De novo prediction of protein folding pathways and structure using the principle of sequential stabilization

Aashish N. Adhikari1,2, Karl F. Freed1,2,3, and Tobin R. Sosnick4,5,6

1Department of Chemistry, University of Chicago, Chicago, IL 60637; 2The James Franck Institute, University of Chicago, Chicago, IL 60637; 3Computation Institute, University of Chicago, Chicago, IL 60637; 4Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, IL 60637; and 5Institute for Biophysical Dynamics, University of Chicago, Chicago, IL 60637

Edited by S. Walter Englander, The University of Pennsylvania, Philadelphia, PA, and approved September 10, 2012 (received for review May 26, 2012)

Motivated by the relationship between the folding mechanism and the native structure, we develop a unified approach for predicting folding pathways and tertiary structure using only the primary sequence as input. Simulations begin from a realistic unfolded state devoid of secondary structure and use a chain representation lacking explicit side chains, rendering the simulations many orders of magnitude faster than molecular dynamics simulations. The multiple round nature of the algorithm mimics the authentic folding process and tests the effectiveness of sequential stabilization (SS) as a search strategy wherein 2° structural elements add onto existing structures in a process of progressive learning and stabilization of structure found in prior rounds of folding. Because no a priori knowledge is used, we can identify kinetically significant non-native interactions and intermediates, sometimes generated by only two mutations, while the evolution of contact matrices is often consistent with experiments. Moreover, structure prediction improves substantially by incorporating information from prior rounds. The success of our simple, homology-free approach affirms the validity of our description of the primary determinants of folding pathways and structure, and the effectiveness of SS as a search strategy.

TertItFix | folds | kinetic traps | Monte Carlo simulation

Despite numerous advances since the original sequence-to-structure folding paradigm was proposed over 50 years ago (1), we still lack a general framework that enables simultaneous prediction of the folding mechanism and structure using only the amino acid (aa) sequence [notwithstanding recent successes of all-atom simulations to fold small, fast-folding proteins (2)]. An obvious obstacle is the astronomical number of conformations available to a polypeptide. Proteins overcome this obstacle by sampling a limited set of conformations, guided by the folding process itself. However, most successful structure prediction methods do not consider the folding mechanism when sampling conformations. Conversely, many methods for predicting folding mechanism rely on knowledge of the final structure (e.g., Gō models).

Another obstacle emerges because many non-native and near-native conformations often differ by only a few RT, which is at or beyond the ability of current energy functions to reliably distinguish. A related difficulty arises because the native state is the global free energy minimum even if three competing properties—local backbone torsional angle preferences, hydrogen bonded 2° structure, and 3° packing—are not individually optimized. For example, 3° context can override local biases in determining the final 2° structure (3). Hence, a successful framework should couple 3° context to 2° structure formation, rather than relying on a strict hierarchical approach.

Sequential stabilization (SS) provides one mechanism for coupling 2° and 3° structure formation during folding and guiding the search process (4, 5). Supported by native state hydrogen exchange experiments, ψ analysis, and other observations (6, 7), this view argues that proteins predominantly fold along one or a few low energy pathways determined by the stepwise addition of cooperative units of structure or foldons (e.g., a helix or a strand). Prior emergence of hydrogen bonded structure serves as a template for the formation of additional structure that may only exist as a minor population in isolation.

Here we describe an iterative framework, termed TertItFix, to test whether the combination of SS with basic principles of protein chemistry can be used to predict folding pathways and structure using only the sequence as input. The principle of SS is implemented by using the statistics of folding trajectories garnered from prior rounds of simulation to bias the subsequent sampling of backbone dihedral angles (8) and the energies of tertiary contacts and hydrogen bonds. The approach combines simple backbone torsional ϕ, ψ moves, a polypeptide chain with no side chains beyond Cβ carbons, and multiple rounds of simulation with the progressive learning and building of 3° motifs through constraints imposed by data from prior rounds. We predict the 2° and 3° structures and pathways for 8 proteins using only approximately $10^3$ CPU hours per protein. The results are largely consistent with experimental data, even in the presence of kinetically significant non-native interactions.

Model

Initially, approximately 500 individual Monte Carlo Simulated Annealing (MCSA) folding simulations are performed using specialized ϕ, ψ backbone moves and energy functions appropriate for a reduced chain representation consisting of the backbone plus Cβ heavy atoms, as discussed below. The best final structures (lowest energy quartile) are then examined for recurring 2° structures, backbone hydrogen bonding, and 3° contacts. After modifying the move set and energy functions to promote these recurring features, another round of approximately 500 folding simulations is performed. The passing of information from one round to another is repeated until convergence (Fig. S1). This iterative, multiround learning and biasing procedure equates to a search strategy involving sequential stabilization, as illustrated with the folding of ubiquitin (Fig. 1).

The folding simulations employ move sets and energy functions that are designed to describe three competing protein properties: ϕ, ψ preferences, 2° structure, and 3° packing. Angle preferences are incorporated by sampling conformational space using neighbor-dependent ϕ, ψ distributions derived from the PDB (Fig. S2). These angles are used for pivot moves, where only a single residue’s ϕ, ψ angles are changed, as well as for double crankshaft local moves, where two consecutive peptide groups are rotated (9, 10) (Fig. S3). In the initial round, angles are chosen...
Our energy function is composed of three statistical potentials and two biasing terms (SI Text, Fig. S4), which guide the formation of 2° structure and 3° packing. The first potential (8) describes the residue-residue interactions according to the distance distributions in the PDB, contingent on 2° structure and the relative orientation of the two residues’ Cα-Cβ vectors (Fig. S4). The second potential describes each residue’s burial propensity, as calculated using the number of heavy atoms surrounding each Cβ atom in an 11 Å hemisphere defined by the orientation of the Cα-Cβ vector (glycines are ignored). The third term is associated with backbone desolvation and backbone hydrogen bonding. The desolvation term assigns a penalty for the loss of water-peptide hydrogen bonds when there is no compensating protein-protein hydrogen bond (Fig. S4).

Even for small proteins, the exploration of the folding landscape poses a formidable search problem, and thus requires additional constraints. The principle of SS provides a realistic method of guiding the search by continually refining the φ, ψ sampling distributions, which determine local structure, and biasing the energy for recurring motifs to guide 2° and 3° structure formation. After each round, the lowest energy quartile is evaluated to identify 2° structure preferences and popular 3° contacts and hydrogen bonding. These items are used to restrict the backbone sampling library (Fig. S2) and to generate two energetic biasing terms, E_contact and E_H-Bond, that are employed in the next folding round (see Methods).

This iterative process incrementally fixes 2° structure and biases 3° structure and hydrogen bonding as the rounds proceed, producing a series of sequential steps that may correspond to the authentic folding pathway (Fig. 1). Individual biases may strengthen or weaken in subsequent rounds because of the emergence of competing contacts. In principle, if no major kinetic traps impede the pathway, the final sampling distribution and contact probabilities should converge to produce the native 2° and 3° structures, respectively.

Results
We apply TerItFix to deduce the folding pathways of eight proteins and simultaneously to predict their 3° structures. The different levels of information accessible are demonstrated by studying the fast folding five helix subdomain of lambda repressor (φ–ψ) and comparing the predictions to experiments (12, 13) and molecular dynamics (MD) simulations (2, 14, 15). Next, TerItFix is applied to describe the folding pathways of two homologous immunity proteins, Im7 and Im9, along with a double point mutant, SIm9, to demonstrate that our method is sensitive enough to capture the kinetic consequences of slight aa variations and to predict the presence of kinetic intermediates. Finally, TerItFix is used to describe five other proteins,Ub, chymotrypsin inhibitor 2 (C12), Protein L and two three helix bundle proteins, Protein A and the designed α3d.

λφ-ψ. Starting with a φ, ψ distribution generated from the coil library, a folding pathway emerges after five rounds of folding with a 4.3 Å Cα RMSD (best) structure (Fig. 2). The initial φ, ψ distribution provides little indication of the positions or propensities of the helices or their order of formation because most angles in the initial distribution are nonhelical. A clear pathway emerges as the rounds proceed, with helices H3 and H4 appearing first and interacting. As the probabilities of these two helices increase in progressive TerItFix rounds, H1 gradually appears and docks against the H3-H4 motif by round R3. While the number of helices remains largely unchanged after R3, the helices lengthen, and their contact probabilities continue to increase for the next few rounds, as evident by the evolution of the φ, ψ distributions, average contact maps, and the centroid of the largest cluster formed from the structures generated in each round (Fig. 24). Although H2 and H5 appear in some trajectories, the population of structures containing these helices remains insufficient to justify restricting the sampling distributions in those positions to the helical basin.

These simulations highlight the interplay between 2° and 3° structure formation. The helical probability for the residues of helix H2 in R1 exceeds that in subsequent rounds. This loss of native-like structure suggests that H1 and H2 initially interact, but 3° contacts between H1 with H3 and H4 dominate in later rounds. At the same time, the average contacts between H4 and H5 continue to rise until round R4, even though residues in H5 never become highly helical (Fig. S5). Besides predicting the pathway, incorporating the strategy of SS into TerItFix improves the predicted structures (Fig. 2B).
H3 is unclear (H1 and H4 are formed in the TSE, while the presence of H2 and bond formation suggests that at least one more helix is present because helical content in an unfolded analog is approximately 16% at 310 K (21), which is much lower than the 42% helical content in the MD simulations but accords with the low approximately 15% helical angle content in TerItFix’s initial \( \phi, \psi \) coil sampling library.

A Markov state analysis of 3265 relatively short (<sec) MD simulations identifies a TS structure \( (P_{\text{TS}} = 0.53) \) having only 1–2 turns of helices H1–H4 and two adjacent \( \beta \) strands (14).

Another set of MD simulations using a new tempering method finds that H1–H3 are formed prior to H4 and H5 (15). Further experiments should be performed to permit more accurate assessments of the disparate results obtained by TerItFix and the three different MD simulations.

**Im7,9.** The homologous immunity proteins Im7 and Im9 highlight a case where TerItFix is advantageous over native-biased methods.

Im7 and Im9 display different folding kinetics despite being nearly identical four-helix bundles with approximately 60% sequence identity. Im7 folds in a three-state manner with an intermediate containing helices H1, H2 and H4, while Im9 folds in a two-state manner (22, 23). Im7’s three helix intermediate is misfolded in the sense that the three helices must at least partially separate in order to accommodate H3. The importance of sequence is further highlighted by the fact that only two conservative mutations in Im9 (“\( \text{Slm9} \)” induce a three-state mechanism akin to Im7’s (24). That such slight variations of the aa sequence can alter the folding behavior reflects the challenge of reproducing these results.

After only three rounds of simulations, all four helices form and interact in Im9, whereas H3 fails to form in Im7 (Fig. 3A and Figs. S6 and S7). The sampling distribution for the residues in H3 of Im7 never evolves beyond the coil specification, and the protein becomes “trapped” in an intermediate structure containing H1, H2 and H4 (Fig. S6). Thus, TerItFix correctly captures the energetic frustration of the folding landscape of Im7 that is absent in Im9.

Next, TerItFix is applied to the folding of Slm9 which has the conservative V37L and V71I substitutions in H2 and H4, respectively, and folds with the accumulation of the same three helix intermediates as Im7. Remarkably, the TerItFix results for Slm9 are very similar to those for Im7, successfully predicting the same three helix intermediate as observed experimentally (Fig. 3A) and demonstrating a high level of sensitivity of our method to changes in sequence and the energy landscape.

The origin of the sensitivity to two conservative mutations is deduced from the differences in the H1-H2 contacts for Im9 and Slm9 (Fig. S7B). Slm9’s two mutations promote docking of these helices in a geometry that precludes the addition of H3. Specifically, the two mutations alter the pairwise DOPE-PW energies between the helices (Fig. 3B). In Slm9, the interactions between V37L, which lies in H2, and residues in H1 are stronger, while the interactions between V37L and I53 in H3 are weaker. These two differences provide an explanation for the disparate folding mechanism induced by only two aa mutations.

**Prediction of Early Events, Foldons and Non-native Contacts.** TerItFix simulations begin from a conformation devoid of regular structure. Hence, the method can provide insights by identifying
motifs that form at the earliest stages of folding for five proteins: Ub, CI2, Protein L, α3d (25), and Protein A.

Ub is a 76 residue α/β protein with a relatively complex topology and a folding pathway that has been extensively characterized by ψ analysis and native state hydrogen exchange (5, 7, 16). The TSE contains four adjoining strands, β1–4, and part of the major α helix. Folding from the TSE to the native state occurs in a stepwise manner with the addition of the small 3β4 helix followed by the β5 strand. However, the early events leading to the TSE are difficult to identify due to their intrinsic instability and the ensuing two-state kinetic folding behavior.

The first motif to form in the TerItFix simulations is the β1-β2 hairpin (Fig. 4 and Fig. S8), followed by the addition of the α helix and the interaction between the two terminal strands, β1-β3. The early interaction between the termini is significant because long-range contacts generally form with greater difficulty, especially when 30+ intervening residues are still unstructured. Although β1-β3 form a parallel arrangement in the native structure, we observe some non-native antiparallel arrangements. The subsequent steps include the formation of contacts between β3 and β4 and the strengthening of contacts between the helix and β4 in later rounds. By R6, both the 2° structure distribution and average contact maps plateau. While the two remaining folds, the 310 helix and β5, maintain low populations in the contact maps, enough steps along the folding pathway are resolved to obtain the correct fold and a best Cα RMSD structure of 4.6 Å. The TerItFix results are consistent with experiments; in particular, the folds known to be in the TSE are predicted to form prior to the two folds known to fold after the transition state.

CI2 contains both parallel and antiparallel β strands onto which a single helix and an active site loop G3 are packed. CI2 folds in a two state manner, with a TSE characterized using ψ analysis (26, 27). The helix has the highest ψ values, followed by strands β3 and β4. In the first TerItFix round, interactions appear throughout the protein (Fig. 4). By R2, the helix begins to emerge

---

**Fig. 3.** Im proteins. (A) Contact maps. By the last round, R3, TerItFix predicts that the H3 is formed in Im7 (green circle) whereas H3 is absent in Im7 and Sm9, as observed experimentally (8). In Sm9, the interaction energy between V37L strengthens the H1-H2 contacts but destabilizes the H2-H3 contacts between V37L and I53. By R3, the helix begins to emerge

**Fig. 4.** Ub, CI2 and α3d. The average contact maps in R5 (upper) are compared to the native contact maps (lower). For Ub, the contacts are circled according to experimental ψ values (green: ψ = 0, absent in T5; red: ψ = 1, present in the TS). For CI2, the non-native contacts are noted with the red arrow. The detailed pathways for the three proteins are in Figs. S1–S3.
invokes no assumptions about 2° structure or uses fragments, while running many orders of magnitude faster than MD simulations (CPU hours compared to CPU weeks). In the absence of major kinetic traps, we expect that TerItFix can predict the native structure for many small proteins. A further test emerges from additional simulations for the set of 12 fast-folding proteins recently investigated by the DE Shaw group using all-atom MD simulations (43). As will be described elsewhere, we obtain an average Ca-RMSD$_{best}$ of $2.7 \pm 1.2$ Å as compared to $2.0 \pm 1.3$ Å from the MD simulations, with TerItFix producing lower values for 5 of the 12 targets. However, proteins with complicated folds such as SH3 still pose a challenge for TerItFix.

While the detection of kinetic traps is one success of our method, resolving them remains difficult. To counter this difficulty, we refold the protein in every round starting from an extended conformation, but using the information garnered from the previous round in the form of sampling and energetic biases. Because the prior information is implemented as biases, rather than as enforced contacts, both native and non-native contacts can weaken in successive rounds. For example, the native-like contacts between H1-H2 in λ repressor form early, are lost in middle rounds, and then reappear in later rounds. Im7, however, provides an example where the new contacts cannot override the earlier, non-native ones, and the protein becomes trapped in an intermediate state. A signature of a kinetic trap in our simulations is the presence of region(s) whose structural diversity varies within and between rounds. Potentially, the threshold for fixing 2° structure assignments and biasing 3° contacts can be reduced to drive the escape from the trap.

Another impediment to modeling protein folding is the inherent difficulty of correctly balancing the energies associated with different types of contacts and backbone geometries. Small errors in the energy function, or the lack of explicit hydrogen bonds and backbone $\phi$, $\psi$ dihedral angles, can greatly impact the order of structure formation and the location of the TSE on the reaction surface. These issues contribute to the inability of nearly all prior methods to accurately describe the TSE of Protein L(18) and Protein A (17).

The TerItFix algorithm’s central feature of coupling the 2° and 3° structure by iterative fixing and SS helps identify low energy pathway(s) with the proper order of structure formation. Nevertheless, we experience difficulty identifying the TSE for Protein L and Protein A. Even though simulations for both these proteins converge within three rounds, ascertaining the TSE is difficult and requires auxiliary information. Our prior $\psi$ analysis studies of four proteins with disparate RCO levels indicate that their TSEs acquire a similar fraction of native topology, $RCO_{TSE} \approx 0.7 \cdot RCO_{Native}$ (5, 16–18). Accordingly, we cluster all structures from the TerItFix simulations whose RCOs are between 60% and 80% of the native value to identify a TSE (Fig. S13). The major cluster for Protein L has both hairpins folded, in agreement with experiment. But the amino portion of the helix is also folded, which is not observed experimentally (18). Overestimation of the helical content in Protein L’s TSE is typical of other methods as well (18). The TerItFix-determined TSE for Protein A has H1 and H3 along with a kinked helix H2. This structure is close to experiment, except that in the experimental studies, the ends of H1 and H3 are frayed and H2 is not kinked.

**Conclusion**

We present TerItFix, a holistic approach for predicting pathways and structure that couples basic principles of protein chemistry with a realistic and robust search strategy involving sequential stabilization to find low-energy folding routes. Central to the TerItFix folding algorithm is the progressive learning and biasing of 2° structure, 3° contacts, and backbone hydrogen bonding. Information learned in one round of folding simulations is used in the following round. This work demonstrates that the empirical
principle of SS can be applied as a computational strategy to predict both pathways and structure. By unifying the determination of folding mechanism and prediction of structure, this work has positive implications for both areas. Because no knowledge about the native state is required, we can predict non-native kinetic traps and structures. Our nature-inspired computational search strategy can benefit the prediction of larger proteins, one of the major frontiers of the field. Finally, our work is equally applicable to fast or slow folding proteins and thus provides a suitable alternative for cases that are outside the range of current MD simulations. Moving folding proteins and thus provides a suitable alternative for cases

Methods

2° Structure-Fixing Protocol. The frequencies of helix, strand, and coil structure, as determined by the Dictionary of Protein Secondary Structure (44), are used to update the consensus 2° structure assignments. At each position, one of the three types is eliminated as a sampling option when its frequency falls below a threshold; e.g., remove helix if frequency lies below 1%. The consensus 2° structure restricts the ϕ, ψ sampling library employed in the subsequent folding round (Fig. 52).


Supporting Information

Adhikari et al. 10.1073/pnas.1209000109

SI Text

SI Methods. Monte carlo simulated annealing minimization. Using the algorithm described by Aarts and Korst (1), the Monte Carlo Simulated Annealing (MCSA) procedure for each round starts with a assignment of dihedral angles based on the Ramachandran Maps for each amino acid. Moves and energy functions are described below. The structure converges to a local energy minimum by gradually diminishing the annealing temperature, which controls the fraction of allowed moves that increase the energy of the system. A fixed number of moves are computed in order to achieve thermal equilibrium at each annealing temperature. The annealing temperature is decreased using the formula,

\[ t = \frac{t'}{1 + \frac{t' \log 2}{\sigma_t}} \]  

[S1]

where \( t' \) and \( t \) are the old and new annealing temperatures, respectively; \( \sigma_t \) is the standard deviation for the energy distribution at temperature \( t' \); and \( \delta \) is a tunable parameter for the cooling schedule. The temperature is decreased according to a Cauchy cooling schedule, until either convergence is reached or the total number of Monte Carlo steps reaches a specified value. The convergence criterion is based on the magnitude of the energy fluctuations for each annealing temperature, and annealing stops when the inequality,

\[ \sigma_t \leq \varepsilon t \]  

[S2]

is satisfied, where \( \varepsilon \) is the convergence tolerance, also a tunable parameter. The tunable parameters in the simulations are set to as: \( \delta = 0.05 \) and \( \varepsilon = 1 \). After every move, the energy of the new conformation is evaluated, and the change is accepted with probability,

\[ P = \min\{1, e^{-\Delta E/t}\} \]  

[S3]

where \( \Delta E \) is the energy difference between the new and old conformations.

Sampling using ramachandran maps. In the backbone plus C\_\beta representation used in TerIIFix, only backbone \( \phi, \psi \) angles are changed during the simulation, with all other angles and bond lengths fixed at their ideal values. Backbone conformations are sampled using \( \phi, \psi \) distributions, or Ramachandran maps, obtained from high-resolution PDB structures (resolution <2.5 Å, homology below 90%). Individual Ramachandran maps are generated for each position conditional on both the amino acid and its nearest neighbors’ chemical identity and, and their \( \theta \) structure specification (Fig. S2).

Because TerIIFix proceeds by eliminating \( \theta \) structure options at a given position, six possible categories of \( \theta \) structure are considered for the construction the sampling distributions (H: helix, E: strand, C: coil, A: everything, O: not helix, and Q: not strand). This description requires a total of 1,728,000 possible Ramachandran maps for the 20 possible amino acid triplets. Each Ramachandran map is divided into 72 bins, with each of these 5 \times 5 bins assigned a probability determined by frequency of \( \phi, \psi \) angles in a set of high-resolution structures appropriate for amino acid identities and their \( \theta \) structure. The \( \theta \) structure is ascertained from structures obtained in the prior round, as described in the next section. In the initial round, the \( \theta \) structure for each amino acid is assumed to be in a coil conformation and the Ramachandran maps are generated accordingly.

The sampling during the MCSA simulation is determined by the probabilities prescribed by a particular Ramachandran map. A particular bin in the Ramachandran map is assigned a probability determined by frequency of \( \phi, \psi \) angles in a set of high-resolution structures appropriate for amino acid identities and their \( \theta \) structure. The \( \theta \) structure is ascertained from structures obtained in the prior round, as described in the next section. In the initial round, the \( \theta \) structure for each amino acid...
Energy functions. The energy function contains three statistical potentials as well as two biasing terms:

\[ E_{\text{MC-SA}} = w_1 \cdot E_{\text{DOPE-PW}} + w_2 \cdot E_{\text{Burial}} + w_3 \cdot E_{\text{desolvation}} + w_4 \cdot E_{\text{H-bond}} + w_5 \cdot E_{\text{Contact}} \]

The relative weights, \(w_i\), of the different energy terms for the two stages are listed in Table S1, which were derived semiempirically. The biasing energies are more heavily weighted in the first stage.

Residue-residue interaction potential. DOPE-PW is a pairwise additive statistical potential based on observed distance distributions in the PDB between all atoms in our Cα representation. It is designed to reflect the basic protein structural principles of chemical complementarity, packing, and 2° conformation (Fig. S4). As detailed in our previous studies (2, 3), each interaction in DOPE-PW is classified according to atom type, residue type, secondary structure assignment, and side-chain orientation defined by the Cα-Cβ vector. The orientational dependence is characterized using the relation,

\[ \rho = \sqrt{(\rho_{12} - 90)^2 + (\rho_{21} - 90)^2} \]  

where \(\rho_{12}\) is the angle between the Cα-Cβ vector of residue 1 and the Cα-Cβ vector from residue 1 to residue 2. Similarly, \(\rho_{21}\) is the angle between the Cα-Cβ vector of residue 2 and the Cα-Cβ vector from residue 2 to residue 1 (Fig. S4). The \(\rho\) value is binned into three orientational regions: low (\(\rho < 40°\)), medium (40° < \(\rho < 70°\)) and high (\(\rho > 70°\)), which are accordingly parameterized from a high-resolution database of PDB structures. Fig. S4 illustrates a sample DOPE-PW energy profile for ILE-LEU Cβ atoms for high and low values of \(\rho\).

Burial energy. \(E_{\text{Burial}}\) energy is obtained from the burial properties for each of the 20 aa’s from a database of approximately 1000 single-domain, high-resolution PDB structures with sizes ranging from 50-150 aa (Fig. S4). The burial level for residue type \(i\) is determined from the number of heavy atoms, \(N_i\), around the Cβ in a 11 Å radius hemisphere in the direction defined by the Cα-Cβ vector. The burial energy contribution of \(N_i\) is given by

\[ E_{\text{Burial}}(N_i) = -\ln \left( \frac{f(N_i)}{f(N_{\text{any}})} \right) \]

where \(f(N_i)\) and \(f(N_{\text{any}})\) are frequencies for residue type \(i\) and any residue type, respectively.

**Backbone desolvation.** A similar term is calculated for hydrogen bonded and nonhydrogen bonded amide nitrogen and carbonyl oxygen atoms. The number of heavy atoms is calculated in 5 Å radius hemisphere in the direction defined by the N-H and C=O vectors, respectively. The N-H and C=O hydrogen bond energy terms are specified as whether or not they form protein-protein hydrogen bonds, and an energy penalty is assigned to NH and CO groups with unsatisfied H-bonds (Fig. S4). A hydrogen bond is defined for H-O distances of <3.5 Å and H-O=C angles of >120°. Sample profiles of the burial energies for LEU and GLU Cβ atoms are shown in Fig. S4, along with the NH burial penalty energy profile.

The present approach differs from our earlier ItFix protocol in which the local sampling distribution is fixed (1, 2) but 3° contacts and hydrogen bonds are not explicitly biased. We now include peptide group desolvation and burial potentials in lieu of a globularity term.

**Biased energies used for sequential stabilization.** \(E_{\text{Contact}}\). After each round, the average contact matrix from the ensemble of the lowest 25% energy structures is used to generate the contact bias energy term for the next round,

\[ E_{\text{Contact}} = \sum_{i=1}^{n} \sum_{j=i}^{n} -w_{ij} \]

where \(w_{ij}\) is the probability of contact between residues \(i\) and \(j\) (Cβ separation <7.5 Å).

\(E_{\text{H-bond}}\). The hydrogen bonding energy bias is

\[ E_{\text{H-bond}} = \sum_{i=1}^{n} \sum_{j=1}^{n} \left[ H_{\text{max}} \cdot p_{ij} + H_{\text{min}} \cdot (1 - p_i) \right] \]

where \(p_{ij}\) is the probability that the NH of residue \(i\) bonds to the CO of residue \(j\), and \(p_i\) is the probability that the NH hydrogen bonds to any other residue in the current round. The weights \(H_{\text{max}}\) and \(H_{\text{min}}\) are set to –25.0 and –2.0. Since a hydrogen bond is generally favorable, the above functional form ensures a contribution to the hydrogen bonding energy even when the specific probability \(p_{ij}\) is low.

Software. The simulation protocol is implemented in C, and input/output is managed with PDB tools from the BioPython package. To facilitate job submission in high performance supercomputing architectures, the programs are wrapped using the parallel scripting language, Swift (4). Simulations are performed using the PADS and Beagle clusters at the Computation Institute of the University of Chicago.

The algorithm iterates through multiple rounds of MCSA simulations, where information about 2° and 3° structure from prior rounds is utilized. A single MCSA simulation involves a two-stage energy minimization: First using pivot moves, followed by a refinement stage using doublecrank moves.

Fig. S1. Flowchart outlining TerItFix algorithm. (A) The algorithm iterates through multiple rounds of MCSA simulations, where information about 2° and 3° structure from prior rounds is utilized. (B) A single MCSA simulation involves a two-stage energy minimization: First using pivot moves, followed by a refinement stage using doublecrank moves.

Fig. S2. Restriction of the backbone sampling distribution. The distribution becomes more refined as the dependence on type and secondary structure of the aa and its neighbors are included.
Fig. S3. Two kinds of backbone torsional angle moves are used in TerItFix. The pivot move involving two angle is effective in exploring large regions of conformational space, while doublecrank changes four angles while minimally perturbing the rest of the protein structure.

Fig. S4. Statistical potentials. Examples of the energy profiles for the three statistical potentials used in TerItFix. All simulations use the same DOPE-PW parameters except for Ub, where a noncontinuous parallel strand is set higher to reduce the non-native antiparallel β1-β3 pairing. A burial penalty profile similar to NH group is obtained for the C=O group as well.

DOPE-PW
\[ \rho = \sqrt{\rho_{\phi_2-90}^2 - \rho_{\psi_2-90}^2} \]

DOPE-PW profile for ILE-LEU Cα-Cγ atoms

Burial Energy Profile for LEU, GLU

Penalty for burial of an unsatisfied NH group

Adhikari et al. www.pnas.org/cgi/doi/10.1073/pnas.1209000109
Fig. S5. Evolution of interhelical contacts in λ repressor as a function of TerItFix rounds.

Predictions after last round (R3)

Im7: Folds via H1-H2-H4 intermediate

Native

Secondary Structure

Contact maps

Im9 – Folds directly to native w/o intermediate accumulating

Native

SIm9: Folds via H1-H2-H4 intermediate

Native

Fig. S6. TerItFix prediction of folding intermediates in Im proteins. The lowest Cα-RMSD value is indicated.
**Fig. S7.** (A) Evolution of contact maps in Im7, Im9 and SIm9. H1-H2 contacts (red ovals) in Im7 and SIm9 are more populated than in Im9. (B) H1-H2 contact frequency in rounds R1 and R2, calculated by summing the contact probabilities between the H1-H2 in the low energy population in each round. (C) The interaction energy between V71I and the rest of the sequence strengthens the H1-H4 contacts in SIm9.

**Fig. S8.** Folding pathway for Ub. Secondary structure and tertiary contacts becomes increasingly determined through the six rounds of folding.
Fig. S9. Folding pathway for CI2.

Fig. S10. Folding pathway for Protein L.
Fig. S11. Folding pathway for α3d. Helices H2 and H3 form more extensively than H1 in the early rounds.

Fig. S12. Folding pathway for Protein A.

Fig. S13. Transition state ensemble for Protein L and Protein A selected by clustering structures with 0.6 < RCO/RCO_{Native} < 0.8.
<table>
<thead>
<tr>
<th>Parameters</th>
<th>Stage 1</th>
<th>Stage 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MCSA parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total MC Steps</td>
<td>50,000</td>
<td>100,000</td>
</tr>
<tr>
<td>Temperature Update Interval (Steps)</td>
<td>400</td>
<td>500</td>
</tr>
<tr>
<td>Initial Temperature</td>
<td>120</td>
<td>150</td>
</tr>
<tr>
<td><strong>Energy Function Weights</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOPE-PW, $w_1$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Burial, $w_2$</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Desolvation, $w_3$</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>H-Bond Bias ($E_{\text{H-bond}}$, $w_4$)</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Contact Bias ($E_{\text{contact}}$, $w_5$)</td>
<td>100</td>
<td>40</td>
</tr>
<tr>
<td><strong>MoveSet</strong></td>
<td>Pivots then doublecarnk</td>
<td>Pivots then doublecarnk</td>
</tr>
</tbody>
</table>