Coarse-grained sequences for protein folding and design

Scott Brown, Nicolas J. Fawzi, and Teresa Head-Gordon*

Department of Bioengineering, University of California, Berkeley, CA 94720

Edited by Peter G. Wolynes, University of California at San Diego, La Jolla, CA, and approved July 21, 2003 (received for review April 1, 2003)

We present the results of sequence design on our off-lattice minimalist model in which no specification of native-state tertiary contacts is needed. We start with a sequence that adopts a target topology and build on it through sequence mutation to produce new sequences that comprise distinct members within a target fold class. In this work, we use the α/β ubiquitin fold class and design two new sequences that, when characterized through folding simulations, reproduce the differences in folding mechanism seen experimentally for proteins L and G. The primary implication of this work is that patterning of hydrophobic and hydrophilic residues is the physical origin for the success of relative contact-order descriptions of folding, and that these physics-based potentials provide a predictive connection between free energy landscapes and amino acid sequence (the original protein folding problem). We present results of the sequence mapping from a 20- to the three-letter code for determining a sequence that folds into the WW domain topology to illustrate future extensions to protein design.

protein L | protein G | minimalist model | sequence design

An important insight into the protein folding problem is the recognition that native-state topology often plays a dominant role in the kinetics of the folding process (1, 2). This concept implies that the subtlety of interactions among 20 different amino acids that give rise to cooperative formation of native structure through backbone hydrogen bonding and specific side-chain packing of the native-state core can often be suppressed and effectively replaced by coarse-grained descriptions that capture the overall topology and spatial distribution of local and nonlocal contacts.

Minimalist proteins are coarse-grained models that use an α-carbon trace to represent the protein backbone in which structural details of the amino acids and aqueous solvent have been integrated out and replaced with effective bead–bead interactions. In these models, the potential energy functions for bead–bead interactions are typically Gō model potentials, which require direct knowledge of native-state tertiary contacts (3). These models are particularly useful in the study of proteins for cases when sequence is unimportant relative to the effects of native-state topology for determining folding rate and mechanism.

However Go bead models avoid the more difficult aspect of the protein folding problem, namely its dependence on amino acid sequence. Therefore, it is not surprising that these idealized models can lack a quantitative connection to experiment in some cases (4, 5). Physics-based models return to the original problem of confronting the complexity of amino acid sequence and corresponding interplay of physical interactions that give rise to the particulars of protein stability and kinetics and therefore do not require knowledge of native-state tertiary contacts (6–11). They are more generally applicable, especially when sequence details are equally important to topology, such as that found for two members of the ubiquitin fold class, the Ig-binding proteins L and G.

Proteins L and G are two single-domain proteins that have little sequence identity and yet identical fold topologies, consisting of a central α-helix packed against a four-strand β-sheet composed of two β-hairpins (12). Experimental evidence indicates that protein L folds in a two-state manner through a transition state involving a native-like β-hairpin 1 and largely disrupted β-hairpin 2 (13–15). Protein G, on the other hand, folds through a possible early intermediate (16, 17), followed by a rate-limiting step that involves formation of β-hairpin 2 (18).

The differences between proteins L and G illustrate the delicate balance between energetic and topological frustration in the folding of proteins and suggest that a model capable of distinguishing them would require a level of resolution that captures the differences in primary sequence. The question we explore here is: can a coarse-grained sequence description and design approach be used to capture the differences in folding mechanism between two proteins that adopt the same fold topology?

In this work, we present the results of sequence design on our off-lattice minimalist bead model that uses three residue types: hydrophobic, hydrophilic, and neutral. We start with a minimalist sequence that folds to the correct α/β fold topology and build on it through sequence mutation to produce two new minimalist sequences that reproduce the differences in folding mechanism seen experimentally for proteins L and G. Ultimately, a difference of only three bead types is required to differentiate between minimalist proteins L and G to exhibit the correct difference in their folding mechanism.

When the designed reduced letter code sequences are aligned alongside their respective counterpart sequences of the real protein L or protein G, we find ≈75% sequence identity. This alignment indicates that there is in actuality rather high sequence identity between proteins L and G at a coarse-grained sequence level, which is clearly minimal at the resolution of a 20-letter code (18). This degree of sequence identity shows that the patterning of hydrophobic and hydrophilic beads is largely preserved in the sequence design to fold to the target topology and suggests that it is the patterning that is the physical origin for the success of relative contact-order descriptions of folding kinetics.

We conclude that a minimalist three-letter code sequence is capable of discriminating between the primary structure for proteins L and G that is responsible for the two proteins folding through such distinctly different folding mechanisms. To demonstrate that the sequence mapping from a 20-letter letter amino acid code to the three-letter reduced code is sufficient for determining the folding to a target topology, we perform this sequence mapping for the WW domain and provide preliminary evidence that it folds to the experimentally observed β-sheet topology by using our physically motivated model.

Methods

Inspired by early efforts of Thirumalai and coworkers (6, 19, 20), our group has developed a more comprehensive minimalist model that is general to α-helical, β-sheet, and mixed α/β protein topologies. We have previously explored its use for members of the ubiquitin α/β fold class (8, 9, 21). The protein chain is modeled as a sequence of beads of three flavors:

This paper was submitted directly (Track II) to the PNAS office.

*To whom correspondence should be addressed. E-mail: tlhead-gordon@lbl.gov.

www.pnas.org/cgi/doi/10.1073/pnas.1931882100

Table 1. Sequence mapping between 20-letter (20) amino acid and coarse-grained three-letter (3) code

20	3	20	3	20	3	20	3
Ala	B	Met	B	Gly	N	Asn	L
Cys	B	Val	B	Ser	N	His	L
Leu	B	Trp	B	Thr	L	Gln	L
Ile	B	Tyr	B	Glu	L	Lys	L
Phe	B	Pro	N	Asp	L	Arg	L

hydrophilic, hydrophobic, and neutral; these are designated by L, B, and N, respectively. The total potential energy function is

$$H = \sum_{\theta} \frac{1}{2} k_{\theta}(\theta - \theta_0)^2 + \sum_{\phi}\left[A(1 + \cos\phi) + B(1 - \cos\phi) + C(1 + \cos 3\phi) + D\left(1 + \cos\left[\phi + \frac{\pi}{4}\right]\right)\right] + \sum_{i,j \geq i+3} 4\varepsilon_H S_1\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - S_2\left(\frac{\sigma}{r_{ij}}\right)^{6}\right]. \quad [1]$$

In the above equation, θ is the bond angle, ϕ is the dihedral angle, and r_{ij} is the distance between beads i and j. ε_H sets the energy scale and gives the strength of hydrophobic contact. The bond angle term is a stiff harmonic potential with force constant $k_\theta = 20\varepsilon_H/\text{rad}^2$, and $\theta_0 = 105°$. Each dihedral angle in the chain is designated to be one of the following three types: helical (designated by H), with $A = 0$, $B = C = D = 1.2\varepsilon_H$; extended (designated by E), with $A = 0.9\varepsilon_H$, $C = 1.2\varepsilon_H$, $B = D = 0$; or turn (designated by T), with $A = B = D = 0$, $C = 0.2\varepsilon_H$. The dihedral potentials are in large part responsible for maintaining the secondary structure topology. For both proteins L and G, the $\beta1$ region begins at bead 1 and terminates at bead 21, and the α helix covers the bead range from 23 to 35, whereas the $\beta2$ region covers the bead range from 37 to 56.

The nonlocal interactions are determined by: $S_1 = S_2 = 1$ for B–B interactions; $S_1 = 1/3$ and $S_2 = -1$ for L–L and L–B interactions; and $S_1 = 1$ and $S_2 = 0$ for all N–L, N–B, and N–N interactions. The attractive forces in the model responsible for collapse are due to the interactions between hydrophobic beads (B–B interactions). The interactions among all other combinations of beads are repulsive, although different strengths of repulsion are used depending on the bead types involved.

We use constant-temperature Langevin dynamics in the low-friction limit to perform simulations for characterizing the thermodynamics and kinetics of folding. Bond lengths are held rigid by using the RATTLE algorithm (22). All simulations are performed in reduced units, with mass m, length σ, energy ε_H, and k_B all set equal to unity.

The free energy landscape is characterized with the multidimensional histogram technique (23–25). We collect multiple six-dimensional histograms over energy E, radius of gyration R_g, and native-state similarity parameters χ, $\chi\alpha$, $\chi\beta_1$, and $\chi\beta_2$. χ is given by

$$\chi = \frac{1}{M} \sum_{i,j \geq i+4}^{N} h(\varepsilon - |r_{ij} - r_{ij}^{\text{native}}|); \quad [2]$$

here the double sum is over beads on the chain, and r_{ij} and r_{ij}^{native} are the distances between beads i and j in the state of interest and the native state, respectively. h is the Heaviside step function, with $\varepsilon = 0.2$ to account for thermal fluctuations away from the native-state structure. M is a constant that satisfies the conditions that $\chi = 1$ when the chain is identical to the native state and $\chi \approx 0$ in the random coil state. The remaining χ parameters are specific to their respective elements of secondary structure. That is, $\chi\alpha$ involves summation over beads in the helix, and $\chi\beta_1$ and $\chi\beta_2$ involve summation over beads in the first and second β-sheet regions, respectively.

From the histogram method, we get the density of states as a function of six-order parameters, $\Omega(V, R_g, \chi, \chi\alpha, \chi\beta_1, \chi\beta_2)$, which can be used to calculate thermodynamic quantities. In constructing the free energy surfaces, we collect histograms at 15 different temperatures: 1.20, 0.90, 0.70, 0.62, 0.60, 0.55, 0.50, 0.48, 0.46, 0.44, 0.42, 0.41, 0.40, 0.39, and 0.38. We run three independent trajectories at each temperature and collect 10,000 data points per trajectory.

The kinetics of the folding process can be characterized by calculating a large number of first-passage times (the time required for a folding trajectory to cross into the native basin of attraction). The first-passage times are calculated by taking an initial high-temperature random coil structure and evolving it at the folding temperature of interest until recording the time that it enters the native basin of attraction. We subtract off an initial correlation time in which the high-temperature chain is briefly equilibrated at the target temperature (this is the computational dead time during the kinetics run).

On the basis of theoretical work (2, 26, 27), one criterion for the foldability of heteropolymer sequences is the requirement of having a significant energy gap between the native-state and average misfold energies. Our sequence design strategy makes use of this concept by creating a misfold library generated from multiple trajectories at temperatures near the collapse temperature. To then obtain optimal new sequences, we attempt to maximize the energy gap, $\Delta E_{\text{design}} = \langle E_{\text{misfold}}\rangle - E_{\text{native}}$. The library is quenched before use, ensuring that member structures are at local minima on the potential energy surface. The structures collected in this way are representative of misfold traps and not the barriers separating minima. Maximizing ΔE_{design} for misfolds located at barriers could be expected to adversely affect folding kinetics, because trapping could become more substantial.

Table 2. Sequence alignment for the B1 domain of protein L (Protein Data Bank ID code 2PTL) (first line) against reduced code of minimalist model for protein L/G (second line)

	Protein L mapping
1°2PTL	VTIKANLIFANGSTQTAEFKGTFEKATSEAYAYADTLKKDNGEYTVDVADKGYTLNIKFAG
1°model L/G	LBLBLB**L**BBNNNL **B**BLB**BBB**BNN NLLBL**LB**BLLBN**B** L**B**LBLB**B** **N**NNBBBLBLB**L**BL
1°model L	LBLBLBLBBNNN**B** BBLB**L**BBBNN NLLBLLBBLLBNB LBLBLB**L** NNN**L**BBLBLB**B**BL
2°2PTL	CCCEEEEECSSSCCEEEECCBSSHHHHHHHHHHHHHTCSSSCCEEECCBTTTTEECEEECC
2°model L	EEEEEETEHTHE EEEEEEEHHE HHHHHHHHHHEHT EEEEEEE TTTEEEEEEEE

Possible error candidates in mapping are shown in bold. The third line shows the new sequence for protein L, with differences from L/G shown in bold. Also shown is the secondary structure alignment for the B1 domain of protein L (fourth line) against reduced code of minimalist model for protein L (last line).

Table 3. Sequence alignment for the B1 domain of protein G (Protein Data Bank ID code 2GB1) (first line) against reduced code of minimalist model for protein L/G (second line)

	Protein G mapping
1°2GB1	MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE
1°model L/G	LBLB**L**BL**BB**N**N**NL**B**BLB**B**BB**B**NNN LLB**L**L**B**B**L**L**BNBLB**L**BL**BB NNN**BB**BLBL**B**LBL
1°model G	LBLBLBLBBNNNLBBLB**L**BBBNNN LLBLL**L**BLLBNB**B**BLB**B**BB NNN**L**BBLBLBLBL
2°2GB1	CEEEEEEECSSCEEEEEEEECSSHHHHHHHHHHHHHHTTCCSEEEEETTTTEEEEEC
2°model G	EEEEEETEHTHEEEEEEEEEHHEH HHHHHHHHHEHTEEEEEEE TTTEEEEEEEEE

Possible error candidates in mapping are shown in bold. The third line shows the new sequence for protein G, with differences from L/G shown in bold. Also shown is the secondary structure alignment for the B1 domain of protein G (fourth line) against reduced code of minimalist model for protein G (last line).

Results

Our original work on a minimalist representation of an α/β fold topology produced a sequence that adopts the correct native fold but exhibits at least two folding pathways, each of which corresponds to either a protein-L-like mechanism or a protein-G-like mechanism (8, 9). In light of this fact, this original α/β sequence will be called the "L/G" sequence to indicate it has properties of both proteins L and G, distinguishing it from our two new distinct sequences for proteins L and G.

Initially, we began our search for mutations to improve the L/G sequence by aligning it against the real protein sequences and proposing mutations that move the L/G sequence toward being more L- or G-like. The alignments are performed by using secondary structure as a rough guide. This process also requires categorization of amino acids as either L, B, or N. Shown in Table 1 is the general mapping between the 20- and three-letter codes, with the interpretation of the three flavors as simply hydrophobic, hydrophilic, or neutral. There is, of course, ambiguity in the mapping from a 20- to a three-letter code. Lysine can be viewed as either hydrophilic or hydrophobic, depending on the structural context. The long lysine side chain can participate in the hydrophobic core through its four methylene groups, whereas its amino group typically resides on the surface, which imparts its typical hydrophilic character. Important hydrogen bonds in the native structure might result in two hydrophilic amino acids (designated as an L bead) becoming two hydrophobic B beads in our reduced-letter code, i.e., emphasizing that an attractive interaction exists. However, in general, the mapping used in Table 1 is a good first approximation to the original sequence.

Tables 2 and 3 contain the sequence alignments for the L/G minimalist reduced code against the amino acid sequences of proteins L (2PTL) and G (2GB1), as well as possible errors in the assignments. We do not consider an L→N or N→L mapping as an error, because both L and N interact through repulsive potentials (see *Methods*); errors arise only from L→B, N→B, B→N, and B→L mappings, because the fundamental interaction potential changes from attractive to repulsive, or vice versa. Overall, we find ≈65–75% sequence identity between 2PTL and 2GB1 and our model L/G sequence.

Further improvements in sequence consist of proposing mutations and threading the new sequences through a (quenched) misfold database and subsequently looking for those mutations that give a favorable ΔE_{design}. To address the feasibility of using errors in the sequence alignments as a guide for proposing new mutations, we began initially by aligning protein L and using solely these errors to improve the sequence. In the case of protein G, all possible single mutations were investigated during the design process, and we did not restrict ourselves to errors solely from the alignment.

Starting from the original L/G sequence and proceeding through a series of mutations, we arrived at the new sequences for proteins L and G. Note that these mutations are only in the primary sequence, and that all proteins share the same secondary structure topology (dihedral sequence). Each new sequence required a series of five point mutations from the original L/G sequence. Shown in Tables 2 and 3 are the mutations used to obtain the new sequences. The energy of the original L/G native state is $-32.4\varepsilon_H$, whereas for the new protein L, the native-state energy is $-28.8\varepsilon_H$, and for protein G the native-state energy is $-26.9\varepsilon_H$. The energy distribution of the misfold library is well separated from these native state energies.

Two of the five point mutations for L and G are common to both sequences (B18L and B47L), which serve to make the proteins more foldable and appear to clean up thermodynamic aspects of the original sequence. This is evident in a comparison

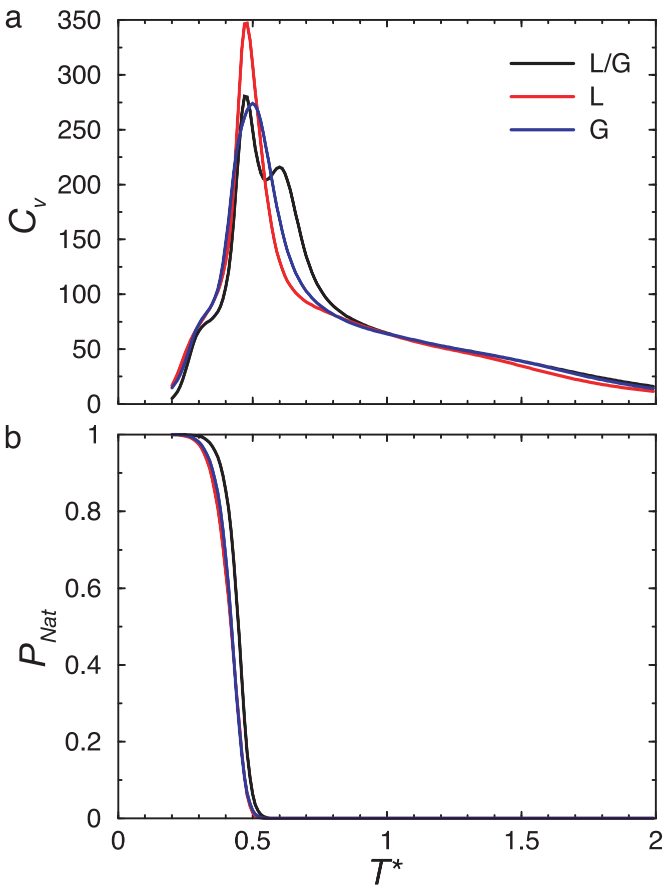

Fig. 1. Comparison of thermodynamic data for the new optimized sequences of proteins L and G to that of original L/G sequence. (*a*) Heat capacity C_v vs. temperature T for the original nonoptimal L/G, L, and G sequences. The new L and G sequences show increased cooperativity. (*b*) Fraction of chains in the native state, P_{Nat}, vs. temperature, T, for the L/G, L, and G sequences. At the folding temperature, T_f, the distribution of population between the folded and unfolded states is equal, that is, $P_{\text{Nat}}(T_f) = 0.5$.

of the heat capacity curves (Fig. 1*a*) for proteins L/G, L, and G. The sharper transitions in these thermodynamic signatures are evidence for greater cooperativity in folding. From the melting curves (Fig. 1*b*), we see that the folding temperature [the value of T for which $P_{\mathrm{Nat}}(T) = 0.5$] is lowered for L and G relative to the L/G sequence.

The remaining three mutations are different for proteins L and G and hence are responsible for differentiating the folding of protein L from G. Their main consequence appears to be changing the relative thermodynamic stability of elements of secondary structure. Fig. 2 shows the native-state similarity parameters for the α-helix and two β-sheet regions as a function of temperature. For $\chi\alpha$, the stability is reduced in the new L and G sequences relative to L/G. Of more interest are the changes in stability of $\chi\beta_1$ and $\chi\beta_2$. By looking at the mutations listed in Table 2, we see that the protein L reduced letter code introduces a net attraction into β-hairpin 1 and a net repulsion into β-hairpin 2. From Fig. 2, it can be seen that the result is a new L sequence in which the second β-sheet region (β_2) is destabilized relative to the β_2 region in the original L/G sequence. Likewise for the new G sequence, it can be seen that the mutations for the reduced code in protein G appear to introduce net stabilization into β-hairpin 2. Consequently, we get a sequence that favors a greater degree of β_2 ordering at higher temperature relative to the original L/G sequence (Fig. 2). This behavior is also reflected in the free-energy projections along $\chi\beta_1$ and $\chi\beta_2$ for L and G (shown in Fig. 3). From these projections there is a minimum free-energy path from the unfolded to folded ensembles that involves either formation of β-hairpin 1 then β-hairpin 2 (protein L), or β-hairpin 2, and then β-hairpin 1 (protein G).

Looking at folding kinetics around the folding temperature, shown in Fig. 4*a*, we see different behavior for L and G (values

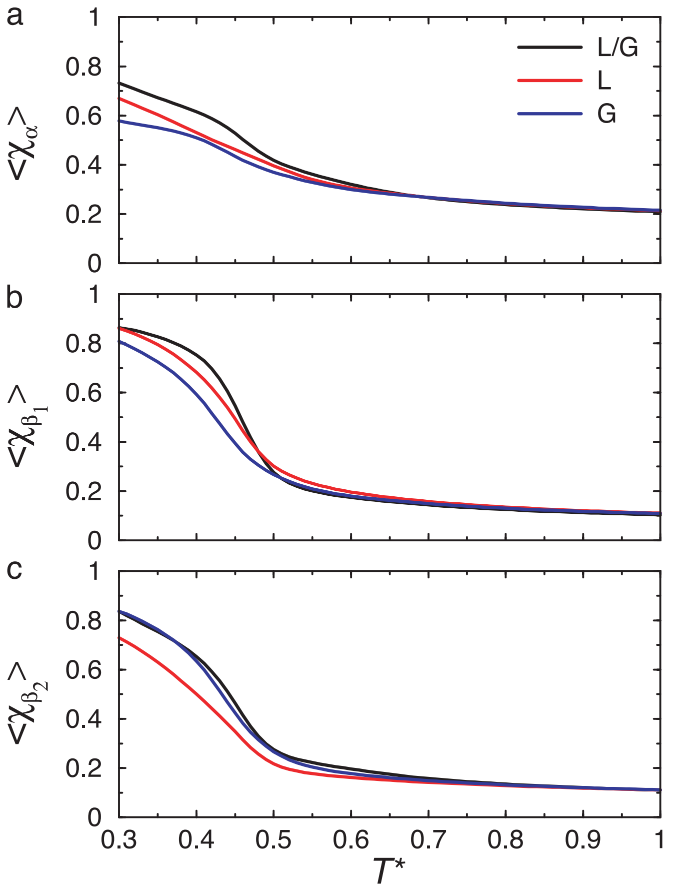

Fig. 2. Thermodynamic measures of the formation of native-state secondary structure. (*a*) Average native-state similarity of the α helix $\langle\chi\alpha\rangle$ vs. temperature T. Stability of the α helix is reduced for new L and G sequences compared with original L/G sequence. (*b*) Average native-state similarity of the first (N-terminal) β-sheet region $\langle\chi\beta_1\rangle$ vs. temperature T. Stability of β_1 region is reduced for new L and G sequences relative to original L/G sequence. Note that the stability is reduced further for G than for L. (*c*) Average native-state similarity of the second (C-terminal) β-sheet region $\langle\chi\beta_2\rangle$ vs. temperature T. Stability of β_2 region is similar for L/G and G sequences but reduced for the L sequence.

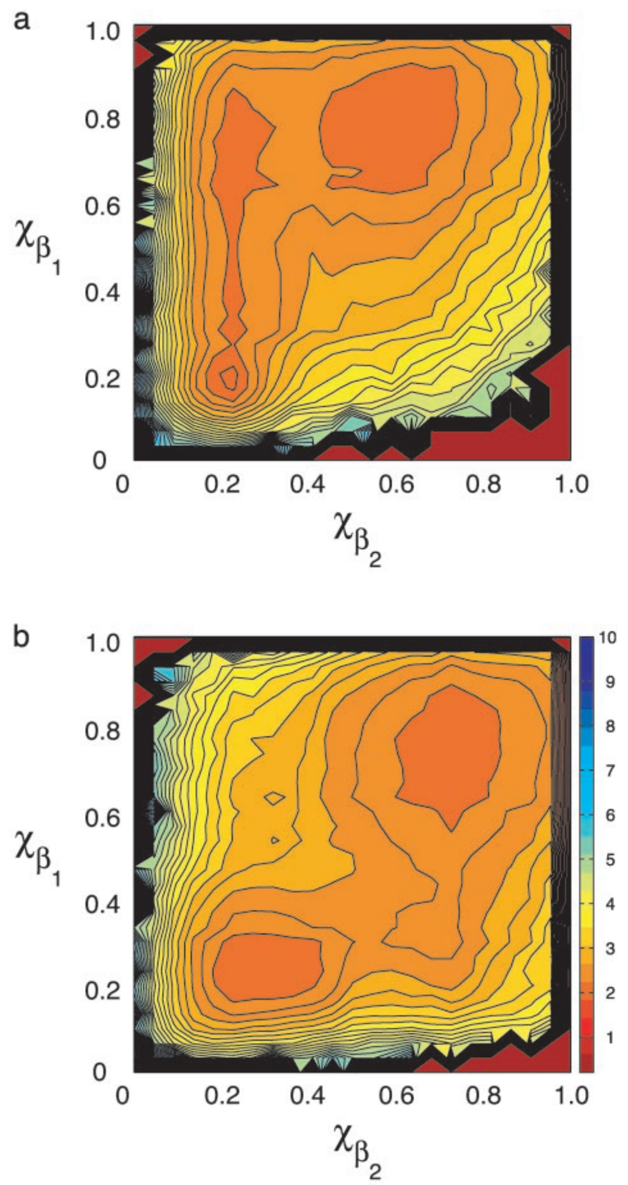

Fig. 3. Projection of free-energy surfaces onto order parameters $\chi\beta_1$ and $\chi\beta_2$. (*a*) Free-energy contour plot for protein L as a function of native-state similarity of the second (C-terminal) β-sheet region $\chi\beta_2$ and first (N-terminal) β-sheet region $\chi\beta_1$ at the folding temperature. Note the minimum free-energy path connecting the unfolded and folded ensembles proceeds through a transition state in which the β_1 region is native and the β_2 region is largely disrupted. (*b*) Free-energy contour plot for protein G as a function of native-state similarity of the second (C-terminal) β-sheet region $\chi\beta_2$ and first (N-terminal) β-sheet region $\chi\beta_1$ at the folding temperature. For G, the minimum free-energy path connecting the unfolded and folded ensembles proceeds through a transition state in which the β_2 region is native-like and the β_1 region is disrupted. Contour lines are spaced k_BT apart.

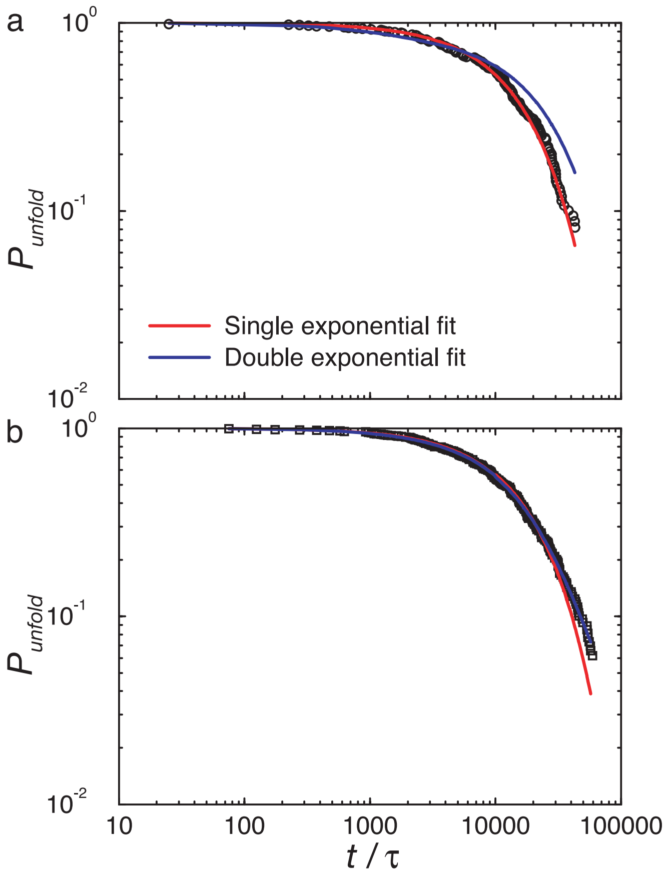

Fig. 4. Kinetic data with fits for proteins L and G. (*a*) Fraction of unfolded states P_{unfold} as a function of time *t* for protein L at the folding temperature. The best fit of the data is to a single exponential. (*b*) Fraction of unfolded states P_{unfold} as a function of time *t* for protein G at the folding temperature. The best fit for these data is to a double exponential. Fit parameters are given in Table 4.

given in Table 4). The kinetics for protein L is well fit by a single exponential. Thus, as reported for protein L in the literature, our minimalist-model protein L shows all of the signs of being a cooperative two-state folder. For protein G, the story is not as straightforward. We find that protein G folds slower than protein L by a factor of two, qualitatively consistent with experiment. The kinetics in protein G is better fit by a double rather than single exponential (Fig. 4*b*). The fit shows a fast folding event, involving ≈80% of the population, and a slow folding event that involves the remainder of the population. These results are consistent with differences in folding kinetics reported for proteins L and G, which we will analyze in greater detail elsewhere (S.B. and T.H.-G., unpublished work).

To illustrate that the mapping from the 20-letter amino acid sequence to three-letter reduced code can result in a protein model that folds to the correct topology, we consider preliminary studies on the prototype β-sheet protein, the WW domain. We focus on one specific member, the Pin WW domain, for which extensive experimental thermodynamic and kinetic analysis of its folding has been reported by Gruebele and coworkers (28). The 20-letter amino acid sequence, KLPPGWEKRMSRSSGRVYYFNHIT-NASQWERPSGN, was mapped to the following three-letter code: LBBBNBLLLBNLNNNLBBBBLLBNLBNLBLLBNNL.

Table 4. Parameters for kinetic fits at folding temperature

	T	A_0	$1-A_0$	τ_0	τ_1	$\chi^2/10^{-4}$
L	0.42	1.0	0	15700	0	3.43
G	0.41	0.81	0.19	13700	46400	0.353

Equation fit is $A_0\exp(-t/\tau_0) + (1 - A_0)\exp(-t/\tau_1)$.

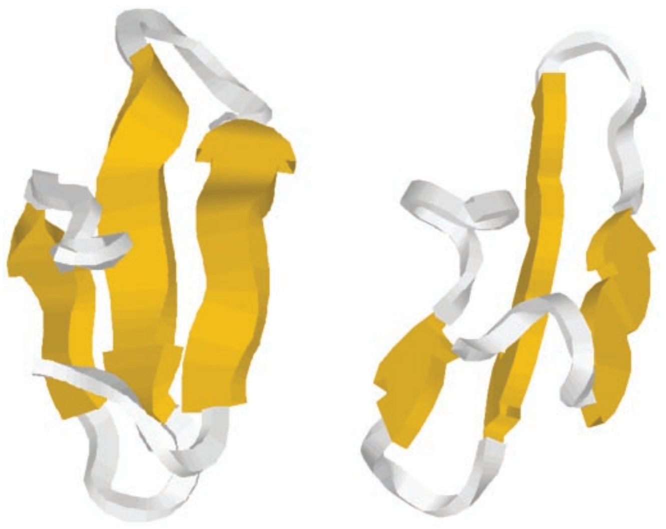

Fig. 5. Minimalist model of the native state topology for Pin WW domain (*Right*) and the NMR solution structure (*Left*), showing the very similar arrangement of secondary and tertiary structure.

Sequence design with the misfold library described in the *Methods* section was used to determine the following sequence: LBBN-NBLBLBNLNNNLBBBBLLNNLBNBBLLBNNL, which reliably folded to the correct β-sheet topology as the lowest energy structure (Fig. 5). Furthermore, we find the same network of residue–residue contacts of hydrophobic groups in the strand regions with the coil ends that is thought to stabilize the Pin WW domain (28).

Conclusion

Experiments on proteins L and G have previously shown that protein L favors a pathway involving formation of the first (N-terminal) β-hairpin, followed by formation of the second (C-terminal) β-hairpin (13–15). Protein G involves the sequential formation of secondary structure in the opposite order, namely β-hairpin 2 followed by β-hairpin 1 (18), and may involve the formation of populated intermediates (16, 17). This situation prompts the question of what changes of the protein G sequence are necessary to modify its kinetics to be more two-state with rate-limiting formation of the first β-hairpin, or alternatively, what changes to protein L would induce its mechanism of folding to be like that of protein G. Nauli *et al*. (29) address this question by using a computer-based design strategy that allows them to reengineer the protein G sequence, resulting in a new protein that folds 100 times faster and by a mechanism more faithful to wild-type protein L.

We answer this question in the context of our model by using a library of misfolds to quantify the consequences of sequence mutations and show that it is possible to design sequences that involve distinctly different folding mechanisms for minimalist representations of proteins L and G. It is important to note we do not *a priori* design a sequence to match the primary structure of naturally occurring protein. However, we find that our sequence design protocol has evolved sequences faithful to the primary structure of the real protein, assuming reasonable definitions of residues considered to be hydrophilic, hydrophobic, and defining small or ambiguous amino acids as neutral. After mapping the 20-letter code to a three-letter code and seeking optimal alignments in secondary and primary structure, we find ≈70–75% sequence identity in the cases we have considered within the ubiquitin fold class. Errors in sequence mappings were used to suggest mutations that actually corresponded to new sequences exhibiting differences in protein L and protein G folding mechanisms. Despite looking at all possible point mutations for protein G, the final outcome resulted in the selection of mutations (based on the ΔE_{design} criterion) that were in fact all beads corresponding to errors in the alignment. Therefore, it further strengthens our case that we can perform a sequence mapping onto our minimalist code, which could allow for the study of novel proteins whose structure is not yet known.

The main result of this paper is that sequence design with a three-letter reduced code is capable of translating the differences in primary sequence for proteins L and G into the expected differences in thermodynamic and kinetic properties. Heretofore, minimalist models have been appreciated for their ability to capture attributes relating to fold topology. Our physics-based model provides insight into the physical origin of how amino acid sequence patterning of hydrophobic and hydrophilic beads favors the formation of a target topology. This work restores the connection between free energy landscapes and amino acid sequence, the original protein folding problem. Another possible implication drawn from our simple model is that evolutionary perspectives of sequence conservation of a folding nucleus might be reexamined by considering the conservation of the overall patterning in the sequence (30, 31). Perhaps the patterning is tolerant to a small number of mutations, which would manifest as residues poorly conserved being found in the folding nucleus (31), also consistent with the robustness of fold topologies to mutation (32, 33).

We show that our minimalist model can also capture attributes explicitly depending on primary sequence that favors one folding mechanism over another. The differences in amino acid sequence could be manifested as explicit atomic representation of side chains (34, 35). However, studies have shown that simple patterning of polar and nonpolar residues is sufficient to produce compact states with significant secondary structure (36, 37). Additionally, active mutants of barnase have been found in which the hydrophobic core has been completely redesigned by incorporating random hydrophobic reassignments (38). These results point to the fact that explicit representation of side-chains may not be necessary.

The applicability of a three-letter code is also of interest because of the computational complexities inherent in protein design (39). For a sequence of N amino acid residues and 20 choices per residue with r rotamers each, the number of design possibilities scales as $(r \times 20)^N$. This neglects any additional complexity introduced by including backbone rearrangements, the incorporation of which can change results appreciably (40). The cost of protein design can be greatly reduced by the mapping of amino acid sequences into our three-letter minimalist code. Note that the idea of using a reduced set of amino acids in protein design is not new and in fact has been shown to be successful in a number of experimental studies (41). However, because our simplified model is computationally tractable, we are able to incorporate not only native-state structure into our design technique but also unfolded structures (misfolds) as well. This ability is significant in light of emerging experimental evidence that unfolded structures can have nontrivial influence on protein design (42).

Finally, we provide preliminary evidence that the mapping from the 20-letter amino acid sequence to a three-letter reduced code results in a sequence that reliably folds into one of the members of the WW domain topology. Furthermore, examination of this lowest-energy structure from simulated annealing reveals that it forms a specific network of hydrophobic contacts consistent with the origin of stabilizing contacts observed experimentally for the Pin WW domain (28). We will report on our results on WW domain folding elsewhere.

T.H.-G. acknowledges financial support from the University of California (Berkeley).

1. Shea, J. E., Onuchic, J. N. & Brooks, C. L. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 12512–12517.
2. Onuchic, J. N., Lutheyschulten, Z. & Wolynes, P. G. (1997) *Annu. Rev. Phys. Chem.* **48,** 545–600.
3. Go, N. (1983) *Annu. Rev. Biophys. Bioeng.* **12,** 183–210.
4. Koga, N. & Takada, S. (2001) *J. Mol. Biol.* **313,** 171–180.
5. Plaxco, K. W., Simons, K. T., Ruczinski, I. & David, B. (2000) *Biochemistry* **39,** 11177–11183.
6. Guo, Z. & Thirumalai, D. (1996) *J. Mol. Biol.* **263,** 323–343.
7. Sorenson, J. M. & Head-Gordon, T. (1999) *Proteins Struct. Funct. Genet.* **37,** 582–591.
8. Sorenson, J. M. & Head-Gordon, T. (2000) *J. Comput. Biol.* **7,** 469–481.
9. Sorenson, J. M. & Head-Gordon, T. (2002) *J. Comput. Biol.* **9,** 35–54.
10. Karanicolas, J. & Brooks, C. L. (2002) *Protein Sci.* **11,** 2351–2361.
11. Head-Gordon, T. & Brown, S. (2003) *Curr. Opin. Struct. Biol.* **13,** 160–167.
12. Wikstrom, M., Drakenberg, T., Forsen, S., Sjobring, U. & Bjorck, L. (1994) *Biochemistry* **33,** 14011–14017.
13. Scalley, M. L., Yi, Q., Gu, H. D., McCormack, A., Yates, J. R. & Baker, D. (1997) *Biochemistry* **36,** 3373–3382.
14. Gu, H. D., Kim, D. & Baker, D. (1997) *J. Mol. Biol.* **274,** 588–596.
15. Kim, D. E., Fisher, C. & Baker, D. (2000) *J. Mol. Biol.* **298,** 971–984.
16. Park, S. H., Oneil, K. T. & Roder, H. (1997) *Biochemistry* **36,** 14277–14283.
17. Park, S. H., Shastry, M. C. R. & Roder, H. (1999) *Nat. Struct. Biol.* **6,** 943–947.
18. McCallister, E. L., Alm, E. & Baker, D. (2000) *Nat. Struct. Biol.* **7,** 669–673.
19. Honeycutt, J. D. & Thirumalai, D. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 3526–3529.
20. Guo, Z. Y. & Thirumalai, D. (1995) *Biopolymers* **36,** 83–102.
21. Sorenson, J. M. & Head-Gordon, T. (2002) *Proteins Struct. Funct. Genet.* **46,** 368–379.
22. Andersen, H. (1983) *J. Comput. Phys.* **52,** 24–34.
23. Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H. & Kollman, P. A. (1995) *J. Comput. Chem.* **16,** 1339–1350.
24. Ferrenberg, A. M. & Swendsen, R. H. (1989) *Phys. Rev. Lett.* **63,** 1195–1198.
25. Ferguson, D. M. & Garrett, D. G. (1999) *Monte Carlo Methods Chem. Phys.* **105,** 311–336.
26. Bryngelson, J. D. & Wolynes, P. G. (1989) *J. Phys. Chem.* **93,** 6902–6915.
27. Sali, A., Shakhnovich, E. & Karplus, M. (1994) *J. Mol. Biol.* **235,** 1614–1636.
28. Jager, M., Nguyen, H., Crane, J. C., Kelly, J. W. & Gruebele, M. (2001) *J. Mol. Biol.* **311,** 373–393.
29. Nauli, S., Kuhlman, B. & Baker, D. (2001) *Nat. Struct. Biol.* **8,** 602–605.
30. Mirny, L. & Shakhnovich, E. (2001) *J. Mol. Biol.* **308,** 123–129.
31. Larson, S. M., Ruczinski, I., Davidson, A. R., Baker, D. & Plaxco, K. W. (2002) *J. Mol. Biol.* **316,** 225–233.
32. Li, H., Tang, C. & Wingreen, N. S. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 4987–4990.
33. Li, H., Tang, C. & Wingreen, N. S. (2002) *Proteins Struct. Funct. Genet.* **49,** 403–412.
34. Shimada, J. & Shakhnovich, E. I. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 11175–11180.
35. Clementi, C., Garcia, A. E. & Onuchic, J. N. (2003) *J. Mol. Biol.* **326,** 933–954.
36. Kamtekar, S., Schiffer, J. M., Xiong, H. Y., Babik, J. M. & Hecht, M. H. (1993) *Science* **262,** 1680–1685.
37. Xiong, H. Y., Buckwalter, B. L., Shieh, H. M. & Hecht, M. H. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 6349–6353.
38. Axe, D. D., Foster, N. W. & Fersht, A. R. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 5590–5594.
39. Voigt, C. A., Gordon, D. B. & Mayo, S. L. (2000) *J. Mol. Biol.* **299,** 789–803.
40. Larson, S. M., England, J. L., Desjarlais, J. R. & Pande, V. S. (2002) *Protein Sci.* **11,** 2804–2813.
41. Plaxco, K. W., Riddle, D. S., Grantcharova, V. & Baker, D. (1998) *Curr. Opin. Struct. Biol.* **8,** 80–85.
42. Hill, R. B. & DeGrado, W. F. (2000) *Struct. Folding Des.* **8,** 471–479.