Protein folding by zipping and assembly

S. Banu Ozkan*†, G. Albert Wu*‡, John D. Chodera§¶, and Ken A. Dill*‖

*Department of Pharmaceutical Chemistry and §Graduate Group in Biophysics, University of California, San Francisco, CA 94143

Communicated by Carlos J. Bustamante, University of California, Berkeley, CA, May 2, 2007 (received for review April 13, 2006)

How do proteins fold so quickly? Some denatured proteins fold to their native structures in only microseconds, on average, implying that there is a folding "mechanism," i.e., a particular set of events by which the protein short-circuits a broader conformational search. Predicting protein structures using atomically detailed physical models is currently challenging. The most definitive proof of a putative folding mechanism would be whether it speeds up protein structure prediction in physical models. In the zipping and assembly (ZA) mechanism, local structuring happens first at independent sites along the chain, then those structures either grow (zip) or coalescence (assemble) with other structures. Here, we apply the ZA search mechanism to protein native structure prediction by using the AMBER96 force field with a generalized Born/surface area implicit solvent model and sampling by replica exchange molecular dynamics. Starting from open denatured conformations, our algorithm, called the ZA method, converges to an average of 2.2 Å from the Protein Data Bank native structures of eight of nine proteins that we tested, which ranged from 25 to 73 aa in length. In addition, experimental Φ values, where available on these proteins, are consistent with the predicted routes. We conclude that ZA is a viable model for how proteins physically fold. The present work also shows that physics-based force fields are quite good and that physics-based protein structure prediction may be practical, at least for some small proteins.

protein structure prediction | replica-exchange molecular dynamics

There are two protein folding problems: one is physical and one computational. The physical problem is a puzzle about how proteins fold so quickly. In test-tube refolding experiments, protein molecules begin in a disordered denatured state (a broad ensemble of microscopic conformations) and then fold when native conditions are restored. On the one hand, folding must be stochastic: a protein's native structure is reached via many different microscopic trajectories from the broad ensemble of different starting denatured conformations. On the other hand, folding happens quickly, sometimes averaging only microseconds to reach the ordered native conformation (1). How does the process of searching and sorting through the protein's large conformational space of disordered states happen so rapidly? And how is the same native state reached from so many different starting conformations? This puzzle has been called "Levinthal's Paradox" (2). Even the simplest disorder-to-order transitions, like the crystallization of sodium chloride, take days. It follows that the conformational search, although stochastic, cannot be random.

The second folding problem is computational: predicting a protein's native structure from its amino acid sequence. Success in this area could lead to advances in computer-based drug discovery. Predicting protein structures has become increasingly successful (3–7). Most current protein structure prediction methods make some use of database-derived conformational preferences. However, for the following reasons, it would be desirable to achieve high-resolution protein structure prediction in models that are purely physics-based, i.e., those that do not rely on information contained in protein structure databases. First, it would put our understanding of protein structures and driving forces on a deeper and more physical foundation. For example, such methods could elucidate the physical routes of protein folding. Second, it would allow the prediction of non-native states, too, those that are of interest for protein folding kinetics and stability, or for the induced-fit binding of ligands, or other conformational changes.

A longstanding viewpoint has been that solving the physics problem of how proteins physically fold up can help to solve the computational problem of protein structure prediction. Some proteins require as little as microseconds to fold into their native structures, yet supercomputers cannot fold them, even in times requiring tens of years, so what insights are missing from our computer prediction methods? If we had sufficient insight about how proteins fold, could we use them to speed up protein structure prediction algorithms?

Historically, mechanistic insights into folding processes have come from experimental studies of folding kinetics on model proteins. It has been suggested that folding is hierarchical, that secondary structures form earlier than tertiary structures, and/or that secondary structures nucleate tertiary contacts (8–14). Some of these features have been observed in computer unfolding simulations (15).

However, experimentally derived folding routes are not sufficient to provide the kind of folding principle that is needed to inform conformational search algorithms. A folding route is a description of a sequence of events, typically the formation of secondary structures, that has been observed for one protein under one set of conditions. In contrast, a folding principle would involve a quantitative model that starts from any microscopic chain conformation, for any sequence, and predict fast routes to the native state. The latter requires vastly more information, and therefore is not derivable from the former.

Protein folding experiments have yet to definitively prove or disprove any particular folding mechanism because such experiments "see" only highly averaged ensemble structures, rather than microscopic trajectories. Hence, at the present time, the most definitive strategy for proving or disproving any putative folding mechanism is rooted in the statement of the folding problem itself: can a computer be taught to fold a protein rapidly by using a purely physics-based model?

To succeed at physics-based protein structure prediction, there have been two questions: (*i*) are the force fields good enough? and (*ii*) is the conformational sampling sufficient? There are well known problems with commonly used molecular mechanics force fields. AMBER94 is known to overstabilize helices, whereas AMBER96 favors extended structures (16, 34). OPLS/AA and GROMOS96

Author contributions: S.B.O. and G.A.W. contributed equally to this work; S.B.O. and K.A.D. designed research; S.B.O. and G.A.W. performed research; S.B.O. and J.D.C. contributed new reagents/analytic tools; S.B.O. and G.A.W. analyzed data; and S.B.O. and K.A.D. wrote the paper.

The authors declare no conflict of interest.

Abbreviations: ZA, zipping and assembly; ZAM, ZA method; REMD, replica-exchange molecular dynamics; GB, generalized Born; SA, surface area; CPU, central processing unit; PMF, potential of mean force; PDB, Protein Data Bank; FFB, force field's best; SH3, Src homology 3.

†Present address: Department of Physics, Arizona State University, Tempe, AZ 85287.

‡Present address: Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720.

¶Present address: Department of Chemistry, Stanford University, Stanford, CA 94305.

‖To whom correspondence should be addressed. E-mail: dill@zimm.compbio.ucsf.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0703700104/DC1.

have difficulty discriminating between the polyproline type II and β-strand basins in the Ramachandran maps of small peptides (17, 18). Yoda *et al.* (19) conducted multicanonical simulations of several small peptides (the α-helical C-peptide of ribonuclease A and the C-terminal β-hairpin of protein G) by using six common force fields (AMBER94, AMBER96, AMBER99, CHARMM22, OPLS/AA/L, and GROM0S96) and concluded that all of these force fields have different propensities to form secondary structures because of the differences in backbone torsional energies. In addition, the popular implicit solvation models have been shown to overstabilize ion pairs (20, 21) or too strongly favor the burial of polar amino acids (22).

However, there are some successes, indicating that the physical force fields are good. In an early milestone paper, Duan and Kollman (23) performed a microsecond molecular dynamics simulation of the 36-residue villin headpiece in explicit solvent starting from an unfolded conformation, reaching a collapsed state 4.5 Å rmsd from the NMR structure. Vila *et al.* (24), starting from a random configuration, folded the 46-residue protein A to within 3.5 Å using Monte Carlo dynamics with an implicit solvation model.

Some groups have recently achieved higher accuracies. The IBM Blue Gene group of Pitera and Swope (25) folded the 20-residue Trp-cage peptide in implicit solvent to within $\approx$ 1 Å by using 92 ns of replica-exchange molecular dynamics (REMD). With Folding@Home, a distributed grid computing system, Pande and coworkers (26–28) folded villin to a rsmd of 3 Å in a computational time of $\approx$300 μs, or $\approx$1,000 central processing unit (CPU) years. Villin is the largest protein that has been accurately folded to date by using a purely physics-based model, to our knowledge. However, we note that the studies by Pande and coworkers and Pitera and Swope were not protein structure predictions, but rather large-scale simulations exploring the nature of folding thermodynamics and kinetics. Larger simulations have also been conducted, but they involve unfolding from the experimental structure; for example, the replica-exchange simulation of the 46-residue protein A in explicit water on HP ASCI Q, one of the world's largest supercomputers (29). To our knowledge, no β-sheet protein beyond 20 residues has previously been successfully folded (30). However, importantly, the studies cited here do show that current state-of-the-art force fields are adequate for protein structure prediction, at least in the few small proteins tested so far.

The implication is that the main bottleneck to physics-based protein structure prediction is that conformational search methods are too slow. It is believed that stochastic simulation methods, Monte Carlo or molecular dynamics, for example, cannot reach sufficiently long time scales on current computers. Therefore, a putative folding principle would be found to be most useful and predictive if it specifies folding routes that could substantially speed up the computer-based prediction of native protein structures in physics-based simulations.

We ask here whether high-resolution protein structure prediction can be achieved in a physical model in more than one or two small proteins. We use the the AMBER96 force field as implemented in the AMBER 7 package (31), with the generalized Born (GB)/surface area (SA) implicit solvent model of Tsui and Case (32) and sampling (REMD) (33). Although there are well known flaws in various force fields, we found AMBER96 to be better balanced for various secondary structures than other force field/solvation models we tested with the solution model used here. In addition, of key importance here, we use a mechanism-based search strategy we call zipping and assembly (ZA), which, we believe, is the strategy that proteins use to fold. ZA samples only a very small fraction of the conformational space that traditional methods would otherwise sample; it is this mechanism-based searching that allows us to efficiently sample the relevant parts of conformational space.

According to the ZA mechanism, upon the initiation of folding conditions, an unfolded chain first explores locally favorable structures at multiple independent points within the chain. These local structures are conformational basins of low free energy, typically stabilized by one or two hydrophobic contacts and often containing small α-helical or β-turn structures. While only transiently stable on their own, such local structures can then recruit neighboring amino acids in the chain sequence to form additional contacts, growing individual local structures (zipping) or combining them by coalescence (assembly); in either case, the protein chain becomes increasingly ordered. This mechanism is supported by studies of lattice-model proteins showing that this type of nonexhaustive greedy searching can find globally optimal states for a large fraction of sequences and by studies of master-equation models showing consistency with Φ-value experiments (35, 36).

In the ZA method (ZAM), the chain is first chopped into 8- to 12-residue fragments with overlapping residues. Each segment subjected to 5 ns per replica of REMD (33) starting from a fully extended conformation. To sample the fragment conformations adequately, our REMD temperatures span from 270 to 690 K (37). We analyze the results by using weighted histogram analysis (38, 39). Most fragments sample a broad ensemble of structures, but some fragments form stable hydrophobic contacts with well formed turns or helical shapes, as determined from the potential of mean force (PMF) for each possible pair of hydrophobic residues in the segment. For the fragments with stable hydrophobic contacts, we then loosely enforce those contacts with added restraints, then grow the fragment by adding more residues in extended form. New REMD simulations are then performed on those larger fragments. A new PMF analysis is performed to see whether new hydrophobic contacts are formed. Such growth attempts are continued to identify additional stable contacts until no further such contacts can be found. Then the algorithm switches to fragment assembly, a process that brings together two or more pieces to attempt further structure formation (see *Methods* for details).

Table 1. Convergence properties (C_α rmsds)

Protein name	Length, Å	rmsd, Å	
		Experiment	Simulation
Protein A (1BDD) fragment (residues 11–56)	46	1.9	1.5
Albumin-binding domain protein (1PRB) fragment (residues 10–53)	44	2.4	2.2
α3D (2A3D)	73	2.85	2.9
Protein G (2GB1)	56	1.6	1.7
Ubiquitin (1UBQ) fragment (residues 1–35)	35	2.0	1.2
YJQ8 (Pin) WW domain (1E0N) fragment (residues 7–31)	25	2.0	1.0
FPB28 WW domain (1E0L) fragment (residues 6–31)	26	2.2	2.5
α-spectrin SH3 (1SHG) fragment (residues 6–62)	57	2.3	1.3
src-SH3(1SRL) (fragment residues 9–64)	56	6.0	3.8

Results and Discussion

Control Simulations Starting from Experimental Structures. Before we attempt to predict a native structure, we must first verify that the force field is an adequate model, i.e., that computer simulations do not drift away from the known experimental native structure under native conditions under extensive conformational sampling. The force field was tested on the proteins listed in Table 1. We conducted a number of simulations initiated from the experimental native structures, running REMD (33) for 10 ns per replica (details described in *Methods*). Although stability over the finite duration of such simulations does not prove that there is no more stable region of phase space, it is a minimal requirement that the force field must pass. To prove that a force field is not adequate, it would suffice to show that a protein that is known to be stable from experiments would unfold under the force field under native conditions.

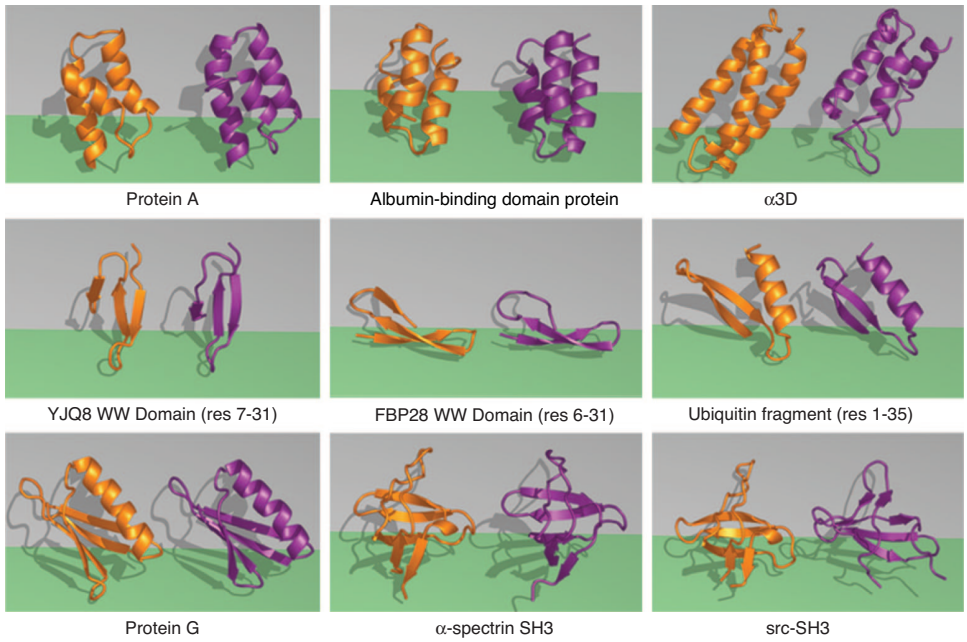


Fig. 1. Ribbon diagrams of the predicted protein structures using the ZAM (purple) vs. PDB structures (orange). The backbone C^{α} rmsds with respect to PDB structures are: protein A, 1.9 Å; albumin-binding domain protein, 2.4 Å; α3D, 2.85 Å (excluding the residues in the loops) or 4.6 Å; 1–35 residue fragment of ubiquitin, 2.0 Å; protein G, 1.6 Å; FBP26 and YJQ8 WW domains, 2.2 Å and 2.0 Å; and α-spectrin SH3, 2.2 Å. Our method fails to find the src-SH3 structure. Shown here is a conformation that is 6 Å from native. The problem in this case appears to be in the GB/SA implicit solvation model.

The main results are as follows. Seven of the nine proteins considered here did not drift >3 Å C^{α} rmsd away from the experimental starting structures over the course of the simulation, suggesting that this force field is adequate for these proteins. In the case of α3D, the molecule deviated from the Protein Data Bank (PDB) structure by up to 4 Å rmsd, but the mobility was mainly in the loop regions; the rmsd over the helical regions, excluding these loops, remained within 2.6 Å.

Comparison with the Experimental PDB Structures. We performed two kinds of tests. First, Fig. 1 (ZA vs. PDB) compares the structures predicted by ZAM with the corresponding experimental native structures from the PDB. ZAM produces an ensemble of structures. We use the centroid of the dominant cluster as representative. The comparison of ZA vs. PDB is a combined test of both the force field and the search method. Second, a more direct test of the search method alone is to compare ZA with the force field's best (FFB) structure (ZA vs. FFB). The FFB structure is a representative conformation taken from the most populous cluster that was reached by REMD simulations that were started from the experimental structure.

In general, the differences between ZA and FFB structures average only ≈1.6 Å rmsd, indicating that for those seven proteins, the search method has converged to the native basin of the force field (Table 1). For src-Src homology 3 (SH3), ZA fails to find any structure better than 5 Å from the experimental structure (ZA vs. PDB). In that case, the problem appears to be the GB/SA implicit solvation model. Similar problems have been observed before of ion pairs that are too stable in proteins having charged side chains (20).

Some Predicted Folding Routes. In our computational process, ZAM generates one or more folding routes, depending on the protein. In our simulations of protein A, helices 2 and 3 form first, then pack together, followed by the addition of the C-terminal helix 1. This result is consistent with the relative stabilities of the helices and intermediates that were found by Garcia and Onuchic in their explicit-solvent REMD simulations (29) and with experiments (41, 42).

ZAM finds that the albumin binding domain folds by first forming the C-terminal helix (helix 1), which then extends this helix. Helices 2 and 3 form independently, then assemble onto helix 1. The three helices of α3D form independently, and then assemble to form the full helix bundle.

The C-terminal 16-residue fragment of protein G is known from experiments to be stable by itself and has been studied by physics-based computational modeling (27, 43, 44). However, the full 56-residue protein has previously been beyond the range of high-accuracy predictions by molecular mechanics force fields. In our simulations, the folding of protein G (described in detail in *Methods*) begins with hydrophobic contacts forming in the N- and C-terminal β-turns, followed by hairpin formation. The C-terminal hairpin recruits the hydrophobic residues of the helical region to its core, causing the helix to form and pack against it. Finally, the N-terminal hairpin and the helix-C-terminal hairpin folded units assemble, completing the core.

We also studied the 35-residue N-terminal fragment of ubiquitin. In its ZA folding route, the 10-residue inner hairpin (Val-5–Thr-14) containing the β-turn is the first to form, followed by growth of the hairpin to its full length of 17 residues. The α-helix forms independently, then assembles onto the β-hairpin.

For both of the WW domains we studied (1E0N residues 7–31 and 1E0L residues 6–33), the two hairpins (β_1–β_2 and β_2–β_3) form independently at first, but the only successful folding route not ending up in a nonproductive trap involves adding the third strand (β_3) onto the β_1–β_2 hairpin.

The α-spectrin SH3 domain contains five antiparallel β-strands packed to form two perpendicular β-sheets. Folding begins by zipping the three-stranded β_2–β_3–β_4 sheet, followed by the addition of the seven-residue diverging turn (DT) and four more residues to the end of the DT, before a favorable hydrophobic contact between the RT loop and β4. Then the chain zips from the N terminus to form hydrophobic contact Tyr-15–Met-25 within the RT loop. The rest of the chain is zipped up in the last step, completing the structure. Folding steps for three additional proteins (α3D, protein

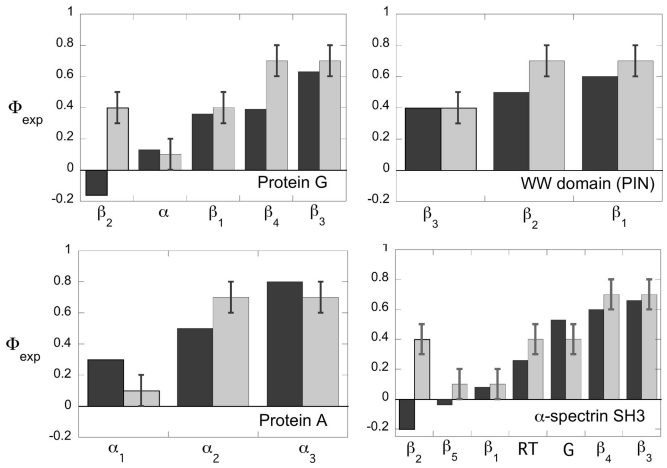


Fig. 2. Experimental average Φ values (black bars) and estimated kinetic impact values (gray bars) based on ZA folding routes for protein G, WW domain (Pin), protein A, and α-spectrin SH3. The kinetic impact value ranges are high (0.6–0.8), medium (0.3–0.5), and low (0–0.2).

A, and α-spectrin) can be found in supporting information (SI) Fig. 4.

Kinetic Impact vs. Experimental Φ Values. To determine whether our ZAM folding routes are consistent with experiments, we computed Thomas Weikl's (45) kinetic impact quantity for each of the secondary structural elements based on stability and order of emergence of the units. According to his quantity, secondary structures are assigned a high kinetic impact value if those structures form early in the folding process and lead to the formation of other structural units, a low kinetic impact value if they form late, and an intermediate value if they form at some stage of folding in between. The relative stability of a structure is also taken into account: when two structures form in parallel in a kinetic model, a medium kinetic impact value is assigned to the one that is less stable. Following Weikl, we do not attempt finer discrimination here than high, medium, and low. To compare kinetic impacts with averaged experimental Φ values, we use the following values for the kinetic impact factor: high (0.6–0.8), medium (0.5–0.3), and low (0.2–0). We compared the computed kinetic impact with average experimental Φ values for their secondary structural elements (see Fig. 2). We applied the method to protein A, the Pin (YJQ8) WW domain, α-spectrin SH3, and protein G.

Based on these simple rules, the kinetic impact values that we computed from our ZA simulations for α-spectrin SH3 are as follows: the three-stranded β-sheet β_2–β_3–β_4 rapidly forms first by zipping, and, because β3 and β4 are more stable than β_2, β_3 and β_4 are assigned a high kinetic impact value and β_2 a moderate value. The formation of β_2–β_3–β_4 leads to the diverging turn and RT turns, which have moderate experimental Φ values. The strands β_1 and β_5 form last in our simulations, so they have low kinetic impact values, also consistent with experimental Φ values.

The kinetic impact values computed from the ZA folding routes correlate well with average Φ values measured for the proteins tested (see Fig. 2). The exceptions are that the β_2 strands of protein G and α-spectrin are estimated to have moderate kinetic impact, whereas the experimental average Φ values are negative, although negative values also imply kinetic importance in the folding pathway (46). The good general correlation between kinetic impact factors and experimental Φ values suggests that the ZAM routes in our simulations are consistent with experiments. However, as noted above, the experiments are, by their nature, much too highly ensemble-averaged to prove or disprove that the ZAM routes are physically correct.

Comparison with Other Protein Structure Prediction Methods. David Baker and colleagues (7), using their Rosetta algorithm, have recently achieved an important milestone in protein structure prediction. Their predictions are "high resolution," which we define to mean: (*i*) a backbone rsmd from the experimental structure over the entire protein (rather than just selected parts) of <3 Å, (*ii*) achieving this level of experimental agreement routinely (i.e., for a significant fraction of proteins tested), rather than rarely, and (*iii*) consistent performance over different classes of folds. For 10 of 16 proteins <85 aa in length, Baker's group (7) recently reported the prediction of native structures to 3 Å or better, averaging 4.7 Å over the entire set of 16. The 3-Å threshold is important because at this level computer-based models of proteins may be as good as experimental x-ray or NMR structures for initiating drug discovery (47). The Rosetta high-resolution method requires ≈0.5 CPU years to predict each protein structure. It appears to represent the state of the art in protein structure prediction for modeling that incorporates database-derived insights and does not require a template having high sequence identity.

For comparison, straightforward molecular dynamics simulations combined with molecular mechanics force fields have folded small proteins without using database-derived information. For example, the Pande group (26–28) has used the distributed computing platform Folding@Home to fold the 16-residue C-terminal β-hairpin from protein G and the 36-residue villin headpiece to within 3 Å backbone rsmd of the experimental NMR structures; Pitera and Swope (25) have used REMD to fold the 20-residue Trp-cage miniproteins; and the Simmerling group (30) has simulated the folding of a designed three-stranded β-sheet by using multiple independent trajectories. The aim of those studies using explicit water was not rapid prediction of protein structures. We have shown here that the ZAM algorithm predicts native structures for eight of nine proteins tested, up to 74 aa in length, to an average of 2.2 Å from their experimental structures, using just a physics-based force field without database-derived preferences.

Conclusions

Here, we have explored ZA, which is both a hypothesis about the routes by which a protein folds and a mechanism-based conformational search method for predicting the native structures of proteins from their amino acid sequences. We give the most extensive evidence to date that physics-based force fields are adequate for protein structure prediction. We show that physics-based methods can achieve the same high resolution currently found otherwise only in the best bioinformatics methods, at least for small proteins. However, so far, we have tested only nine proteins, so we do not yet know whether the method handles larger proteins. We have not tested it in blind tests such as the CASP protein structure prediction event, and we believe that limitations of the force field will be problematic in some cases, as we found for the src-SH3 domain.

The ZA model can explain how physical protein folding can be so efficient, despite the diversity of microscopic trajectories and differences among protein structures. It hypothesizes that proteins use a divide-and-conquer strategy: small local independent peptide fragments of the chain establish conformational preferences, upon which further structure then grows and assembles. The premise is that the earliest time scales of folding are too short for the chain to explore more nonlocal aspects of its conformational space, and thus that early-stage folding is a greedy process of minimal conformational entropy loss per step. Those local structures that have sufficient metastability on the fast time scales are then able to grow and assemble increasing amounts of structure on longer time scales.

Because no experimental method is yet available that captures sufficient microscopic detail to give the relative probabilities of the different microscopic folding trajectories, the best evidence for a folding mechanism is simply its utility: can the method predict routes that speed up computer-based protein folding? The ZA conformational sampling method is substantially faster than

straightforward Monte Carlo or molecular dynamics simulations. Using ZA, protein G, which has 56 aa, folds computationally in ≈360 CPU days, or ≈1 CPU year, on a single 2.8-Ghz Xeon Intel machine.

Methods

Molecular Mechanics Model and Simulation Protocol. REMD (33) was used to sample the conformation space during the growth and assembly stages of the ZA algorithm. REMD periodically attempts to exchange conformations between independent molecular dynamics simulations running in parallel at different temperatures, based on a Metropolis-like criterion. This allows individual replicas to heat up to overcome barriers and then cool back down to temperatures of interest. It has two main advantages: (*i*) REMD explores more conformational space than conventional molecular dynamics techniques (48) and (*ii*) REMD samples from the canonical ensemble at each temperature, giving estimates of free energies, not just energies.

Proteins and fragments were modeled with the AMBER96 force field (31) with a GB implicit solvent model (32) and a SA penalty term of 5 cal·mol^{-1}·Å^{-2}. All fragments were capped at the N and C termini with acetyl and *N*-methylamine blocking groups, respectively, to avoid undue influence from the zwitterionic termini. All simulations were conducted by using a custom Perl script wrapper around the sander program from the AMBER7 molecular dynamics package. Replica temperatures were exponentially distributed over the range 270 to 690 K, with the number of replicas chosen to give average exchange acceptance probabilities of ≈50%. Exchanges were attempted every picosecond, between which energy-conserving molecular dynamics was used with a 2-fs time step. Velocities were randomized from a Maxwell-Boltzmann distribution after each exchange attempt to ensure sampling from the canonical ensemble at the appropriate temperature.

The ZAM Algorithm. We illustrate ZAM by describing how it folds protein G (shown in Fig. 3). According to the ZA model, on the shortest time scales, the chain does not have time to explore more than a few local degrees of freedom in any chain segment. Thus, in the ZA strategy, small peptide fragments of the chain first search for metastable structures, independently of other segments, i.e., in the absence of the rest of the chain. The computer first determines the locations within the protein chain at which structure might preferentially begin to form.

Because of limitations in our computer resources, we have done this first step of parsing into peptides in two different ways. For a few of the proteins (i.e., protein G, protein A, and α-spectrin SH3 domain), the process has been random, systematic, and not guided by knowledge of the native structure. In those cases, the chain is chopped into overlapping fragments 8–12 residues in length (e.g., residues 1–12, 5–16, 9–20, etc.). The fragments are chosen to be eight residues in length unless it is necessary to extend the chain up to 12 residues in length to include two hydrophobic residues. We then determine whether each peptide finds a metastable structure, by the method described in ref. 49. We believe that this strategy may work in general, because 133 different peptides from six different proteins are found to be structured by using our same force-field approach (49). For other proteins, we accelerate the first stage by simply choosing sites that we guessed would be nucleation points, based on knowing the native structure. The ZA algorithm then found routes to the folded states from those initiation sites. The latter tests prove that there are routes to the native structure that we would have found from the systematic random chopping process described above, but it does not show whether there might have been alternative routes, and it does not show what dead-ends or kinetic traps or misfolded structures might have hindered the search, if we had allowed a broader set of nucleation sites.

After discarding the first 2.5 ns per replica to equilibration, most such peptide fragments are found to populate a broad ensemble of

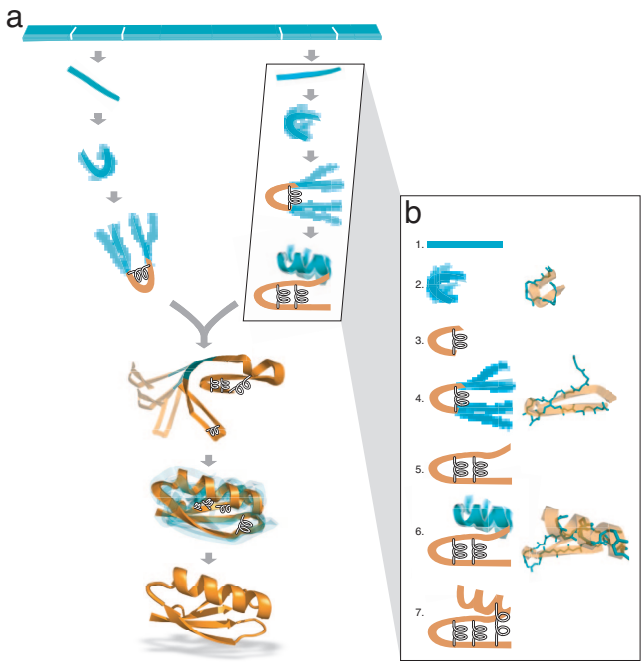


Fig. 3. ZA process for protein G. The chain is parsed into fragments of 8–12 residues. For each fragment, REMD simulation is performed for 5 ns per replica. PMFs are computed to determine whether a fragment is structured or unstructured. For protein G, the PMFs reveal that the C-terminal hairpin and the N-terminal β hairpin each form favorable hydrophobic contacts independently. For each segment that is structured, a spring is added to enforce that structure. Then new residues are added to the fragment ends ("growth") for another round of REMD simulation. For protein G, this results in the complete formation of both hairpins, with a helix packing onto the C-terminal β hairpin (*b*). When growth is no longer possible, as in protein G, the two folded units attempt to assemble, which, in this case, successfully leads to the native structure (*a*).

structures. Some fragments, however, have strong conformational preferences based on a coarse-grained ϕ–ψ angles of backbone representation called mesostrings (49). One can calculate mesostring entropy by using the Boltzmann formula $S = -k \Sigma_i p_i \ln p_i$, where p_i is the probability that the peptide is in the mesostring i. The fragments having low mesostring entropy can be considered as early folding nuclei to initiate the zipping (49).

Stable hydrophobic contacts within the fragments where the zipping is initiated are located by computing the PMF as a function of C^α distance between each possible pair of hydrophobic residues using weighted histogram analysis (38, 39). We identify all hydrophobic contacts that exhibit a substantially deep minimum in the distance range 0 to 8 Å in their PMF plots (see SI Fig. 5 for details). At this stage, evaluation of the PMF plots and the decision about which hydrophobic contacts needs to be restrained are not yet automated. The first tests show that automation can be done in two ways: (*i*) we compute the probability of a contact (i.e., integrating the PMF plots based on the contact distance ≥7 Å), hydrophobic residue pairs with contact probabilities >50% are considered as the contacts to be restrained; and (*ii*) we divide the PMF plot in two regions, the contact-forming region of distance separation (the region between 0 and 8 Å of distance separation) and contact breaking region (the region between 10 and 14 Å of the distance separation between two hydrophobic residues). We locate the minima in these two regions and compute the difference in free energy of these two minima. All contacts with contact free energies more stable than 2.0 kcal/mol are considered stable and become proposals for restraints.

To speed equilibration without affecting the final structure, we have found that seeding the REMD simulations with configura-

tions taken from the Baker I-sites library to be useful in accelerating the convergence of computed PMFs if the fragments appear in the library with >80% confidence (50). For protein G shown here, we did not use this acceleration process.

For protein G, two fragments from this initial stage, Tyr-45–Phe-52 and Ile-6–Gly-15, contain stable hydrophobic contacts. To explore whether these transient contacts are able to recruit additional contacts, we conduct a growth phase whereby the fragment is extended to incorporate additional hydrophobic residues. More simulation is conducted to compute a conditional PMF to determine whether additional contacts may be stable in the presence of the first ones. The contact identified in the previous step is restrained by using a potential applied to the distance between C_β atoms that is zero over the distance range [0,6.0) Å, harmonic with a force constant of 0.5 kcal·mol^{-1}·Å^{-2} over [6.0,6.5) Å and linear thereafter with continuous slope at 6.5 Å. In a small fraction of cases, such as α spectrin SH3, examination of all interresidue PMFs indicated that a hydrogen bond may have a smaller variance than a hydrophobic contact; in these cases, hydrogen bond distances are restrained instead. At each growth step, four additional residues are added in the extended configuration to each end of the fragment. Another set of REMD simulations is performed on all fragments that have been grown in this way, and a new set of conditional PMFs is computed. Such growth attempts are then repeated to identify additional stable contacts until no further such contacts can be found, at which point the algorithm switches to fragment assembly, described below.

In the case of protein G, the first fragment identified for growth initially spans residues 45–52 and contains contact Tyr-45–Phe-52. We then enforce this contact with a restraint as described above, and four additional residues are added to each end, so that the new fragment contains residues 41–56, the entire C-terminal β-hairpin. Similarly, the other fragment that has metastable structure, which includes the contact Ile-6–Gly-15, is subjected to the same treatment. Iterating this procedure leads to a highly structured segment (residues 28–56) containing a helix packing onto a strand and a β-hairpin in the N-terminal segment (residues 1–20).

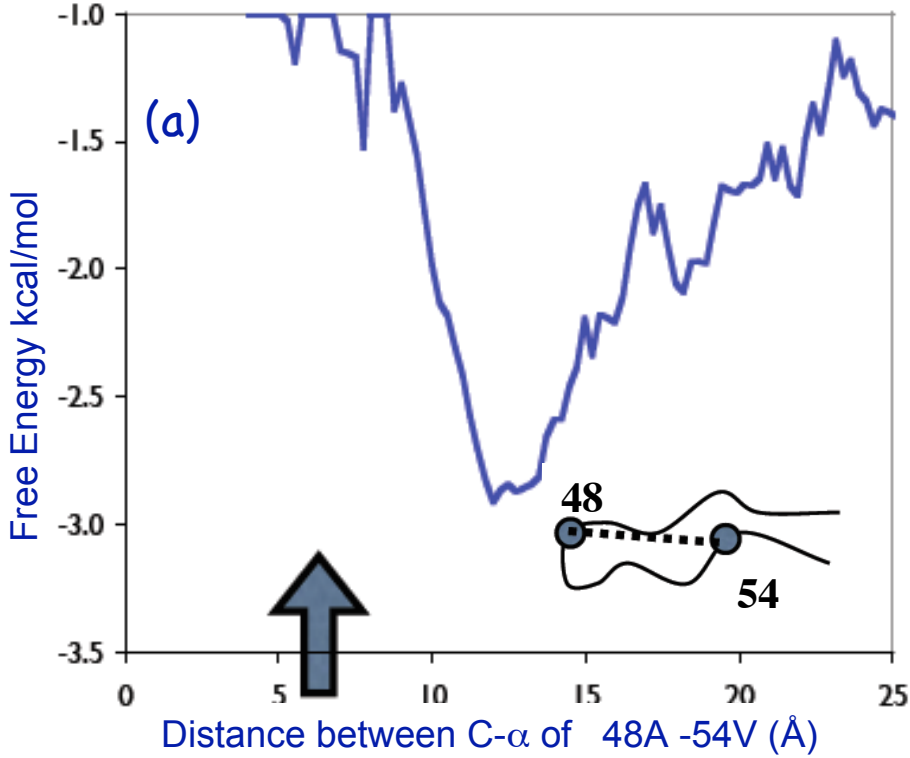
In some cases, for example, protein A and α-spectrin SH3, this zipping procedure alone is sufficient to reach the native state. In other cases, the fragments grow only to a point at which the addition of new residues to the segment forms neither additional structure nor favorable hydrophobic contacts. In those cases, the ZA algorithm then attempts to assemble the existing structures. To assemble, two nonoverlapping fragments are selected from the growth stage, the intervening residues are incorporated, and an ensemble is generated of different relative conformations of those pieces. A new REMD simulation is then initiated from this ensemble of configurations, and more simulations are conducted, retaining all of the previously imposed restraints, until new contacts and additional structure are formed, as determined by the PMF criterion.

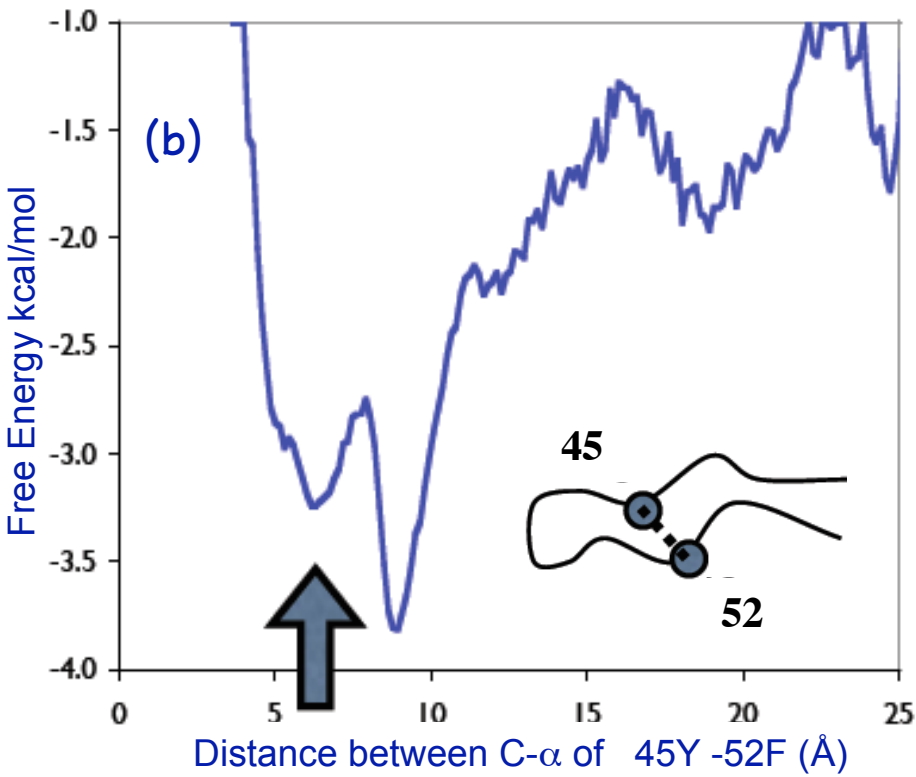
Either of two different methods has been found satisfactory for generating an ensemble of relative orientations of these fragments: (*i*) the anisotropic network model (ANM) or (*ii*) uniformly sampling the backbone torsion angles of the intervening chain before a new round of REMD sampling. ANM begins with coordinates of a structure, connects adjacent residues, based on a cut-off distance, with a spring, and computes elastic models by using a Hessian connectivity matrix (40). We diagonalize this matrix, decompose it into its eigenvectors, and take the two slowest (i.e., most global) eigenmodes. We add this (fluctuation) vector to the current atomic coordinates of the chain, up to the inverse force constant with a scaling factor of 10, to generate an ensemble of relative positions of the two pieces. Or, one residue near the center of the intervening chain in the assembly unit can be chosen randomly, from which we generate nine conformers by letting the ϕ and ψ angles each take on three values, -180 or $\pm 60°$.

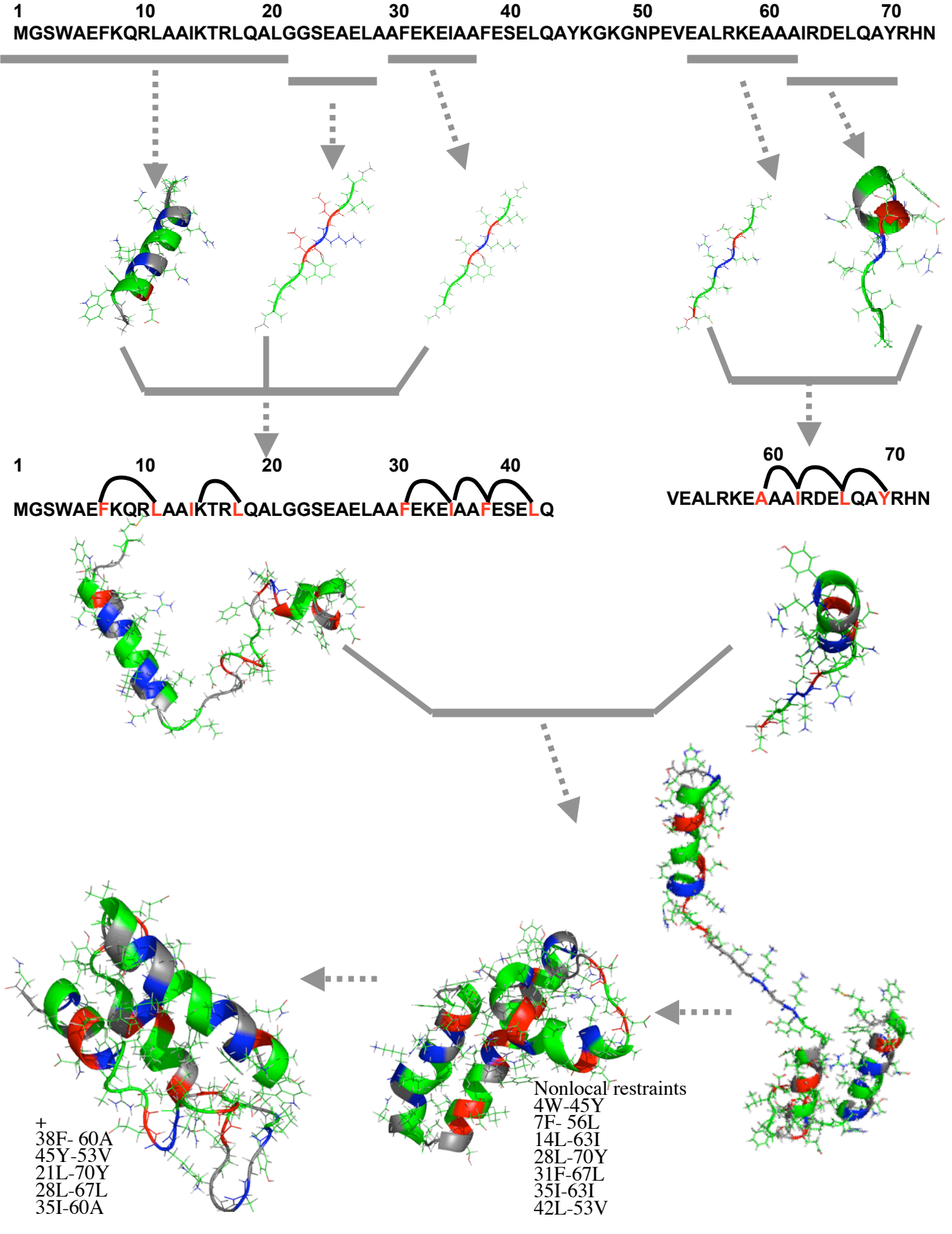
Protein G reaches a stage where growth terminates when the fragments span residues 1–20 and 28–56, so the algorithm then attempts to assemble these fragments by using an anisotropic network model. To speed up assembly, we add additional spring constraints either when PMFs of hydrophobic interactions indicate a contact, as described above, or when the side-chain centroid distance between a pair of hydrophobic residues is <10 Å and closing to half of its initial distance in 600 ps. We then restart the REMD simulation in the presence of these springs until convergence. For protein G, PMFs for possible hydrophobic contacts obtained by weighted histogram analysis indicate further pairing of nonlocal hydrophobic contacts between Leu-5–Phe-30, Tyr-3–Ala-26, Leu-5–Phe-52, Leu-5–Trp-43, Tyr-3–Phe-52, Leu-7–Val-54, and Ala-20–Ala-26. At the end of the process, all restraints are removed and an additional REMD simulation is conducted to ensure the resulting structure is stable on its own, irrespective of the ZA folding pathway.

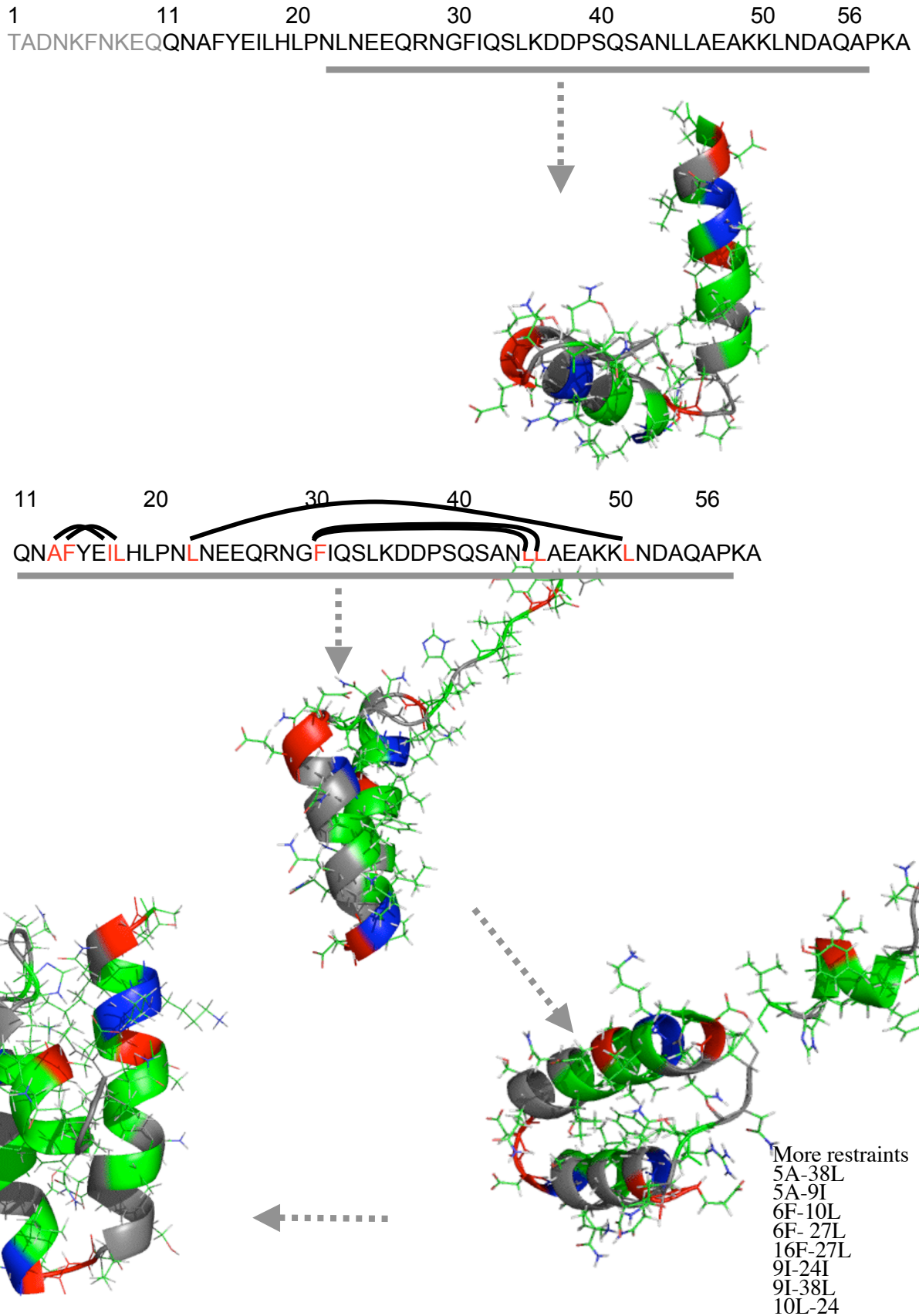
We thank B. Ho, V. Voelz, A. Narayanan, and K. Lau for assistance and I. Bahar, C. Camacho, R. Jernigan, H. Meirovitch, R. Baldwin, C. Dobson, V. Pande, and several reviewers for helpful comments. This work was supported by National Institutes of Health Grant GM34993 and the Sandler Foundation. J.D.C. was supported by Howard Hughes Medical Institute and IBM predoctoral fellowships. G.A.W. is supported by a National Institutes of Health National Research Service Award fellowship.

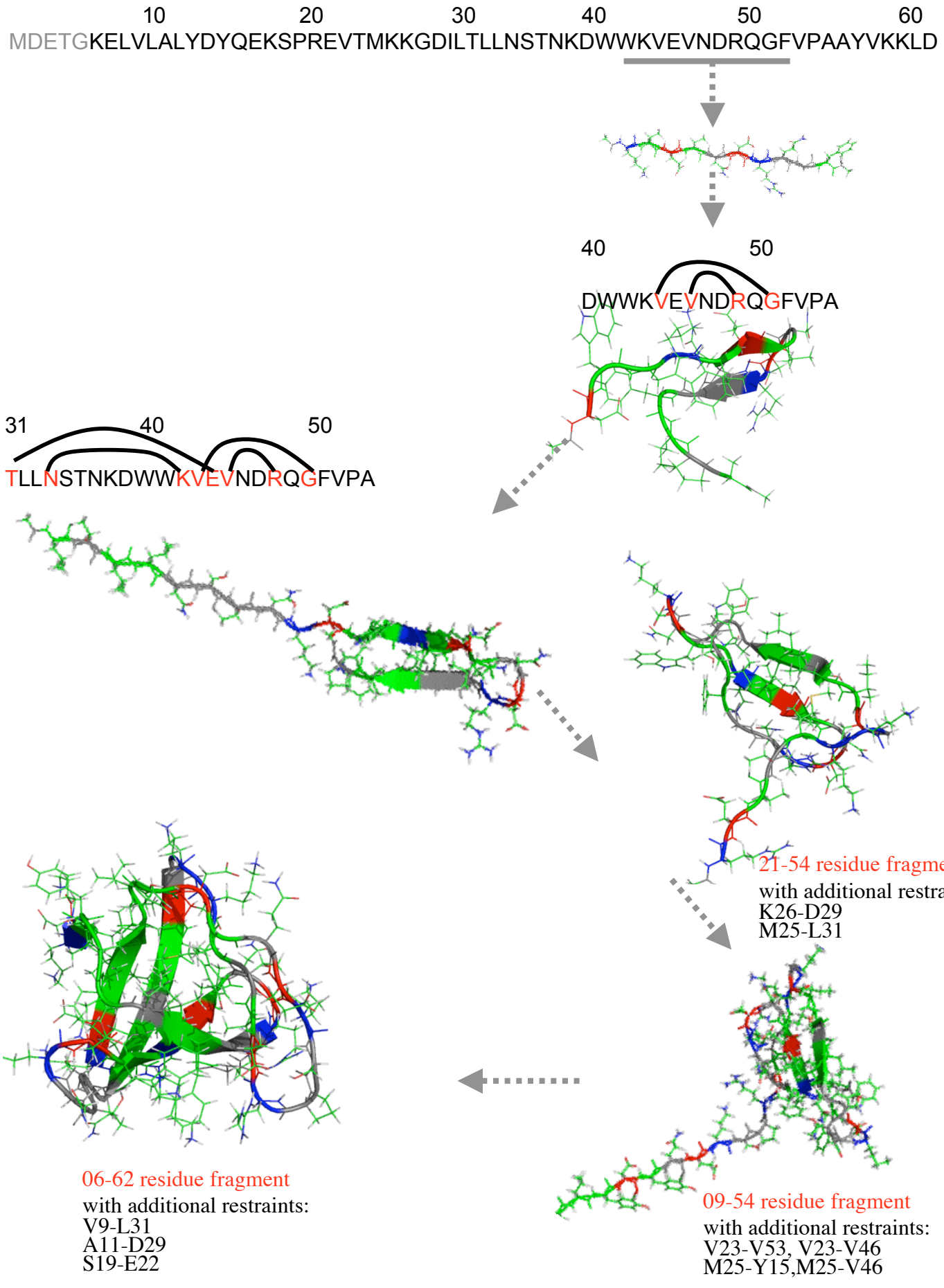
1. Kubelka J, Hofrichter J, Eaton WA (2004) *Curr Opin Struct Biol* 1:76–88.
2. Levinthal C (1968) *Extrait J Chimie Phys* 1:44–45.
3. Baker D, Sali A (2001) *Science* 294:93–96.
4. Liwo A, Khalili M, Scheraga HA (2005) *Proc Natl Acad Sci USA* 102:2362–2367.
5. Oldziej S, Czaplewski C, Schafroth HD, Kazmierkiewicz R, Ripoll DR, Pillardy J, Saunders JA, Kang YK, Gibson KD, Scheraga HA (2005) *Proc Natl Acad Sci USA* 102:7547–7552.
6. Hubner IA, Deeds EJ, Shakhnovich EI (2005) *Proc Natl Acad Sci USA* 102:18914–18918.
7. Bradley P, Misura KM, Baker D (2005) *Science* 309:1868–1871.
8. Karplus M, Weaver DL (1976) *Nature* 260:404–406.
9. Rose GD (1979) *J Mol Biol* 134:447–470.
10. Kim PS, Baldwin RL (1982) *Annu Rev Biochem* 51:459–489.
11. Dill KA, Fiebig KM, Chan HS (1993) *Proc Natl Acad Sci USA* 90:1942–1946.
12. Fersht AR (1997) *Curr Opin Struct Biol* 7:3–9.
13. Maity H, Maity M, Krishna MG, Mayne L, Englander SW (2005) *Proc Natl Acad Sci USA* 102:4741–4746.
14. Weikl TR, Dill KA (2003) *J Mol Biol* 329:585–598.
15. White GWN, Gianni S, Grossmann JG, Jemth P, Daggett V, Fersht AR (2005) *J Mol Biol* 350:757–775.
16. Gnanakaran S, Garcia AE (2003) *J Phys Chem B* 107:12555–12557.
17. Hu H, Elstner J, Hermans M (2003) *Proteins* 50:451–463.
18. Mu YS, Kosov DS, Stock G (2003) *J Chem Phys B* 107:5063–5073.
19. Yoda YT, Sugita Y, Okamoto Y (2004) *Chem Phys* 307:269–283.
20. Felts AK, Harano Y, Gallicchio E, Levy RM (2004) *Proteins* 56:310–321.
21. Zhou RH, Berne BJ (2002) *Proc Natl Acad Sci USA* 99:12777–12782.
22. Jaramillo A, Wodak SJ (2005) *Biophys J* 88:156–17.
23. Duan Y, Kollman PA (1998) *Science* 282:740–744.
24. Vila JA, Ripoli DR, Scheraga HA (2003) *Proc Natl Acad Sci USA* 100:14812–14816.
25. Pitera JW, Swope W (2003) *Proc Natl Acad Sci USA* 100:7587–7592.
26. Zagrovic B, Snow C, Shirts MR, Pande VS (2002) *J Mol Biol* 323:927–937.
27. Zagrovic B, Sorin EJ, Pande VS (2001) *J Mol Biol* 313:151–169.
28. Snow CD, Zagrovic B, Pande VS (2002) *J Am Chem Soc* 124:14548–14549.
29. Garcia AE, Onuchic JN (2003) *Proc Natl Acad Sci USA* 100:13898–13903.
30. Roe D, Hornak V, Simmerling C (2005) *J Mol Biol* 352:370–381.
31. Kollman PA, Dixon R, Cornell W, Vox T, Chipot C, Pohorille A (1997) *The Development/Application of a "Minimalist" Organic/Biochemical Molecular Mechanic Force Field Using a Combination of ab Initio Calculations and Experimental Data* (Kluwer, Boston), Vol 3, pp 83–96.
32. Tsui V, Case DA (2000) *J Am Chem Soc* 122:2489–2498.
33. Sugita Y, Okamoto Y (1999) *Chem Phys Lett* 314:141–151.
34. Zhou R (2003) *Proteins* 53:148–161.
35. Weikl TR, Dill KA (2003) *J Mol Biol* 332:953–963.
36. Voelz VA, Dill KA (2007) *Proteins* 66:877–888.
37. Klimov DK, Thirumalai D (2005) *J Mol Biol* 353:1171–1186.
38. Kumar S, Bouzida D, Swendsen RH, Kollman PA, Rosenberg JM (1992) *J Comput Chem* 13:1011–1021.
39. Chodera JC, Swope WC, Pitera J, Seok C, Dill KA (2007) *J Chem Theor Comput* 3:26–41.
40. Atiligan AR, Durrell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I (2001) *Biophys J* 80:505–515.
41. Bai Y, Karimi A, Dyson HJ, Wright PE (1997) *Protein Sci* 6:1449–1457.
42. Sato S, Religa TL, Daggett V, Fersht AR (2004) *Proc Natl Acad Sci USA* 101:6952–6956.
43. Zhou RH, Berne BJ, Germain R (2001) *Proc Natl Acad Sci USA* 98:14931–14936.
44. Rao F, Caflisch A (2003) *J Chem Phys* 119:4035–4042.
45. Weikl TR (2005) *Proteins* 50:701–711.
46. Ozkan SB, Bahar I, Dill KA (2001) *Nat Struct Biol* 8:765–769.
47. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, Friesner RA (2004) *Proteins* 55:351–367.
48. Zhang W, Wu C, Duan Y (2005) *J Chem Phys* 123:154105–154113.
49. Ho BK, Dill KA (2006) *Plos Comp Biol* 2:1–10.
50. Bystroff C, Baker D (1997) *Proteins* 1(Suppl):167–171.


(a)
Free Energy kcal/mol
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
0
5
10
15
20
25
48
54
Distance between C-α of 48A -54V (Å)


(b)
Free Energy kcal/mol
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
0
5
10
15
20
25
45
52
Distance between C-α of 45Y -52F (Å)


1 10 20 30 40 50 60 70
MGSWAEFKQRLAAIKTRLQALGGSEAELAAFEKEIAAFESELQAYKGKGNPEVEALRKEAAAIRDELQAYRHN
1 10 20 30 40
MGSWAEFKQRLAAIKTRLQALGGSEAELAAFEKEIAAFESELQ
60 70
VEALRKEAAAIRDELQAYRHN
Nonlocal restraints
4W-45Y
7F- 56L
14L-63I
28L-70Y
31F-67L
35I-63I
42L-53V
+
38F- 60A
45Y-53V
21L-70Y
28L-67L
35I-60A


1 11 20 30 40 50 56
TADNKFNKEQQNAFYEILHLPNLNEEQRNGFIQSLKDDPSQSANLLAEAKKLNDAQAPKA
11 20 30 40 50 56
QNAFYEILHLPNLNEEQRNGFIQSLKDDPSQSANLLAEAKKLNDAQAPKA
More restraints
5A-38L
5A-9I
6F-10L
6F- 27L
16F-27L
9I-24I
9I-38L
10L-24


10
20
30
40
50
60
MDETGKELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFVPAAYVKKLD
40
50
DWWKVEVNDRQGFVPA
31
40
50
TLLNSTNKDWWKVEVNDRQGFVPA
21-54 residue fragment
with additional restraints:
K26-D29
M25-L31
09-54 residue fragment
with additional restraints:
V23-V53, V23-V46
M25-Y15,M25-V46
06-62 residue fragment
with additional restraints:
V9-L31
A11-D29
S19-E22