Balancing energy and entropy: A minimalist model for the characterization of protein folding landscapes

Payel Das, Silvina Matysiak, and Cecilia Clementi*

Department of Chemistry, Rice University, Houston, TX 77005

Coarse-grained models have been extremely valuable in promoting our understanding of protein folding. However, the quantitative accuracy of existing simplified models is strongly hindered either from the complete removal of frustration (as in the widely used Go-like models) or from the compromise with the minimal frustration principle and/or realistic protein geometry (as in the simple on-lattice models). We present a coarse-grained model that “naturally” incorporates sequence details and energetic frustration into an overall minimally frustrated folding landscape. The model is coupled with an optimization procedure to design the parameters of the protein Hamiltonian to fold into a desired native structure. The application to the study of src-Src homology 3 domain shows that this coarse-grained model contains the main physical-chemical ingredients that are responsible for shaping the folding landscape of this protein. The results illustrate the importance of nonnative interactions and energetic heterogeneity for a quantitative characterization of folding mechanisms.

free energy landscape | molecular dynamics simulation | ϕ-value analysis | frustration | folding rate

The folding process of a protein into its biologically functional structure spans a large range of time and length scales. Molecular dynamics (MD) atomistic simulations aim to reproduce in detail the interactions of all of the atoms of the protein and solvent and are the tool of choice to study short time scales and/or more localized motions in protein systems at equilibrium.

Recently, the use of distributed computing (1) and the development of theoretical/algorithmical approaches improving the sampling of rare events (such as crossing large energy barriers) have significantly extended the reach of detailed simulations over longer time scales. Full folding trajectories have been recently obtained for small proteins (2, 3). However, extensive sampling of the overall folding landscape associated with large protein systems requires computational resources exceeding what is currently possible. Because of the computational impasse, theoretical studies on folding dynamics are oftentimes based on coarse-grained protein models that sacrifice most of the atomistic details to explore time scales and system sizes that are currently inaccessible to standard all-atom MD simulations. Simple protein models where residues are represented by one (or a few) bead(s) have been very valuable in advancing our understanding of the folding process, particularly when used in combination with experiments, or complemented by more detailed atomistic simulations on particular regions of a protein landscape. Some of the most recent simplified models incorporate experimental data and appear promising for strengthening the quantitative connections between theory and experiment (4–6). However, key issues remain unsolved. Problems in the present modeling scenario can be grossly divided into two classes: either the models artificially remove most of the interactions between pairs of residues that are not “in contact” in the native state (i.e., the so-called Go-like models), or they compromise with the minimal frustration principle and/or a realistic protein geometry (e.g., simple on-lattice models), thus hindering the possibility of a direct investigation of real proteins.

We present a coarse-grained modeling strategy and an optimization procedure for the model parameters that allow folding a minimalist off-lattice protein model into a desired native structure (within 1-Å rms deviation from the crystal structure). As the parameters of the model are optimized to fold into a known protein structure, the resulting model Hamiltonian contains information on the native structure. In this sense the proposed procedure maintains in spirit some analogies to the Go-like approach. However, by modeling the protein energy as a sum of pairwise interactions between 20 amino acid “colors,” this model does not a priori remove the contribution of nonnative interactions and naturally incorporates the energetic modulation of a protein sequence.

Go-like models have become very popular in the last few years for coarsely characterizing folding processes (see e.g., refs. 7–10). Their popularity is ascribable mostly to their ability to fold smoothly into any given (a priori known) native structure and their straightforward forward-modeling. Go-like models and their progeny (see e.g., refs. 11–14) base their roots on the commonly accepted hypothesis that efficient folding sequences minimize frustration (15, 16). The minimal frustration principle implicitly invites one to disregard nonnative interactions between residues as a zeroth order level of approximation. The complete removal of frustration can, however, significantly distort the representation of the folding landscape, even if its overall topography is oftentimes qualitatively preserved. It has been pointed out that Φ-values (17) obtained from simulations of existing coarse-grained protein models generally do not correlate well with experimentally determined Φ-values (see ref. 18 and references therein for a thorough analysis of this topic). Moreover, it has been shown that nonnative interactions may play an important role in shaping at least the early stages of the folding landscape, in the formation of on-pathways intermediates (19), in restricting the accessible configurational space in the unfolded state (20, 21), or in the formation of the transition state (22).

Alternative models have been proposed to account for the physicochemical information lost in a Go-like approximation (23, 24); however, if these models can correctly reproduce the overall arrangement of secondary and tertiary structure, they significantly deviate from the real native protein structure (24).

The main issue to tackle is to keep the ability to fold relatively smoothly into any real protein structure, without a priori removing the contribution of nonnative interactions, focuses on the design of effective parameters for a minimally frustrated Hamiltonian. Pairwise potentials in geometrically simplified protein representations are notoriously difficult to design (25). In this article we present a strategy that circumvents this problem. We show that it is possible to dress a minimalist Cα protein representation with a pairwise model Hamiltonian able to fold smoothly into the native state of a selected protein. The effective residue–residue interactions are modulated by 20 colors, representing the amino acids’ chemical and physical diversity.

In Model and Methods we present the general framework that can in principle be coupled with any coarse-grained Hamilt...
nian. We thoroughly test the model in close comparison with the available experimental data on the folding of src-Src homology 3 (SH3) protein domain (see Application: The Folding of the Src-SH3 Domain).

The results strongly support the validity of the model to quantitatively map a protein folding landscape. The high correlation obtained between the simulated and experimental Φ-values shows that all of the parameter sets obtained within this modeling framework produce a folding mechanism closely resembling the experimentally inferred mechanism.

As the proposed model incorporates sequence details and energetic frustration into an overall minimally frustrated protein landscape, it opens the possibility for a realistic study of the delicate balance between energetic (e.g., frustration, sequence dependence) and topological factors [i.e., the roughness in the configurational entropy (10)] in shaping the folding free-energy landscape that was not possible with a plain Gö-like model.

**Model and Methods**

**Basic Ideas and Assumptions.** A crucial ingredient of the proposed strategy consists of the design of the model parameters not only by enforcing standard minimum-energy requirements for the native state (see below and refs. 26–28), but also by selecting the potential functions to maximize native-state packing. We had shown that the design of equilibrium distances of continuous off-lattice potentials maximizing the native packing is a decisive factor for the definition of efficient folding models in a coarse-grained geometry (27). Moreover, it has been shown that the requirement of efficient packing is, by itself, sufficient to reproduce native protein motifs, such as helices and β-sheets (29). It is well known that proteins are extremely well packed and that mutations introducing cavities generally have a destabilizing effect (30, 31). The physicochemical rationale behind this fact is that efficient packing maximizes the number of interactions in the hydrophobic core. Recent studies have provided support for the so-called jigsaw-puzzle model (32), advocating that stereo-specific association of interacting side chains is a crucial ingredient for tight packing.

Some of the information on the specific shape and size of the amino acid is inevitably lost in a coarse-grained representation of the protein. Particularly, the possibility to fit tightly together the amino acid geometries is strongly hindered in a protein model based on isotropic interactions centered on coarse-grained atomic coordinates (as, for instance, the Cα atoms). If the high steric fit of interacting side chains can be likened to the geometric complementarity of irregularly shaped pieces in a jigsaw puzzle, clearly the same effect cannot be reproduced by modeling the residues as interacting spheres. If the equilibrium distance in the interaction between two residues is assumed to be a constant over all residue pairs (as it is in all non-Gö models previously proposed), then the model implicitly represents all residues as interacting spheres. A jigsaw-like effect can be reintroduced into a minimalist model, without necessarily increasing its complexity, by effectively incorporating some information on the nontrivial amino acid geometry into the definition of equilibrium distances of interresidue interaction potentials. In practice, this goal can be accomplished by using a distribution of equilibrium distances for each amino acid pair, rather than a constant value.

The importance of anisotropic coarse-grained potential functions has also been shown in the analysis of statistical-based effective interaction potentials for native fold recognition (33). Our strategy builds on similar premises but aims to the definition of a viable off-lattice minimalist model, naturally incorporating sequence detail and energetic frustration, while preserving an overall smooth folding landscape that ensures fast and reliable folding into the selected protein structure. The latter is the characteristic feature that has made popular the application of off-lattice Gö-like models. It is worth stressing that the goal of the proposed procedure is to advance the modeling and characterization of folding mechanisms beyond the Gö-like approach, and it is not intended as a tool for structure prediction. However, similar procedures have been used in the past to optimize energy parameters in the context of structure prediction, both for simplified protein-like structures (34) or real protein structures (35, 36).

**Definition of the Model Potential.** The potential energy function comprises a local and nonlocal term, $V = V_{\text{local}} + V_{\text{nonlocal}}$, where $V_{\text{local}}$ encompasses bond, angle, and torsional energy terms and is here designed to have its absolute minimum in the native state, as in previous work (8). However, the details of the local potential do not significantly affect the results presented here, provided that low energy values are associated with physically plausible local structures.

The nonlocal potential function for a protein of $N$ residues is chosen to be in the form:

$$V_{\text{nonlocal}} = \sum_{i<i,j=1}^{N} e(c_i, c_j) \left[ 5 \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 6 \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{10} \right].$$

The amino acid sequence is encoded in the function $c_i$ that associates a specific type, or color, to each residue $i$, $i \in (1, \ldots, N)$. The equilibrium distances $\sigma_{ij}$ for nonlocal interactions [between each pair of residues $(i,j)$ separated by more than three other residues along the chain] contains the information on the “shape” of the interacting residues. As discussed above, if the values of $\sigma_{ij}$ were taken to be constant, then the protein is reduced to a chain of spherical beads, and precious information on the geometric complementarity of interacting residues would be lost. To capture some of this information into the model, the value for the parameter $\sigma_{ij}$ for a pair of residues $(i,j)$ is extracted from a distribution $P(\sigma; c_i, c_j, \|i - j\|)$.

For each pair of residue types $(c_i, c_j)$, three different distributions are considered according to the relative distance of residues $i$ and $j$ along the sequence: the instances $\|i - j\| = 4$, $\|i - j\| = 5$, and $\|i - j\| > 5$ are considered separately, as a statistical analysis on the protein database shows that local steric effects can introduce significant deviations from the overall distribution of contact distances for the same pair of residues. The distribution $P(\sigma; c_i, c_j, \|i - j\|)$ is obtained by considering the occurrence of different values for the $C^\alpha - C^\alpha$ distance over all native contacts formed by pairs of residues of types $c_i$ and $c_j$, with the desired distance $\|i - j\|$ along the protein sequence, from a large (>4,000) database of nonredundant protein structures.

The parameters $e(c_i, c_j)$ and $\delta(c_i, c_j)$ are functions of the residue types $(c_i, c_j)$ only and are generated through an iterative procedure. The parameter $e(c_i, c_j)$ is positively defined and is constrained not to exceed a cutoff value (that determines the energy scale). The parameter $\delta(c_i, c_j)$ can only take one of two values: if $\delta(c_i, c_j) = 1$, the interaction between residues of type $c_i$ and $c_j$ is attractive; if $\delta(c_i, c_j) = 0$, it is repulsive. The iterative procedure consists of the following steps:

(i) **Generation (for the first iteration) or increment (for each following iteration), see step iv:** of a set of decoy structures $(\Gamma_i')$, $i = 1, \ldots, N_{\text{decoys}}$;

(ii) **Definition of an optimal set of parameters $(\{e_{\text{opt}}\}, \{\delta_{\text{opt}}\}) by using the maximum energy gap criterion on the set of decoys (26–28, 35, 57):**

$$\Delta E(\{e_{\text{opt}}\}, \{\delta_{\text{opt}}\}) = \max_{\{e\}} \left[ E_{\text{nat}} - E_{\text{refolding}} \right],$$

where $E_{\text{nat}}$ indicates the native structure (with energy $E_{\text{refolding}}$), and $E_{\text{refolding}}$ is the energy corresponding to the decoy structure $\Gamma_i$.

(iii) **Multiple “heat and quench” unfolding/refolding MD simulations with the obtained parameter set.**

(iv) **If the native state of the considered protein is consistently recovered during the unfolding/refolding process, the parameter set is considered effective and the dynamics is further investigated.**
Otherwise, the compact misfolded structures obtained in simulations are added to the decoy set and another iteration is implemented (step 1). Any protein configuration with rms deviation $> 3$ Å from the crystal structure is considered a decoy structure. A refolding simulation is considered successful if the lowest energy configurations in the folded state are closer than rms deviation $< 1$ Å to the crystal structure. An initial set of decoys is obtained by running short MD simulations with a random choice for the parameter values. The selection of parameters maximizing the energy gap is implemented by using a combination of Monte Carlo simulated annealing and perceptron learning (38) algorithms.

We used the SANDER module (properly adapted to deal with the minimalist protein representation) of the simulation package AMBER (version 6) (39) for all MD simulations presented in this work. All programs used in the optimization of the model parameters and analysis of the data were developed in-house and are available on request.

**Application: The Folding of the Src-SH3 Domain**

The abundance of experimental data and previous results on the folding of the SH3 domain makes this protein an ideal system to test our model. More than 10 high-resolution structures of different SH3 domains have been identified, and the folding thermodynamics and kinetics have been extensively characterized for most of them. The SH3 fold is relatively small, consisting of two β-sheets orthogonally packed around an hydrophobic core. Within the sheets, strands are joined together by the RT, n-src, and distal loops (40). We selected the sequence of the src-SH3 domain. The native coordinates for src-SH3 were taken from residues 84–140 of Protein Data Bank 1D code 1FMK.

Experimental studies have established that src-SH3 folding and unfolding can be modeled by simple two-state kinetics (41). Mutational studies (40) have shown that the transition state is strongly polarized, with a cluster of high $\phi$-values localized on strands 3 and 4 and the distal loop. Previous work has shown that the degree of “topological frustration” of src-SH3 is strong enough that its folding mechanism can be qualitatively mimicked by using completely unfrustrated models (8, 42, 43). However, recent experimental (44, 45) and theoretical studies (20, 21, 46) have highlighted the important role played by nonnative interactions in the folding of the SH3 domain. Particularly, Cobos et al. (44) have shown that the formation of non-specific hydrophobic interaction in an engineered SH3 domain results in an increase of both folding and unfolding rate constants, suggesting an increased stability of the transition state (with respect to the folded and unfolded states). This result is consistent with the counterintuitive theoretical prediction that a moderate degree of frustration introduced by the formation of nonnative contacts may actually assist the folding (20, 21). In general, Paci et al. (22) have shown for a set of short fast-folding proteins that, while it is reasonable to use pairwise interaction potentials to model the protein energy in folding/unfolding simulations, the use of Gō-type models completely disregarding non-native interactions may be misleading in the characterization of the transition state.

Because our model naturally incorporates both energetic heterogeneity and frustration in the protein energy function, it allows a quantitative investigation of the interplay between native and nonnative energy in the folding of SH3.

**Definition of the Model Parameters.** The iterative optimization procedure outlined above was used to obtain several sets of $(\xi, \delta)$ parameters that allow the protein model to fold the src-SH3 sequence into the src-SH3 native structure (within 1 Å from the x-ray structure). A total of five iterations was performed, and for each iteration cycle the Monte Carlo simulated annealing and perceptron learning minimization was repeated 100 times, producing 100 independent sets of parameters per iteration. For each set of parameter, multiple short heat and quench simulations (i.e., a very high-temperature unfolding simulation followed by increasingly low-temperature refolding simulations) were performed. If an unfolding/refolding simulation for a given set of parameter produced misfolded structures, instead of correctly recovering the native structure, the parameters were discarded, and the misfolded structures were added to a set of decoys. An initial set of 100 decoys was used to start the iterative procedure. Foldable sets of parameters are found after three iterations.

Over the iterations performed, a total of 16 sets of parameters were found that enable the simplified model to reliably fold back to the native state from high-temperature unfolded configurations. Given the high dimensionality of the parameter space, it is not surprising that several solutions are obtained. These 16 parameter sets are strongly correlated to each other (with an average correlation between pairs of parameter sets of 0.77, standard deviation of 0.12) and exhibit a similar degree of frustration (as defined below).

However, the small variations in the energy parameters reflect on slightly different $\Phi$-values for the different parameter sets, and on a fluctuation in their agreement with the experimental data (see below). The fact that this modeling framework leaves room to accommodate variations in the parameters is important for the potential applications, particularly to incorporate experimental data into a coarse-grained model (as in ref. 5). The latter issue goes beyond the purpose of this article.

**Thermodynamics and Kinetics Analysis.** Simulations were performed with all 16 sets of parameters obtained. Because of the high correlation between different parameter sets, very similar thermodynamic and kinetic behavior is found across the 16 sets of simulations. For instance, the correlation between probabilities of contact formation (considering both native and nonnative contacts) in given regions of the landscape (i.e., the transition-state ensemble) resulting from simulations with different parameter sets is found to be $0.8–0.95$. For this reason, we limit discussion to the results obtained for one particular set of energy parameters, the one producing the best agreement with the experimental $\Phi$-values. The effect of the small variations in the energy parameters across the 16 sets is discussed below, where a quantitative comparison with the experimental data is presented.

**Free-energy landscape.** Fig. 1 shows the projection of the folding free-energy landscape on the 2D subspace spanned by the reaction coordinates $Q$ and $A$, defined as the fraction of native contacts and nonnative contacts formed, respectively. Native and nonnative contacts are defined on the base of their probability of formation in the native state: if a given contact is formed in the folded state with a probability higher than in the transition state, then the
contact is considered native; otherwise it is considered nonnative. Moreover, to eliminate the noise that may arise from fluctuations of contacts that are not crucial for the stability of the native structure and may easily form or break in the folded state, the coordinate A is corrected by adopting the procedure introduced in ref. 20.

A single free-energy barrier separates the unfolded and folded states, in agreement with the two-state folding behavior observed experimentally. The definition of a free energy on subspaces defined by different sets of reaction coordinates shows a similar two-state behavior (data not shown).

Fig. 1 shows that the amount of nonnative contacts formed throughout the folding is not negligible, nevertheless it is not introducing significant roughness into the free energy. The free-energy surface closely resembles what was obtained in ref. 20, when a random, Gaussianly distributed interaction energy, with zero mean and variance $\sigma_{\text{nat}} = \epsilon_{\text{nat}}$, where $\epsilon_{\text{nat}}$ is the energy per native contact, was assigned to each native contact as a perturbation to a C$^{\alpha}$ Gō model of src-SH3. However, the analysis presented below shows that the effect of nonnative interactions in the present model goes beyond random noise, by inducing the formation of a cluster of nonnative contacts at the transition-state level, a result consistent with the experimental data (see below).

Glass temperature ($T_g$) and activation temperature ($T_a$). To measure the global and local “roughness” on the landscape induced by nonnative contacts, we estimated the $T_g$ and $T_a$. Below $T_g$ the protein remains frozen in a local minimum corresponding to one of the few compact misfolded (or at least partially disordered) configurations with low energy, and cannot evolve toward the global minimum (i.e., the native state). The ratio $T_g/T_f$ (between $T_g$ and folding temperature, $T_f$) quantifies the degree of frustration of the system, as it is related to the ratio $\Delta E/\delta E$ (16, 47), between the average energetic roughness $\Delta E$ (the energetic variance over different low-energy compact configurations) and the stability gap $\delta E$ (the average energy difference between the folded and unfolded state).

The presence of local roughness on the folding landscape can also be detected by the existence of another critical temperature $T_a$ ($T_a > T_g$), below which the dynamics becomes activated (48). In other words, the folding kinetics at $T < T < T_a$ is affected by the hopping dynamics of escape from local traps, while for $T > T_a$ the local escape barriers disappear (47).

$T_g$ can be estimated by assuming that the energy of compact misfolded structures is Gaussianly distributed. This assumption allows us to use the results obtained in the framework of the random energy model (15, 49) to obtain (see ref. 20 for details)

$$T_g = \Delta E_{\text{nn}}/\sqrt{2\kappa},$$

where $\Delta E_{\text{nn}}$ is the variance of the nonnative interaction energy over compact misfolded structures and $\kappa$ is the entropy associated to this ensemble of structures.

To obtain a large set of compact misfolded structures we performed multiple quenching simulations. Different initial configurations were created by running short simulation at extremely high temperature ($T = 40\ T_f$); the resulting unfolded configurations were then quenched to very low temperature $T < T_f$ to explore the structure of the misfolded low energy states. The expression above gives

$$T_g \approx 0.4\ T_f,$$

which is a realistic value for minimally frustrated proteins (47). $T_f$ is defined as the temperature correspondence to the peak in the heat capacity [obtained from standard thermodynamic analysis of simulation (8)].

$T_a$ can be estimated by monitoring the folding kinetics at increasingly low temperatures in the range $T_a < T < T_f$. Above $T_a$ the kinetics is single-exponential as it is regulated by only one major barrier-crossing event (corresponding to the transition state). Below $T_a$ the kinetics is expected to deviate from the single exponential behavior as the effect of local trap hopping dynamics is not negligible. The analysis over a large range of temperature gives an estimate for $T_a \approx 0.8\ T_f$. Remarkably, this value is fully consistent with the theoretical prediction of a ratio $T_a/T_f = 2$ for a protein of this size (47, 48). Kinetics analysis is provided in Fig. 5, which is published as supporting information on the PNAS web site.

**Folding rate: Effect of energetic heterogeneity and nonnative interactions.**

Both nonnative interactions and energetic heterogeneity are expected to have an effect on the folding landscape (20, 21, 50). In our model the two effects are intertwined, and the comparison with the corresponding results for a plain Gō model (where both effects are absent) allows us to estimate only their cumulative effect. However, it is of interest to evaluate separately the relative contribution of frustration and heterogeneity on the folding rate and folding mechanism. This estimate is possible by introducing an heterogeneous Gō-like model, defined by replacing all interactions in the model Hamiltonian involving nonnative residue pairs with an hard-core repulsive interaction (as customarily used in Gō-like models to maintain the self-avoidness of the protein chain, see e.g., ref. 8), while leaving unchanged the interactions involving native pairs. By comparing the results both with a plain, homogeneous Gō model and the above-defined heterogeneous Gō model, we can disentangle the effect of native energy heterogeneity and frustration.

The comparison of folding rates for the three models is presented in Fig. 2. The results are also compared with the theoretical prediction for the sole effect of either energetic heterogeneity (50) or frustration (20, 21).

Consistent with theoretical results, both heterogeneity and frustration contribute to increase the folding rate with respect to the homogeneous Gō model. Moreover, the observed rate increase agrees remarkably well (i.e., within the error bar) with the theoretical prediction in both instances, namely:

(i) The effect of native heterogeneity on the rate can be estimated...
by considering the change on the free-energy barrier $\delta F$, obtained as a perturbation by means of free-energy functional methods (50):

$$
\delta F = \sum_{i,j} (\langle Q_{ij} \rangle_{TS} - \langle Q_{ij} \rangle_{U}) (e_{ij}^{\text{homo}} - e_{ij}^{\text{hetero}}),
$$

and

$$
k_{\text{homo}}^{\text{hetero}} = \exp\left(\frac{-\delta F}{kT}\right),
$$

where the sum runs over all interacting residue pairs, and $\langle Q_{ij} \rangle_{X}$ is the probability of finding a given native contact $(ij)$ formed, computed over the ensemble $X$ ($X$ = transition state, TS; or unfolded, U) of the homogeneous model. The equations above yield a value log ($k_{\text{hetero}}/k_{\text{homo}}$) $\approx 0.85 \pm 0.15$, which is consistent with the value log ($k_{\text{hetero}}/k_{\text{homo}}$) $\approx 0.75 \pm 0.2$ obtained from simulation (see Fig. 2).

(ii) A small amount of nonnative energy can assist the folding by creating a more compact transition state. By using the same procedure as ref. 20, and assuming that the nonnative energy is Gaussianly distributed, the change on the free-energy barrier upon introduction of nonnative interactions to the heterogeneous Gō model can be estimated as:

$$
\delta F = \frac{\flow_{NN}}{kT} - \frac{b^2}{2kT} \left(A_{TS} - A_{U}\right),
$$

where $A_X$ is the average number of nonnative contacts formed in the ensemble $X$ ($X$ = TS or U), $e_{NN}$ and $b^2$ are the mean energy and energetic variance of a nonnative interaction, respectively. Expression above predicts an increase in the folding rate log ($k_{\text{hetero}}/k_{\text{homo}}$) $\approx 0.34 \pm 0.15$ that is in agreement with the value log ($k_{\text{hetero}}/k_{\text{homo}}$) $\approx 0.31 \pm 0.15$ obtained from simulation (see Fig. 2).

Quantitative Comparison with Experimental Data. To test the predictive power of our model, we computed the $\Phi$-values for all 16 sets of parameters that enable the model to correctly fold to the correct SH3 structure (see above), and compare the results with the experimental data (40, 51). $\Phi$-values and associated uncertainties are computed by using free-energy perturbation theory (to the first order), as discussed (5, 8, 13).

All 16 foldable parameter sets produce $\Phi$-values exhibiting a significant correlation with the experimental data (correlation coefficients 0.55 < $r$ < 0.78, $P$ values $10^{-13} < P < 10^{-7}$). A histogram of the values of correlation coefficients as obtained over all 16 parameter sets is provided in Fig. 6, which is published as supporting information on the PNAS web site. The best set of $\Phi$-values is plotted together with two sets of experimental data in Fig. 3. The correlation between simulated and experimental $\Phi$-values increases up to $r \approx 0.85$ when the experimental data used for comparison is restricted to mutations destabilizing the protein more than $\Delta\Delta G_{\text{H}} > 4$ kJ/mol. This selection is suggested by recent experimental tests to eliminate the most noisy experimental measurements while retaining the highly reproducible data (K. W. Plaxco and P. E. Wittung-Stafshede, personal communication).

It is worth stressing that, although energetically unfrustrated models can reproduce qualitatively well the folding landscape of a large set of proteins, the $\Phi$-values obtained by using a plain Gō-like model give a lower correlation with the experimental data (18). Particularly, for the protein considered here the Gō model $\Phi$-values give $r \approx 0.45$, $P \approx 10^{-5}$.

The $\Phi$-values obtained with the heterogeneous Gō-like model (as defined in the previous section) yield an intermediate correlation, in between the results obtained with a plain Gō model and the non-Gō model presented here. This evidence suggests that, for this protein, nonnative contacts do play a role, at least in shaping the details of the energy landscape around the transition state. This result is fully consistent with the analysis of Paci et al. (22), that showed that, although a major contribution to the energy comes from native interactions, nonnative interactions cannot be neglected.

A nonnegligible amount of nonnative energy has a 2-fold influence on the calculation of the $\Phi$-values. First, it can affect the
of different contacts at different stages of the folding process by perturbing the folding landscape (therefore affecting the folding dynamics). Second, even if the folding landscape were preserved exactly the same, it can still affect the Φ-values by changing the energetic balance among the interacting pairs involved in the mutation. Fig. 3 quantifies the latter effect. The area below the Φ-values was computed by using the full Hamiltonian which is shaded in blue in Fig. 3, while the red area corresponds to the Φ-values computed by considering only the native interaction on the configurational ensembles sampled by simulating the non-Go model with the full Hamiltonian. Interestingly, the largest difference between the two sets is found in regions of high Φ-values. Particularly, native contacts do not significantly contribute to the high Φ-value obtained at position 28 of our model protein (corresponding to Val-35 in src-SH3), that is then caused mainly by nonnative contacts. Experimental data confirm that mutation V35A yields a large Φ-value (0.77 ± 0.05) (51). The role of nonnative interactions at the transition state is apparent from Fig. 4, which shows the probability of formation of native and nonnative contacts at the transition state. A main cluster of nonnative contacts (circled in red in Fig. 4) appears formed at the transition state. The residues involved in this cluster correspond to the hydrophobic core of SH3; therefore, these nonnative contacts can be interpreted as a nonspecific hydrophobic cluster. The very same cluster of nonnative contacts has been identified as important for the folding of SH3 domains by mutational analysis (44, 52) and previously suggested by the analysis of nonfunctionally relevant conserved residues (46).

Conclusions

We propose a minimalist model and a procedure for the definition of the parameter model that allows the consideration of off-lattice simplified protein representations into their correct native state. As the parameters are optimized to fold into a known native structure, the model potential energy is not “ab initio.” However, by modeling the protein energy as a sum of pairwise interactions between 20 amino acid colors, this model does not a priori remove any contribution from nonnative interactions and naturally incorporates the energetic modulation of a protein sequence. The effect of nonnative interactions and energetic frustration is strongly minimized by maximizing the energy gap between native and misfolded structures. The ability of the model, despite its extreme simplicity, to fold a protein sequence relatively smoothly into its correct native structure relies into a “mean-field-like” definition of equilibrium distances for the interacting amino acid pairs, which allows us to effectively capture their nonspherical geometry. We have presented a detailed analysis of the folding of src-SH3 as obtained from simulation within the framework of this model, in close comparison with the experimental results. The results from our analysis are in good agreement with all available experimental data and theoretical predictions. The picture emerging from this synoptic analysis of theory, experiment, and simulation results supports the idea that nonnative interactions as well as energetic heterogeneity play a nonnegligible role in determining the detailed folding mechanism of SH3. Particularly, we show that, while most of the large Φ-values measured for this protein can be interpreted in terms of an high probability of formation of a set of native interactions in the transition-state ensemble, some large Φ-values may be predominantly determined by the stabilization of nonnative interactions. Similar observations have been reported before (22, 46).

We believe that the model presented in this article introduces a significant improvement into the current protein folding minimalistic modeling scenario and provides a solid starting point for the development of synergistic theoretical/experimental methods (4, 5) aimed at a more quantitative characterization of folding free-energy landscapes.

We thank Giorgio Favrin and David Maxwell for their contributions to the initial stage of this project, Steve Plotkin and Charlie Brooks for precious advice and suggestions, and Giovanni Fossati for assistance with computer-related problems. This work was funded by the Welch Foundation (a Norman Hackerman Young Investigator Award and Grant C-1570), Texas Advanced Technology Program Grant 03604-0010-2003, and National Science Foundation CAREER Award CHE-0349303.