High-resolution protein folding with a transferable potential

Isaac A. Hubner*, Eric J. Deeds†, and Eugene I. Shakhnovich‡*

Departments of *Chemistry and Chemical Biology, and †Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138

Edited by Peter G. Wolynes, University of California at San Diego, La Jolla, CA, and approved November 7, 2005 (received for review March 16, 2005)

A generalized computational method for folding proteins with a fully transferable potential and geometrically realistic all-atom model is presented and tested on seven helix bundle proteins. The protocol, which includes graph-theoretical analysis of the ensemble of resulting folded conformations, was systematically applied and consistently produced structure predictions of ∼3 Å without any knowledge of the native state. To measure and understand the significance of the results, extensive control simulations were conducted. Graph theoretic analysis provides a means for systematically identifying the native fold and provides physical insight, conceptually linking the results to modern theoretical views of protein folding. In addition to presenting a method for prediction of structure and folding mechanism, our model suggests that an accurate all-atom amino acid representation coupled with a physically reasonable atomic interaction potential and hydrogen bonding are essential features for a realistic protein model.

Monte Carlo | all-atom model | folding simulation | structure prediction | graph theory

Protein folding is easy. Without effort, every living organism completes the process innumerable times. Unfortunately, modeling the process is notoriously difficult. Since Anfinsen’s experiment (1), we have known that a protein’s tertiary structure is defined by its primary sequence. However, the question of sequence-structure mapping remains unsolved. Although researchers in the field have risen to the challenge and continue to make incremental progress (2), a complete solution remains among the great outstanding problems in computational biology. The problem has two aspects. First, given a protein’s amino acid sequence, can one reliably predict its tertiary structure? Second, can one accurately understand and describe at a detailed atomic level the physical process by which a protein reaches its native conformation and the dynamics of the folded conformation?

Approaches to the protein-folding problem fall into two major categories: bioinformatics methods attempt to model the structure of a protein primarily through homology to known structures and methods that rely on modeling the physical process by which the polymer chain attains its native conformation. Although homology-based approaches have generally yielded more accurate structure predictions and are more readily applied to larger proteins (3), they do not provide physical insight into the folding or conformational dynamics of proteins. The “holy grail” of the protein-folding community thus remains a computationally efficient model that both accurately predicts structure and provides physical insight into the folding and function of any protein given only its amino acid sequence.

Over the past decade, various models have been applied to protein folding and structure prediction. In an important study, the 36 residue villin headpiece fragment was folded to ∼4-5 Å from the native structure, demonstrating that the dream of ab initio protein folding is becoming a reality (4). Other highly successful methods (5) combine sequence and structural homology with incremental physical model-building for structure prediction. Many detailed physical studies use computationally intensive molecular dynamics simulations with complex potentials such as CHARMM and AMBER. Although they provide a

Conflict of interest statement: No conflicts declared.

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

Abbreviations: GC, Giant Component; MC, Monte Carlo; PDB, Protein Data Bank; rmsd, rms deviation.

†To whom correspondence should be addressed: Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138. E-mail: eugene@belok.harvard.edu.

© 2005 by The National Academy of Sciences of the USA

PNAS | December 27, 2005 | vol. 102 | no. 52 | 18914–18919

www.pnas.org/cgi/doi/10.1073/pnas.0502181102
means of MC dynamics. An advantage of this method's computational efficiency is that it allows for hundreds of fully independent simulations, resulting in representative statistics and that allows one to test ensemble kinetics and thermodynamics. The simulation requires no knowledge of the protein's native structures and may be applied in a systematic prescribed manner to any amino acid sequence. Although hierarchical clustering has previously been used in protein structure prediction (23–25), we employ a different graph-theoretic approach that retains topological features of the relationships between members of a cluster and allows us to interpret the results in the context of landscape theories of folding, overcome noise in the potential, and identify high-resolution structure predictions from simulation.

The resulting predictions are significant by the criteria of the protein-folding community (2) and on estimates based on studies of structural homology (26). We also show that the results are meaningful when compared with a set of control simulations. Although conclusions regarding folding mechanism may be made from this model, we limit the present discussion to demonstrating the feasibility of folding and interpreting the graph-theoretic analysis in terms of landscape theory. The successes of the model, which was not optimized or parameterized to any specific protein or training set, are derived from a realistic representation of the topological effects of folding: chain connectivity and side-chain packing combined with a simple, physically reasonable two-body potential that governs specific collapse and a generic H-bonding potential that ensures secondary structure formation.

Model and Methods

Protein Representation and Dynamics. Simulations used structures from the Protein Data Bank (PDB). For randomized sequence controls, all-atom models were built from the FASTA (http://fasta.bioch.virginia.edu) sequence by using SWISS PDB VIEWER (www.expasy.org/spdbv). Nonhydrogen atoms are explicitly modeled as impenetrable hard spheres. The move set includes global and localized backbone moves and side-chain torsions. Bond length and connectivity, as well as excluded volume, are always maintained. The move set and MC simulation have been described in detail (27) and have been shown to behave ergodically and satisfy detailed balance (26, 27). Although the thermodynamics and kinetics of the μ-potential have not been calibrated in detail, cooperative two-state (single-exponential) folding and unfolding behavior is observed in all test proteins.

Form and Derivation of the Potential. We have adapted and modified a transferable, knowledge-based pairwise contact potential from earlier work (14) that has also been used in a hybrid potential to introduce physically realistic interaction in simulations of SH3 (29) of the form:

$$E_{AB} = \frac{-\mu N_{AB} + (1 - \mu)\tilde{N}_{AB}}{\mu N_{AB} + (1 - \mu)\tilde{N}_{AB}}$$  \[1\]

where A and B are two interacting atoms, $N_{AB}$ and $\tilde{N}_{AB}$ are the number of AB pairs in contact and not in contact in the database, and $\mu$ is a parameter balancing attraction and repulsion. As is clear from the above equation, the potential becomes the Go potential at the limit where the number of atom types goes to the number of atoms in a single protein. The (nonoptimized) value of $\mu$ (0.9979) was chosen such that the $\langle E_{AB} \rangle = 0$. To calculate $N_{AB}$ and $\tilde{N}_{AB}$, we use a database (30) of 103 proteins (see Data Set 1, which is published as supporting information on the PNAS website) with <25% sequence homology that are >50 and <200 residues and define a set of transferable atom types where backbone atoms are typed as peptide N, Cα, carbonyl C, and O regardless of residue. Each side-chain atom has its own type, with the exception of those atoms related by symmetry (methyl group carbons in Val, for example) yielding a total of 84 atom types (see Table 6, which is published as supporting information on the PNAS website). Importantly, all proteins in our test set and their homologues are excluded from the database used to compute the potential. These two-body interactions are represented by a square well potential, where atoms A and B with hard-sphere radii $r_A$ and $r_B$ separated by distance $d$ are in contact if $0.75(r_A + r_B) < d < 1.8(r_A + r_B)$. (Potential available from authors upon request.)

In addition to the above pairwise interaction potential, we consider a backbone H-bonding ($E_{HB}$, see Fig. 3, which is published as supporting information on the PNAS web site) function to ensure proper secondary structure formation. The relative strength of H-bonding and pairwise interaction is controlled by $\alpha$, which balances the forces of polymer elongation and collapse.

$$E_{total} = \alpha E_{HB} + (1 - \alpha)E_{AB}$$  \[2\]

To perform effective simulations, the relative energy scale between $E_{AB}$ and $E_{HB}$ must be set by $\alpha$. When $\alpha$ is very high, the total energy is dominated by H-bonding and extended helix conformations are formed. When $\alpha$ is too low, hydrophobic interactions, which are the strongest among $E_{AB}$, are overwhelmingly represented, leading to collapsed conformations with a well-packed hydrophobic core but without secondary structure. At extreme ($\alpha \sim 0$ or 1) values, protein behavior is sensitive to this parameter. However, it is possible to systematically identify an appropriate value of $\alpha$ by beginning with a high value and annealing until the majority of structures collapse to globular conformations (see Fig. 4, which is published as supporting information on the PNAS website). The values of $\alpha$ that induce collapse are 0.92 for PDB ID codes 1BDD, 1BA5, 1ENH, and 1GAB and 0.89 for 1GAB and 1GJS. Further tuning does not improve the results, and lowering $\alpha$ significantly beyond this point worsens the structures. As previously observed by other research groups (15), attempts at parameter optimization (in our case $\mu$ or $\alpha$) did not improve the results.

Simulation Protocol. First, the native PDB structure is unfolded for $10^8$ random (without energy but maintaining excluded volume) MC steps at $T = 1,000$ to create a random, fully unfolded (extended and without correlations to native $\phi$ and $\psi$ angles) starting conformation. Each folding simulation is initiated from an independent random conformation and propagated at $T = 1.75 \sim T_i$ for $10^8$ MC steps. Next, the minimum energy conformation from the folding simulation is annealed from $T = 1.75$ to 0 for $5 \times 10^7$ MC steps to improve side-chain packing. The minimum energy conformation from this refinement simulation is then collected as the structure prediction. The above protocol was repeated 400 times for each protein to create an ensemble of predictions.

Test Proteins. Seven independently folding domains of short length were selected to test the folding model (see Table 7, which is published as supporting information on the PNAS website). The albumin binding (‘‘up-down-up’’ three-helix bundle) topology, which includes: Staphylococcus aureus protein A, Ig-binding B domain (PDB ID code 1BDD), Escherichia coli albumin-binding domain surface protein (PDB ID code 1GAB), and Streptococcus IgG binding protein G (PDB ID code 1GJS). The DNA binding (‘‘helix-turn-helix’’ or homeodomain-type) topology, including: DNA binding domain of human telomeric protein HTRF1 (PDB ID code 1BAA), an engrailed homeodomain-DNA complex (PDB ID code 1ENH), and c-Myc R1 protooncogene (PDB ID code 1GUU). An FF domain protein (PDB ID code 1UZC), with a more
The “All” column describes the average and range of rmsd values for the E_{min} conformation from each of 400 independent simulations; GC values are calculated at the transition midpoint.

Table 1. The GC presents a significant enrichment of the data, eliminating misfolds and reducing the average rmsd while retaining the best predictions

| Protein ID | PDB Code | All (rmsd) | Range | GC (rmsd) | Range |
|------------|----------|------------|-------|-----------|-------|
| 1BDD       | 8.75     | 3.42–11.44 | 7.72  | 3.42–11.44 |
| 1GAB       | 7.26     | 2.56–11.99 | 6.35  | 2.56–9.86  |
| 1GJS       | 7.51     | 2.43–16.63 | 5.25  | 2.43–8.73  |
| 1BA5       | 8.33     | 3.79–16.76 | 6.31  | 3.79–8.94  |
| 1ENH       | 9.19     | 2.44–18.40 | 5.42  | 2.44–9.41  |
| 1GUU       | 8.07     | 3.28–13.95 | 7.14  | 3.28–10.06 |
| 1UZC       | 8.75     | 3.42–12.15 | 7.72  | 3.42–11.44 |

Table 2. Comparison of native and mirror misfold clusters

| Protein | GC Prediction |
|---------|---------------|
| ENH1    | 5.42          |
| ENH2    | 9.85          |
| GUU1    | 4.59          |
| GUU2    | 8.82          |

Table 3. rmsd of structure predictions from the GC by E and k

| Protein | Three lowest E | Three highest k |
|---------|----------------|-----------------|
| 1BDD    | 4.00, 4.25, 6.82 | 4.77, 5.55, 6.27 |
| 1GAB    | 7.47, 7.71, 8.20 | 3.07, 3.10, 3.41 |
| 1GJS    | 4.51, 5.07, 5.36 | 3.76, 3.97, 4.65 |
| 1BA5    | 5.82, 5.92, 6.59 | 5.63, 5.72, 5.85 |
| 1ENH    | 2.96, 4.63, 5.72 | 2.44, 2.96, 3.01 |
| 1GUU    | 7.73, 8.62, 9.17 | 4.05, 8.25, 8.99 |
| 1UZC    | 5.05, 6.10, 7.09 | 4.25, 5.22, 1.90 |

complex four-helix (one of which is 3_{10}) topology, also was included. 1ENH and 1GUU were compared with x-ray structures, whereas all others were compared with the NMR structures. We calculate rmsd for the helical and turns regions (ignoring long, disordered regions at the termini), corresponding to F6-A60 in 1BDD, N9-A53 in 1GAB, D16-A62 in 1GJS, L7-L53 in 1BA5, F8-I56 in 1ENH, T44-L86 in 1GUU, and K14-T96 in 1UZC. Limitation of the present work to shorter proteins is, at least in part, due to the sampling restrictions presented by a fixed-temperature MC search that, unlike replica exchange, may be more suitable to studying the folding mechanism.

Graph-Theoretical Analysis. The rmsd in Cα coordinates is computed for all pairs of the lowest-energy structures obtained from 400 independent simulations of each protein. A graph is then created from these comparisons by considering each minimum energy structure as a node and connecting any two nodes that exhibit a rmsd less than a particular cutoff (r). The clusters in this graph are defined as any set of nodes where a path exists in the graph between any two members of that set. These disjoint clusters are obtained by using a standard depth-first search algorithm. At any given value of r, the giant component (GC) of the graph is defined as the largest disjoint cluster. As has been observed in many systems, this GC undergoes a transition as a function of r (see Fig. 5, which is published as supporting information on the PNAS web site). We analyze the GC for each protein at the midpoint of the transition, the cutoff r at which half of the total structures are contained within the largest cluster.

Results and Discussion

Separating Native Folds from Misfolds. From examining the 400 independent trajectories for each protein, it is clear that the native state is well sampled. Most minimum energy conformations fall in the 2–6 Å range. Examining all minimum energy structures (Table 1), it is clear that along with native folds, there are a number of low energy decoys. Because of the approximate nature of our energy function, this result is not surprising. Previous computational studies also resulted in the minimum of various energy functions not corresponding exactly to the native state (3, 4, 11–15, 19, 20). Many of the decoys we observe involve undocked helices or poorly formed secondary structure. These misfolds usually exhibit higher energy than near-native structures and are thus easily identifiable. There is a second class of misfolds that are protein-like and are broadly describable as “mirror images” of the native structure. These misfolds represent a more difficult case because they have energies comparable with native-like structures despite high (8–10 Å) rmsd from the native state. Other researchers also have noted the presence of mirror misfolds, because a three-helix bundle exhibits twofold topologic degeneracy (3, 11, 15).

Given that our energy function cannot a priori distinguish these low-energy decoys from native conformations, we must rely on other objective analyses to identify native conformations. To accomplish this goal, we employ a graph-theoretic clustering procedure. The largest cluster in this graph, the GC, undergoes a sharp transition as r, the structural cutoff used to construct the graph, becomes more stringent. At the midpoint of the transition many of the decoy structures are excluded from the GC as evidenced by the decrease in average rmsd in the GC compared with the entire ensemble of structures at the midpoint of the transition in the GC (Table 1). In most cases the mirror misfold structures are excluded from the GC at this point in the transition. A representative graph (see Fig. 6, which is published as supporting information on the PNAS web site) for 1BDD at the midpoint in the transition is shown in Fig. 2. As evidenced by these graphs, in the predominance of cases no set of decoys obtained from these simulations forms a cluster that is as large and coherent as the native-like structures in the GC. This finding indicates that the predominant structural class sampled by our simulations and identified via clustering is the native basin.

The two exceptions to the above observation are 1GUU and 1ENH. In the case of 1GUU mirror misfolds remain in the GC at the midpoint of the transition. Visual inspection of the 1GUU graph at the transition (Fig. 6) clearly reveals that the GC in this graph consists of two distinct, dense clusters that are connected to one another by only a few edges. At a lower cutoff these two clusters break into a near-native cluster and a mirror misfold cluster (Table 2). In the case of 1ENH the misfolds form a coherent cluster separate from, but of almost the same size as, the GC (Fig. 6 and Table 2). In both of these cases our method cannot identify which represents the native cluster. Graph-theoretic analysis allows us to identify this problem when it does occur, however, providing an objective measure of the degeneracy in our sampling of structural space.

Identifying the Native Fold. Clustering the results of independent folding simulations improves the quality of the prediction by enriching the representation of the well-sampled native state and excluding the disparate misfolds. However, the size of the GC

| Protein PDB | ID code | Three lowest E | Three highest k |
|------------|---------|----------------|-----------------|
| 1BDD       | 4.00, 4.25, 6.82 | 4.77, 5.55, 6.27 |
| 1GAB       | 7.47, 7.71, 8.20 | 3.07, 3.10, 3.41 |
| 1GJS       | 4.51, 5.07, 5.36 | 3.76, 3.97, 4.65 |
| 1BA5       | 5.82, 5.92, 6.59 | 5.63, 5.72, 5.85 |
| 1ENH       | 2.96, 4.63, 5.72 | 2.44, 2.96, 3.01 |
| 1GUU       | 7.73, 8.62, 9.17 | 4.05, 8.25, 8.99 |
| 1UZC       | 5.05, 6.10, 7.09 | 4.25, 5.22, 9.40 |
may be quite large (200 of 400 configurations at the half point of the transition). This finding raises the question of how the best models from the GC can be reliably chosen using objective, quantifiable criteria. Although ranking predictions by energy provides reasonable results, it clearly fails in such cases as 1GAB (Table 3). The success of clustering in eliminating misfolds suggests that some topological features of the graphs may serve to identify the native state. We hypothesize that the most native conformations, if properly sampled, should be the most connected within the GC because a higher population of similar structures results in more connections between those conformations. We find that clustered conformations exhibit a general relationship between the number of neighbors a node exhibits (called $k$, the degree of the node) and the rmsd from the native state (see Fig. 7, which is published as supporting information on the PNAS web site). Although conformations with low rmsd from the native may exhibit a low degree, most nodes of high degree are among the most native conformations observed in simulations. When the three highest-degree nodes from the GC are chosen, they unilaterally include some of the highest-quality structures from the entire simulation. This approach is possible because the graph-theoretic approach we employ preserves topological information about the relationship between nodes in a cluster.

The superposition of the top $k$ prediction and native state for each protein is presented in Fig. 1. These predictions are obtained from a generalized, fully transferable potential, and the graph-theoretic approach represents a completely consistent, objective analysis that requires no knowledge of the native state. The resulting predictions, three at $\sim$3 Å rmsd, three at 4 Å, and a 5-Å “fold prediction” in the worst case, demonstrate the effectiveness of our potential and the utility of the clustering method. As discussed in the preceding section, the 1GUU GC contains native and misfolds at the half point and splits into two clusters at lower cutoffs. Applying the same highest-$k$ criteria to these clusters provides a very good (3 Å) structure prediction in one of the two cases. In every test case the GC at half transition contains the native fold. In every test, the native fold is always among the highest three $k$-conformations. Graph-theoretic analysis thus provides a means for enriching the native fold by means of clustering, choosing a high-quality prediction using $k$, and gauging the reliability of that prediction by identifying the misfold problem when it occurs.

**Controls and the Meaning of the Results.** From the application of our protocol to seven real proteins, it is clear that sequences designed by evolution to fold do so in our model. However, several potential concerns should be addressed through control simulations. First, we show that the $\mu$-potential contains useful and meaningful information by demonstrating that it does not behave like a random potential. Second, we verify that the $\mu$-potential predictions are meaningful in that the model is not constructed to generically fold every sequence into a helix-bundle topology. Lastly, we verify that the predicted topologies result from the interplay between the H-bonding and pair contact potential.

In all control simulations (Table 4), 1BDD is selected to represent albumin-binding domains, and 1ENH is chosen to represent DNA-binding domains. The protocol for 200 folding simulations and clustering analysis for each control are identical to those for folding the test proteins. The $\mu$-potential has interaction energies that range from $-1$ to 1 and have an average of 0. For comparison, we produce a random potential with the same statistical properties and apply it, along with the same H-bond potential, to the folding of the 1BDD and 1ENH sequences. In 1BDD, this potential results in an average of 16.72-Å rmsd from the native state, and for 1ENH the average rmsd of the resulting structures is 16.40 Å. In both cases, the structure nearest to native is $\sim$11 Å and, upon clustering, the GC contains native and misfolds at the half point and splits into two clusters at lower cutoffs. Applying the same highest-$k$ criteria to these clusters provides a very good (3 Å) structure prediction in one of the two cases. In every test case the GC at half transition contains the native fold. In every test, the native fold is always among the highest three $k$-conformations. Graph-theoretic analysis thus provides a means for enriching the native fold by means of clustering, choosing a high-quality prediction using $k$, and gauging the reliability of that prediction by identifying the misfold problem when it occurs.

![Fig. 1. Top $k$ predictions from Table 3, superimposed on the native conformation and colored from N (blue) to C (red) termini.](image)

**Table 4. $C_r$ rmsd results from controls of 1ENH and 1BDD**

| Protein ID | All 200 | GC | Prediction | 3 lowest $E$ | 3 highest $k$ |
|-----------|---------|----|------------|--------------|--------------|
| RS        | 10.70   | 6.40–16.10 | 10.25 | 9.18–11.47 | 9.69, 10.18, 10.37 | 9.95, 10.26, 10.43 |
| RP        | 16.40   | 11.2–21.25 | 16.30 | 14.52–18.33 | 14.51, 15.73, 17.86 | 14.93, 15.20, 16.05 |
| PO        | 9.67    | 5.66–13.07 | 10.24 | 8.78–11.03 | 9.55, 10.63, 11.00 | 10.12, 10.22, 10.52 |
| RS        | 12.72   | 4.77–22.09 | 9.62  | 6.13–12.05 | 11.07, 11.42, 12.05 | 10.18, 10.42, 11.42 |
| RP        | 16.72   | 10.65–22.10 | 17.98 | 15.98–20.46 | 16.26, 18.63, 19.50 | 18.26, 18.63, 18.85 |
| PO        | 10.70   | 7.15–13.68 | 10.52 | 8.60–13.07 | 9.38, 10.94, 11.02 | 9.98, 10.16, 10.47 |

The randomized sequence (RS), randomized potential (RP), and $E_{rat}$ potential only (PO) yield poor results that are not improved by graph-theoretical analysis.
some arbitrary random protein-like sequence it does not produce helix-bundle conformation. The FASTA sequences for 1BDD and 1ENH were each subjected to 6,000 randomizing permutations (see Fig. 8, which is published as supporting information on the PNAS web site). When the resulting sequence was checked by using BLAST against the nonredundant National Center for Biotechnology Information database of protein sequences (31), no known sequence homologues were found. A new amino acid chain of the same amino acid composition and length as two different real proteins, but different sequence, then was constructed by using SWISS PDB VIEWER. From the results it is clear that the 200 control runs result in structures that are much worse than the prediction runs and that no improvement comes from clustering. Within the GC, no helix-bundle topology resembling the native conformation of either protein is identified by energy or clustering (Table 4). Although no conformations resembling the WT are produced, the GC and “prediction” represent a cohesive set of similar structures. Clearly, this model is not contrived to turn any sequence into a helix bundle: a random collapsed false prediction should exhibit a rmsd of \( \approx 10 \, \text{Å} \), highlighting the significance of the \( \approx 3\text{-Å} \) predictions.

Simulations run with only the \( \mu \)-potential result in collapsed structures with a compact hydrophobic core, but without secondary structure, that are \( \approx 10\text{-Å} \) rmsd from the native state. Moreover, clustering analysis fails to identify a dominant GC, even at high \( r \) cutoffs. When the conformations do cluster, they form a large number of small clusters, indicating great structural diversity. If simulations are run with only H-bonding, then the amino acid chain becomes a single extended helix, with a 19.43- and 21.30-Å rmsd for 1ENH and 1BDD, respectively. Neither \( E_{\text{AB}} \) nor \( E_{\text{HB}} \) alone can identify native conformations. Although an atomic interaction term is necessary for collapse, secondary structure formation functions to limit the conformational space available to an amino acid chain, reducing the conformational space necessary to search for the global minimum and contributing to the presence of a cohesive, well defined ensemble conformations in a protein’s folded state. It was recently suggested (32, 33) that compaction, chain geometry, and excluded volume ensure protein-like conformations. The above simulations indicate that these factors alone are not able to produce protein-like conformations. We find that necessary conditions for a protein model to achieve a realistic tertiary structure include a geometrically and spatially realistic side chain and backbone representation, an accurate representation of H-bonding, and a potential that represents specific hydrophobic and other atom–atom interaction in a physically appropriate manner (34).

From the clustering of each control (Table 4) it is clear that the resulting “predictions” are nonrandom but do not resemble helix bundles. We observe primarily compact conformations that may have helices and turns, with the obvious exception of using only the H-bonding potential, which produced a single extended helix. It is not surprising to observe a small number of collapsed conformations resembling helix bundles in the randomized sequence controls, some as close as 4 Å from native. However these conformations cannot be identified by energy or any graph-theoretical criteria introduced in this work. Several studies have based claims of successful folding, at least in part, on the

![Fig. 2. The relationship between graph-theoretic analysis and landscape theory. Proteins (designed heteropolymers) exhibit a deep, pronounced minimum (large, dense native cluster), but the landscape is rugged with low energy traps (other small and disjoint clusters). The randomized sequence behaves as a random heteropolymer, characterized by multiple energy minima and lacking a single, prominent minimum (native conformation). Graphs are calculated with results from simulations; landscape cartoons are only illustrative. Nodes are colored to represent the rmsd (Å) from the native state with <3.5 purple, <4.0 blue, <4.5 green, <5.5 yellow, <6.5 orange, and >6.5 red.](image)

### Table 5. Comparison of representative, contemporary models for protein folding

| Reference | All atom? | Length of test proteins, aa | No. of test proteins | rmsd objective predictions, Å | Lowest rmsd observed, Å | Test and training set separate? |
|-----------|----------|-----------------------------|---------------------|-------------------------------|--------------------------|-----------------------------|
| Present work | Yes | 43–56 | 7 | 2–5 | 2–3 | Yes |
| 16 | No | 28–75 | 7 | 2–5 | 2–4 | No |
| 23 | No | 69–170 | 5 | NR | 3–10 | Yes |
| 18 | Yes | 20–60 | 4 | 2–4 | 2–3 | No |
| 17 | No | 47–76 | 6 | 4–11 | 2–6 | Yes |
| 13 | Yes | 46 | 1 | NR* | 3 | Yes |
| 21 | No | 36–147 | 11 | † | 2–4 | No |
| 12 | Yes | 36–46 | 2 | NR | 3–6 | Yes |
| 24 | Yes | 36 | 1 | NR | NR | Yes |
| 15 | No | 46 | 1 | 9 | 2 | Yes |
| 20 | No | 46 | 1 | 3 | 3 | Yes |

* Some replicas were initiated from <2-Å conformations.
† Utilized experimentally determined secondary structure assignments.

NR, value not reported.
presence of low rmsd conformations (12, 22) in a subset of trajectories. However, from these controls we see that it is possible to sample helix-bundle conformations by randomly collapsing a random sequence heteropolymer chain with the caveat that the resulting native-like structure exists only in a tiny fraction of runs and cannot be identified by clustering or energy. For a protein-folding simulation only identification of low rmsd conformations by quantitative criteria independent of knowledge of the native state constitutes a successful description of a physical sequence-dependent folding event.

Conclusions

These results represent significant progress and promise for understanding protein folding and structure prediction in ab initio high-resolution simulations. Systematically applying a quantitative method that uses an energy function along with clustering to identify the prediction consistently identifies structures in the 3-Å rmsd range for most proteins ($P = 10^{-3}$); even the $<6$-Å ($P = 10^{-5}$) (26) structures are relatively good, indicating that Table 3 (excluding mirror images) presents as set of meaningful predictions. Clustering works in concert with the ensemble of energy-based predictions to reliably eliminate decoys and identify the free-energy minimum structures that closely resemble the native state. Graph-theoretic analysis provides an indication of the quality of the results, identifies the most native folds and misfolds, and provides a conceptually useful link to interpretation of the results in the context of physical theories of folding.

Fig. 2 shows the graph for all structures at the transition in the GC for 1BDD folding simulations and randomized sequence control. It is clear from this figure, and it has been shown analytically (35, 36), that random and designed heteropolymers exhibit different behaviors dictated by their energy landscapes. Whereas the designed polymer largely populates a distinct and deep minimum, the random polymer inhabits multiple, energetically similar, but structurally unrelated, states. This view from heteropolymer theory explains why clustering of conformations improves predictive power of the method: native-like states feature multiple interactions that work in concert to provide minimum energy structure in the native state, consistent with the “principle of minimal frustration” (37). Because our potential is approximate, the energy landscape has features of both design and random heteropolymers. Apparently the native basin of attraction is “broad,” containing many structurally related conformations, in contrast to spurious minima, characteristic of random heteropolymers that may be deep but contain only few structures. A graph is a topological entity and can serve to analyze and conceptualize the multidimensional protein-folding energy landscape (23, 38) without reliance on spatial coordinates and, as such, has potential for representing energy landscapes in a limited set of order parameters. These data show that a high $k$ in a graph corresponds to a high density of states, or energy minimum, on the protein-folding energy landscape. In addition to understanding energy landscape topology at the minimum (structure prediction), we anticipate many applications for this model in understanding global landscape topology (folding).

We have demonstrated that fully atomistic simulations, using a single protein representation from beginning to end, have the ability to fold multiple proteins to their native states with a single transferable potential that is not trained or optimized on the test proteins or decoy sets and relies on absolutely no information of the native structure. Comparison of our results with contemporaneous studies is given in Table 5. Controls show that simply collapsing an amino acid chain with an attractive pair potential is insufficient to fold proteins; the problem of secondary structure also must be addressed. Likewise, we have demonstrated that our potential contains enough specific information to identify the native structures of proteins, without such bias that it introduces false-positive predictions; it does not turn any sequence into a helix-bundle protein. The model works because it accurately captures protein geometry and side-chain packing, H-bonding, and secondary structure formation and presents a physically reasonable pairwise potential for compaction. It is both encouraging and intellectually satisfying that simple physical models reliably represent many of the aspects of protein folding and that graph-theoretic analysis conceptually links the results of folding simulations to energy landscape theory.

We thank C. Brian Roland for simulating discussion and careful reading of the manuscript. I.A.H. and E.J.D. are supported by the Howard Hughes Medical Institute. This work was supported by National Institutes of Health Grant GM52126.