.6 ± 3.9
Additionally allowed (%)	17.0 ± 3.5	16.7 ± 3.2
Generously allowed (%)	3.7 ± 2.1	3.7 ± 2.8
Disallowed (%)	1.0 ± 1.2	1.9 ± 1.8
Bad contacts/100 residues	1.9 ± 1.0	2.0 ± 1.2
Overall dihedral G factor	0.01 ± 0.03	0.06 ± 0.03

*Residues 1–56 for G_B95. Residues 9–51 for G_A95.

†The secondary elements used were as follows: G_B95, residues 1–8, 13–20, 24–37, 42–46, and 51–55; G_A95, residues 9–23, 27–33, and 39–51.

Petitioner Merck, Ex. 1041, p. 21160