PROTEINS: Structure, Function, and Bioinformatics 65:922–929 (2006)

Improved Beta-Protein Structure Prediction by Multilevel Optimization of NonLocal Strand Pairings and Local Backbone Conformation

Philip Bradley and David Baker*
University of Washington, Seattle, Washington

ABSTRACT **Proteins with complex, nonlocal beta-sheets are challenging for de novo structure prediction, due in part to the difficulty of efficiently sampling long-range strand pairings. We present a new, multilevel approach to beta-sheet structure prediction that circumvents this difficulty by reformulating structure generation in terms of a folding tree. Nonlocal connections in this tree allow us to explicitly sample alternative beta-strand pairings while simultaneously exploring local conformational space using backbone torsion-space moves. An iterative, energy-biased resampling strategy is used to explore the space of beta-strand pairings; we expect that such a strategy will be generally useful for searching large conformational spaces with a high degree of combinatorial complexity. Proteins 2006;65:922–929.**

Key words: protein structure prediction; beta-sheet; fragment assembly; Rosetta

INTRODUCTION

Protein structure prediction from sequence information alone is a grand challenge of molecular biology. Considerable progress has been made in recent years, fueled in part by the biennial experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP).[1] Despite this progress, predicting structures for larger, more topologically complex proteins, particularly proteins with nonlocal beta-sheets,[2] remains a formidable challenge.

Structure prediction for beta-sheet proteins is challenging for several reasons. First, long-range (sequence-distant) beta-strand pairings are difficult to sample because of the coarse-grained nature of the conformational search, which is necessitated by the very large size of protein conformational space. Long-range beta-strand pairings require a precise relative geometry which is hard to achieve by random torsion-space moves in the intervening segment (by contrast, local beta-strand pairings, such as beta-hairpins, can be formed by insertion of a small number of compatible fragments). Second, beta-strand pairings have a very high entropic cost once formed: the intervening segment is effectively frozen since any torsion-space move will likely perturb the geometry of the pairing. Third, formation of nonlocal beta-pairings may be prevented by the formation of competing local beta-pairings, which are easier to sample. Finally, the number of alternative nonlocal beta-sheet topologies is very large, leading to a large space of conformations that must be searched.

Here we describe a new approach to predicting the structures of complex, beta-sheet containing proteins. Our solution is to take advantage of regularities of beta-sheet protein structures to radically reformulate structure generation and sampling in a way that makes forming long-range pairings quite simple and reduces the entropic cost of their formation. We employ a multilevel sampling approach to explicitly sample alternative long-range pairings and at the same time explore local conformational space using fragment assembly in torsional coordinates. To accomplish this, we replace traditional folding from the N to C terminus of a continuous chain with "fold tree"-based generation of structures with one or more long-range connections and an equal number of chain breaks. Nonlocal strand pairings are formed and maintained by construction via these long-range connections. Sampling of alternative nonlocal strand pairings is carried out using an iterative, energy-biased resampling approach in which nonlocal pairings observed in previous rounds are explored while local conformations such as hairpins and beta–alpha–beta units are stochastically disfavored.

MATERIALS AND METHODS

Fold Tree Representation

The protein chain is represented by a *fold tree*—a directed, acyclic, connected graph composed of peptide segments together with long-range connections. This tree is constructed from a simple graph in which each residue i is connected to residues $i - 1$ and $i + 1$. A new edge is added to the graph for each long-range connec-

*Correspondence to: David Baker, Department of Biochemistry and HHMI, University of Washington, Box 357350, Seattle, WA 98195. E-mail: dabaker@u.washington.edu

Received 16 January 2006; Revised 2 May 2006; Accepted 23 June 2006

Published online 10 October 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21133

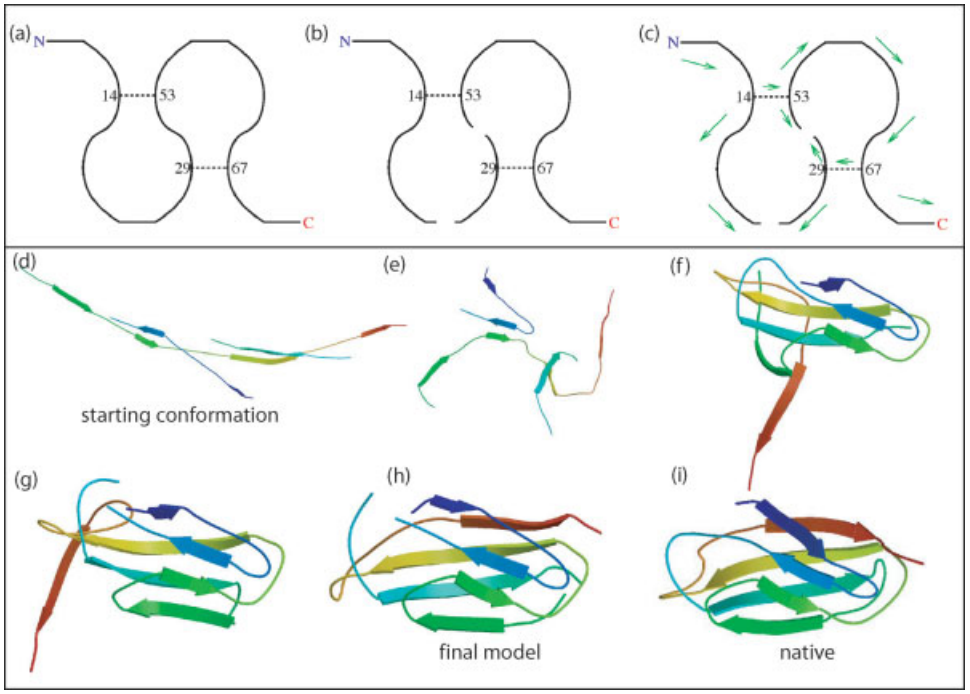

Fig. 1. (**a–c**) Construction of a fold tree for a simulation with two forced pairings (14–53 and 29–67). Starting with a simple graph corresponding to the peptide chain, long-range edges are added between 14 and 53 and between 29 and 67 (a). Chain breaks are chosen randomly to generate a tree [acyclic graph, (b)]. Finally, the tree is rooted at the N-terminus, defining a folding order (green arrows) by which the structure is built (c). (**d–h**) Snapshots from a folding trajectory that uses this fold tree. The target is the benchmark protein 1fna, whose native structure is shown in (**i**). For ease of comparison, the locations of the seven strands in the final model are shown in each frame. In the starting conformation (d), the torsion angles are initialized to extended values. The two nonlocal beta pairings are present by construction, and chain breaks can be seen between the second and third strands and between the third and fourth strands. Early in the simulation (e), the structure is still quite extended. Packing between the two sheets begins to develop in (f). The sheets grow outward from the forced pairings (g), with the C-terminal strand being added last (h).

tion [Fig. 1(a)]. These connections determine how conformational change propagates through the structure. They can represent rigid or semirigid orientational constraints such as beta-sheet pairings or disulfide bonds, or fully flexible linkages such as the connection between two docking partners. In the current application, a new edge would be added for each long-range strand pairing that is being forced. Peptide bond ($i \rightarrow i + 1$) edges are then deleted at random until the graph is acyclic [Fig. 1(b)], while preserving connectivity. Edges are selected for deletion with a bias proportional to the predicted loop frequency. Finally, an ordering of the graph is defined by selecting a root vertex, for example, the N-terminus [Fig. 1(c)]. This graph provides a rule for generating three-dimensional coordinates from backbone torsion angles and a set of rigid-body transformations (one for each long-range connection): starting with an arbitrary location and orientation for the root vertex, traverse the edges of the graph in order, using torsion angles and/or rigid body transformations to build the terminal vertex of each edge given the coordinates of the initial vertex for that edge. The program Undertaker[3] developed by the Karplus group at UCSC implements a similar tree representation.

Beta-Sheet Transforms

Rigid-body transformations between the coordinate systems defined by the $N-C_\alpha-C$ atoms of paired residues in beta-sheets were extracted from proteins of known structure (6246 total transforms). These transformations fall into four classes defined by the strand orientation (parallel or antiparallel) and an additional *pleating* term (-1 or 1) that specifies whether the beta-carbons of the paired residues point into or out of the plane of the beta-sheet when the residues are oriented such that the first residue is on the left with chain direction increasing from bottom to top (or equivalently, whether the NH and CO groups of the first residue point toward or away from the second residue).

Fragment Assembly Protocol

This fold tree representation is incorporated into a modified fragment assembly protocol. As in the standard Rosetta[4] de novo protocol, we perform a large number of independent simulations (typically 4000) with different random number seeds. At the start of each simulation, one or more nonlocal beta-strand residue pairings (described by a pair of residues, an orientation, and a

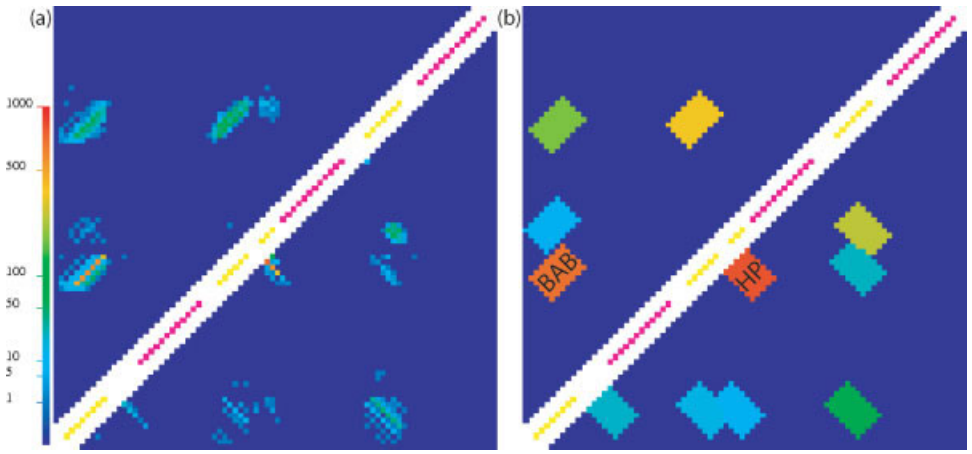

Fig. 2. Feature identification. (**a**) The raw frequencies of residue–residue beta-pairings are plotted for a set of models. Sequence number increases from left to right and from bottom to top. The color of the square at position (*i, j*) indicates the frequency with which residues *i* and *j* are paired in a beta-sheet with parallel ($i < j$) or antiparallel ($i > j$) orientation (i.e., parallel pairings appear above the diagonal, antiparallel below). The raw frequencies are smoothed and the local maxima of the smoothed frequency distribution are defined as feature centers. (**b**) The identified features are plotted as blocks centered on the feature center and colored by frequency. Each feature is classified according to the average length and secondary structure content of the loop between the paired strands in the decoys containing that feature. Two local features, a hairpin ("HP") and a parallel beta–alpha–beta unit ("BAB"), are labeled.

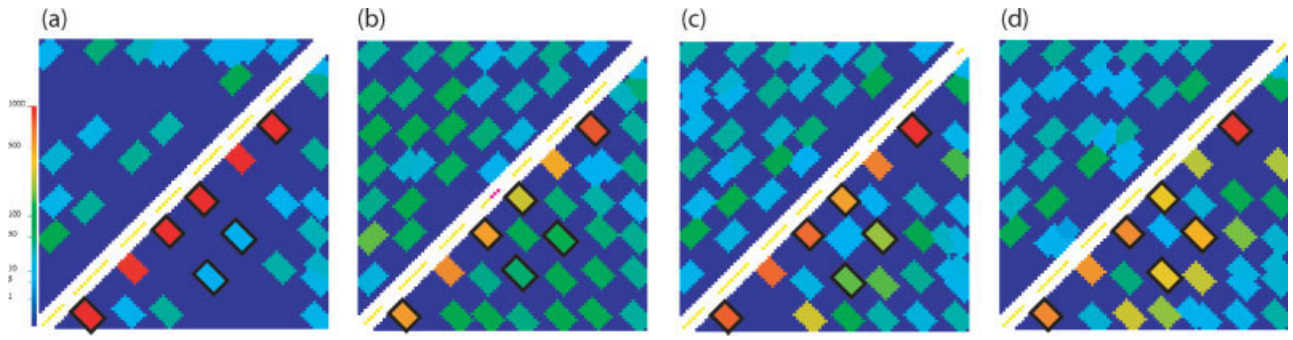

Fig. 3. Beta-sandwich resampling. The feature frequencies are plotted for models in the four rounds of the beta-sandwich resampling protocol for 1who, using the same plotting scheme as in Figure 2(b). Native features are boxed. Features from models built by the standard protocol are shown in (**a**); note the high frequency of beta-hairpins. In the models built with stochastic killing of hairpins, these features are less prevalent and an increase in the frequency of nonlocal features can be seen. A further increase is evident in the decoys built by stochastically resampling a single nonlocal feature (**c**), and in the models built by resampling interlocked feature pairs (**d**). The two native, nonlocal features (strand pairings 2 → 5 and 3 → 6) form an interlock; the models in a–d are built without knowledge of the native structure.

pleating) are chosen according to the stochastic sampling strategy described below. A fold tree is constructed from the residue pairings [Fig. 1(a–c)], the torsion angles are initialized to extended values, and a beta-sheet transform with the specified orientation and pleating is selected for each pairing; this defines the starting conformation [Fig. 1(d)]. The simulation proceeds by fragment-replacement Monte Carlo trials as in the standard Rosetta protocol, while maintaining the relative orientation of the aligned residue pairs [Fig. 1(d–h)]. A pseudoenergy term favoring closure of the chain breaks is included in the Rosetta potential function, with a weight that increases throughout the simulation. This term is proportional to the RMSD between the backbone atoms surrounding the break and a set of pseudoatoms built by folding forward and backward across the junction using the current torsion angles and ideal bond lengths and angles. The weight on this term is negligible in the early stages of the simulation, while at the end of the simulation breaks in the chain are strongly disfavored. Low-energy models tend to have fairly well-closed chains, although minor breaks do remain [Fig. 1(h)].

Sampling Beta Features

Given this machinery for constructing models with specified pairings, we have developed an iterative resampling strategy for beta-sheet structure prediction in which the nonlocal beta-pairings present in low-energy structures from one round of models are resampled to generate the next round of structures. To select beta-pairings for resampling, we first process the raw residue–residue pairings that occur in a set of models into a set of topological features. Featurizing allows us to classify the

strand pairings and apply topology-based filters. Because beta strands may be present in some models and absent, shifted, or broken in others, we identify pairings as features in the two-dimensional space of contacts, rather than first defining consensus strands and pairing them. The residue–residue beta-pairing frequency distribution [Fig. 2(a)] is smoothed (by simple averaging over a local mask oriented parallel to the strand pairing direction), and the local maxima of the smoothed frequency distribution are defined as feature centers [Fig. 2(b)]. Each feature is classified based on the average properties of the loop between the two paired strands in the set of decoys matching that feature, allowing us to identify recurring motifs such as beta-hairpins and parallel beta–alpha–beta units (parallel $i \rightarrow i + 1$ strand pairings with an intervening alpha-helix). Overrepresented local features such as hairpins may then be selected for stochastic killing, while a subset of the nonlocal features are chosen for resampling in the next round of models. Reduction of the beta-pairing information into a discrete set of features also allows visualization and classification of structural models by topology.[5]

Penalizing Local Features

We seed our iterative resampling protocol with models built by standard connected-chain fragment assembly, motivated by the empirical observation that although nonlocal pairings are undersampled by standard fragment assembly simulations, there is a bias toward the native pairings. To ensure that a diverse set of nonlocal pairings is present in this initial set of models, we selectively disfavor competing local pairings when generating these models. This is done in a stochastic fashion, since any given local strand pairing may be present in the native structure. A subset of overrepresented local features is first identified by processing the beta-pairings observed in standard fragment-assembly models. For all-beta proteins, the overrepresented features are beta-hairpins [Fig. 3(a)], while for alpha + beta proteins the overrepresented features are parallel beta–alpha–beta units [Fig. 2(b)]. At the start of each simulation, each overrepresented local feature is independently and with probability 0.5 assessed a score penalty for the duration of that trajectory that effectively prevents its formation.

Resampling Strategies

We consider here two classes of target: (1) beta-sandwich proteins (a subset of SCOP[6] class b); (2) the mainly antiparallel class of alpha + beta proteins (SCOP class d). We found that the difficult targets in our in-house benchmark set of moderate-length proteins (<125 residues) were almost entirely from these two classes.* As described below, the sheet topologies of beta-sandwich proteins show regularities that our sampling protocol is well-suited to exploit. For alpha + beta proteins, a simple strategy involving a single round of resampling was tested. We start with a seed population of connected-chain models built with stochastic disfavoring of overrepresented beta–alpha–beta units, and build one round of models using the fold tree protocol, resampling a single nonlocal antiparallel beta feature in each simulation. The features are sampled with frequencies proportional to their frequency in the starting population; for each feature, alternative registers and pleatings are taken from the starting models.

Beta-sandwich proteins are characterized by a pair of beta-sheets packing face to face, with connecting loops crossing between the sheets at the top and bottom of the sandwich. The topologies of sandwich proteins have been extensively analyzed, and regularities have been discovered that can be used in structure prediction.[7,8] These include an avoidance of parallel strand pairings, particularly in the interior of the sheet, and a preference for a topological arrangement in which strands i and j are paired in one sheet while strands $i + 1$ and $j + 1$ are paired in the opposite sheet (termed an interlock or cross-beta). The analysis of Fokas et al.[8] suggests that nearly every sandwich protein contains at least one interlock. Our prediction strategy consists of two rounds of resampling, starting from a seed population of connected-chain models built with stochastic disfavoring of beta-hairpins [Fig. 3(b)]. In the first round of resampling, a single nonlocal, antiparallel feature is forced in each simulation [Fig 3(c)]. We exclude features for which the average length of the loop between the paired strands is greater than 45 residues, thereby targeting the midrange, $i \rightarrow i + 3$, strand pairings that are enriched in sandwich proteins. In the second round of resampling, two interlocked beta features—taken from the low-energy models in the first round—are forced in each simulation [Fig. 3(d)]. Because peripheral beta-strands are often inserted in beta-sandwiches, we relax the definition of interlock, requiring two strand pairings $i - j$ and $k - l$ for which $i < k < j < l$ (in the original definition, $k = i + 1$ and $l = j + 1$). An example of an interlocked resampling trajectory is illustrated in Figure 1.

Folding Sequence Homologues

It has been demonstrated previously[9,10] that simulations with sequence homologues can be a powerful tool for enhancing sampling in de novo structure prediction. In our protocol, we select up to 30 sequence homologues of the target protein, generating de novo models for each homologue in parallel with the target sequence. The sequence alignment of the target with the homologues is used to analyze features across the full set of models, and for coclustering the models as described in Bonneau et al.[9]

*This is partly a consequence of the length distributions for different structural classes: only 24 of 222 SCOP superfamily representatives (11%) in the alpha/beta class are under 125 residues, whereas 205 of 408 alpha + beta representatives (50%) are under 125 residues.

TABLE I. Summary of Benchmark Results

				Cluster MaxSub		Avg # native pairings		
PDB	Class	SCOP ID	Len	Old	New	Total	Old	New
1wapA	b	b.82.5.1	68	42	56	3	0.09	1.67
1npsA	b	b.11.1.1	88	29	31	4	0.06	0.48
1fna_	b	b.1.2.1	91	44	84	2	0.15	1.66
1who_	b	b.7.3.1	94	45	73	2	0.05	0.64
1tul_	b	b.85.5.1	102	34	65	3	0.10	1.06
1sppB	b	b.23.1.1	112	44	49	7	0.13	0.64
1hdn_	a + b	d.94.1.1	85	43	84	2	0.17	0.95
1ris_	a + b	d.58.14.1	92	46	79	2	0.16	0.80
2acy_	a + b	d.58.10.1	98	69	79	3	0.30	0.86
1fkb_	a + b	d.26.1.1	107	56	52	2	0.15	0.52
1kpeA	a + b	d.13.1.1	108	50	49	1	0.10	0.50
1acf_	a + b	d.110.1.1	125	106	95	1	0.02	0.23

The resampling protocol ("new") is compared with standard fragment assembly ("old") on a benchmark of 12 proteins using two metrics. Columns 5–6: the number of residues superimposable to the native structure with an RMSD under 4 Å for the best of five models selected by clustering. Columns 8–9: the average number of native nonlocal pairings per model over the entire population of models generated by the protocol (the total number of nonlocal pairings in the native structure is listed in Column 7).

Clustering and Model Selection

To select a small number of predictions for comparison of alternative methods, we cluster the models[11] to identify recurring conformations. Since our model sets are rather large (typically 4000 models × 30 homologues = 120,000 conformations), we apply a two-step procedure. For each homologue, the top 1000 models are chosen by score and clustered and the members of all clusters of size >5 are selected, with clustering parameters chosen to give approximately 250 models.[10] These sets of 250 models are pooled and clustered together, using the multiple sequence alignment to calculate RMSDs between models with different sequences.[9] The centers of the largest five clusters are selected as predictions.

Model Evaluation

We used the MaxSub[12] algorithm to calculate for each model the maximum number of alpha-carbons that can be superimposed onto the native structure with an RMSD under 4 Å. To assess correctness at the level of beta-sheet topology, we counted the number of native strand pairings in each model.

RESULTS AND DISCUSSION

We tested our beta-sheet resampling protocol on a benchmark of 12 challenging proteins, six beta-sandwiches and six alpha + beta proteins (Table I). For each target, we generated models using the standard Rosetta fragment assembly protocol as well as our new resampling protocol. The overall quality of the two populations of models was compared by using MaxSub to calculate for each model the number of residues superimposable onto the native structure under an RMSD of 4 Å. To determine whether the resampling protocol was generating more native-like models, we compared the near-native tail (the top 10%) of the MaxSub distribution for the two protocols (Fig. 4, red and blue curves). For the majority of the targets, the blue curve is consistently above the red curve, indicating higher-quality models, with dramatic improvements in several cases. While generating better models does not in itself guarantee that these models can be selected given a small set of predictions, these results suggest that the energy-biased resampling protocol—which does not depend on knowledge of the specific pairings present in the native structure—has improved sampling of near-native structures. The one failure in the beta-sandwich proteins (1 nps) is caused by an error in the secondary structure prediction that prevents the two native interlocks from being formed.[2] The two least-successful alpha + beta proteins (1kpeA and 1acf) also have the smallest number of native nonlocal pairings (Table I, column 7); as expected, improvement over connected-chain fragment assembly is greatest for higher contact-order[†] beta-sheet proteins.

To assess model correctness at the level of beta-sheet topology, we counted the total number of native nonlocal strand pairings present in each model. In Table I, columns 8 and 9 show the average number of native nonlocal features for models built by the standard fragment assembly protocol ("old"), and for models built by the resampling protocol ("new"). A consistent improvement is evident across the set, even on proteins for which the MaxSub distributions for the resampling protocol were not significantly better (suggesting that correctness at the level of three-dimensional coordinates is a more stringent criterion than correctness at the level of beta-sheet topology). Figure 3 provides a concrete example of this sampling improvement for the protein 1who.

While better sampling near the native structure is one goal, we would also like to be able to select more native-like models from a population of structures. We have found that in the low-resolution structure prediction regime (where no model may be within 4 Å of the native structure), clustering based on global structural similarity is a more robust method of selecting models than picking by energy. This is likely due to the fact that the low-resolution energy function appropriate to coarse-grained exploration of the landscape does not capture the structural details that contribute to the very low energy of the native structure; high-resolution all-atom potentials can reliably detect near-native models provided they are within ~2 Å RMSD of the native struc-

[†]Both of the native interlocks in 1nps involve short edge strands, and these strands are completely missed in the secondary structure predictions. As a result, very few strand fragments are present at these positions, and the beta-features that form the native interlocks are not sampled in the initial de novo models.

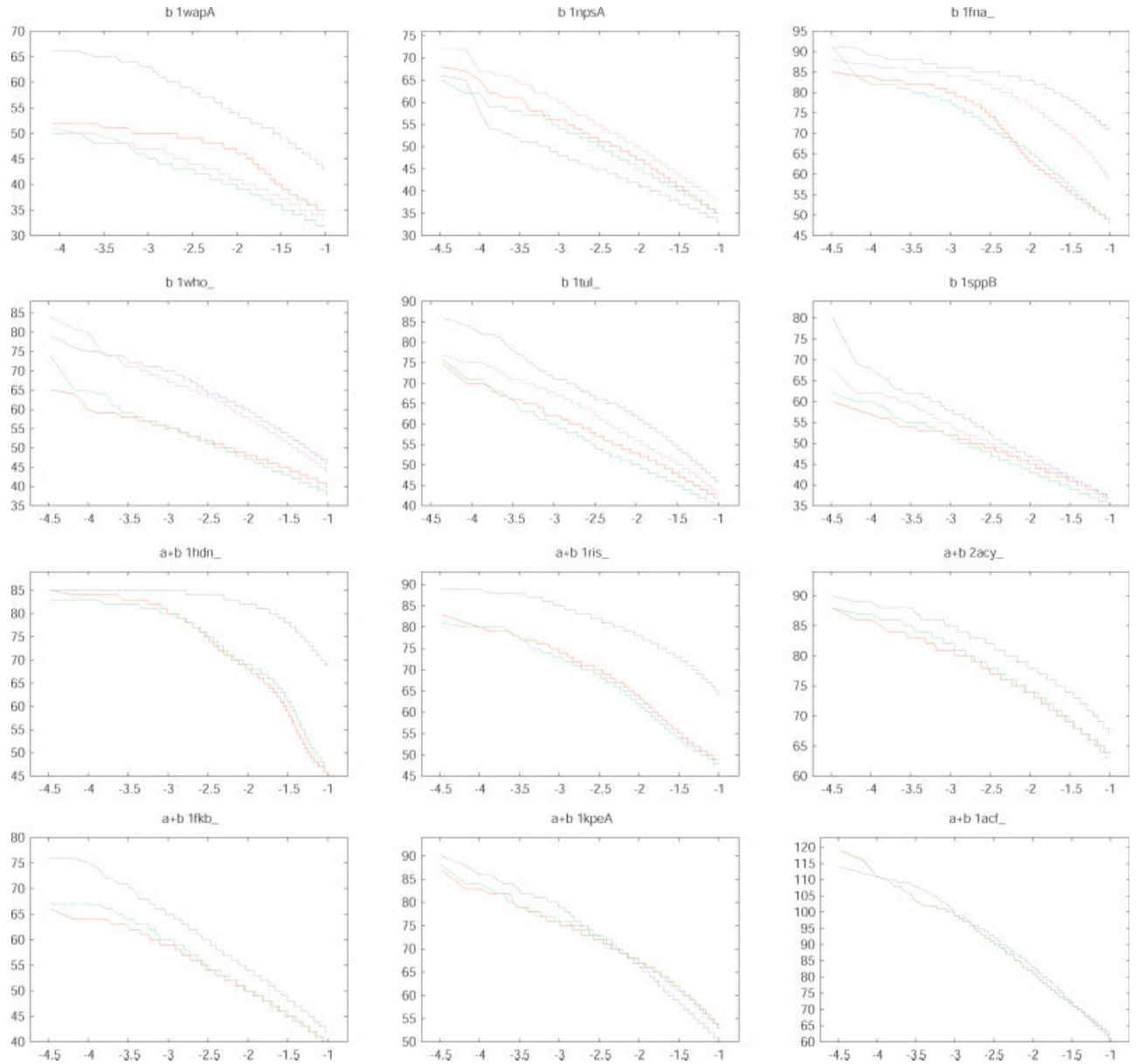

Fig. 4. MaxSub distributions for the 12 benchmark proteins. The score attached to each model is equal to the number of residues that can be superimposed onto the native under 4 Å RMSD by MaxSub. The scores for each population are sorted in decreasing order, and each model in the top 10% is plotted as follows: *x*-coordinate = log10 (rank/#decoys), *y*-coordinate = number of superimposable residues. This yields a visualization of the complementary cumulative distribution function (with log-probability on the *x*-axis and score on the *y*-axis) that focuses on the near-native tail of the distribution. Higher values for the *y*-coordinate correspond to better predictions; when comparing two protocols, the higher curve corresponds to the population of models with better sampling near the native structure. For the beta-sandwich proteins (the top six plots, titled "b <id>"), the four populations plotted are: standard fragment assembly (red lines), models built with stochastic hairpin killing (green), models with one forced pairing (pink), models with two forced pairings (blue). For the alpha + beta proteins (the bottom six plots, titled "a + b <id>"), the three populations plotted are: standard fragment assembly (red lines), models built with stochastic beta–alpha–beta killing (green), and models built with one forced pairing (blue). In all cases, the red lines correspond to the standard fragment-assembly protocol and the blue lines correspond to the final resampling protocol. Protein lengths are listed in Table 1.

ture.[10,13] Structural clustering after multiple rounds of resampling will tend to identify beta-sheet topologies that have been enriched by the energy-biased resampling process. We applied a clustering protocol that included an initial energy filter: the lowest 25% of the structures by energy were selected and clustered by full-chain alpha-carbon RMSD (see Clustering and Model Selection). For each protocol, the centers of the five largest clusters were selected as predictions and compared with the native structure using MaxSub. The MaxSub values for the best model in each set of five cluster centers are reported in Table I, columns 5–6. Of the eight proteins with a MaxSub score difference of 10 residues or greater, seven improved with the resampling strategy.

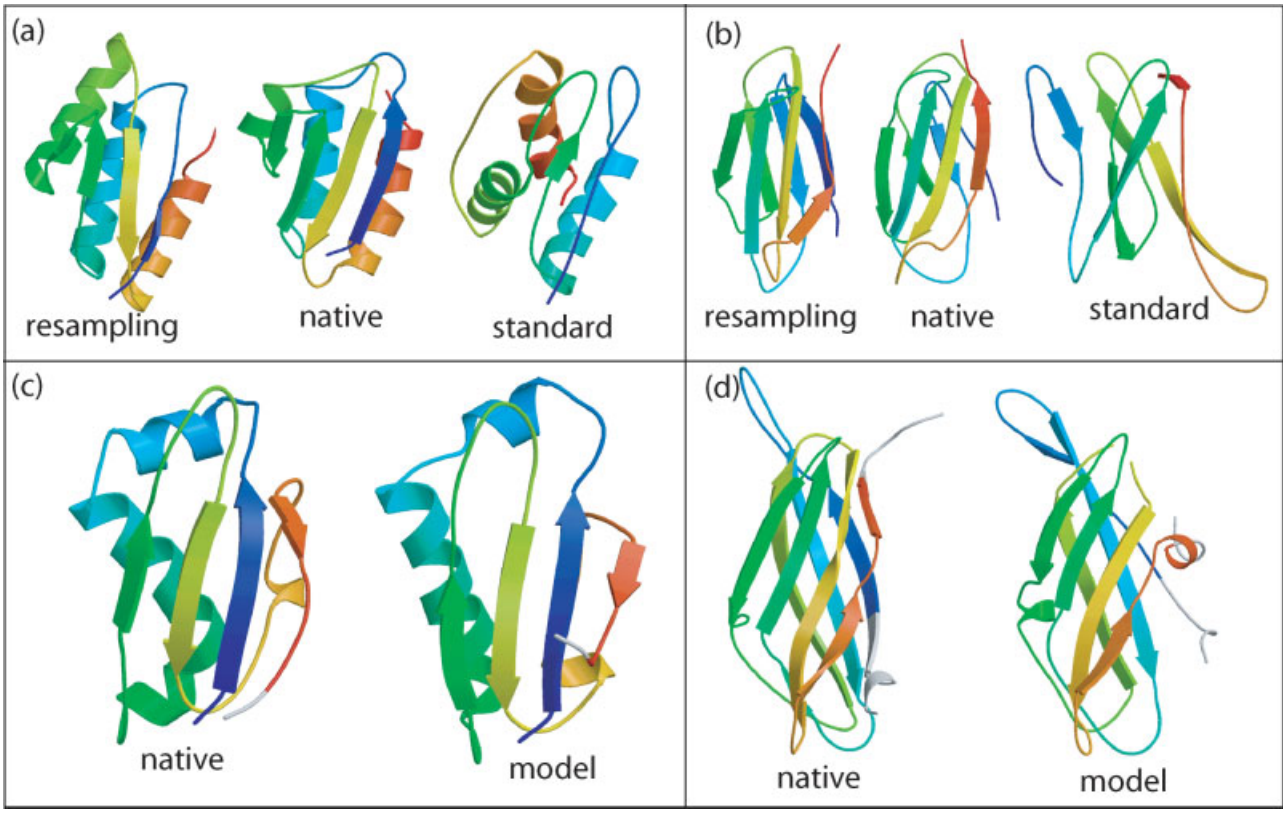

Fig. 5. Structures produced with the beta-sheet resampling protocol. (**a–b**). Comparison of the best of five cluster centers produced by beta-sheet resampling or by standard connected-chain fragment assembly (see Table 1, columns 6–7) for the benchmark proteins 1hdn (a) and 1fna (b). (**c–d**) CASP6 predictions made with the beta-sheet resampling protocol. (c) Predictions for target T0272 domain 1 were made by resampling consensus nonlocal pairings from de novo models. Model 1 is shown, for which 85 residues superimposed onto the native with an RMSD of 3.4 Å. (d) Predictions for target T0212 were made by resampling consensus nonlocal pairings from fold-recognition models generated by automated servers. Model 2 is shown, for which 109 residues superimposed onto the native with an RMSD of 3.97 Å.

Several of the improvements are quite dramatic, with 30–40 additional residues that superimpose well to the native structure. Models for the two proteins with the largest difference between the two protocols are shown in Figure 5(a,b).

The beta-sheet resampling protocol was also tested during the recent CASP6 experiment.[5] Figure 5 (c,d) shows two cases in which the new protocol produced the best prediction submitted by any group. Target T0272 consisted of two domains with an alpha + beta ferredoxin fold. We used the nonlocal resampling strategy for this target, producing the best model for both domains. Target T0212 had a beta-sandwich fold. For this target, we applied the resampling protocol to build models using consensus beta pairings taken from automated fold recognition servers. This illustrates that the resampling strategy can be seeded with fold recognition models as well as de novo models.

CONCLUSIONS

We have developed a novel, multilevel sampling approach to beta-sheet structure prediction. In this approach, we explicitly sample alternative long-range pairings and at the same time explore local conformational space using fragment assembly. Our results suggest that this method can produce more accurate models for proteins with complicated beta-sheets. With our new approach, long-range interactions can form before the intervening local interactions, as has been observed in the folding of naturally occurring proteins.[14,15] The fold-tree framework we have implemented provides a general solution to the problem of simultaneously optimizing torsional coordinates together with rigid body (Cartesian-space) transforms, and should have applications ranging from flexible backbone docking to predicting protein–DNA interactions.

It should be possible to improve the prediction of structures for proteins with complex topologies further by developing resampling strategies tailored for alpha/beta class proteins, and extending the resampling strategy used here to more complicated beta-sheet proteins. Simulations with the most complex protein in the benchmark, 1sppB, suggest (data not shown) that multiple rounds of resampling, forcing an additional nonlocal beta-strand pairing in each round, can successfully generate the native topology—if we know in advance how many nonlocal pairings to construct. The challenge is to

develop a balanced strategy that will work for proteins with a range of beta-sheet complexities. In particular, performance on proteins with few or no nonlocal beta-pairings is somewhat degraded by resampling of nonlocal features; however, these are the proteins for which the connected-chain fragment assembly protocol already performs quite well. A combined strategy should allow us to generate good models for a broad range of targets.

In this paper, we have applied a resampling protocol to search the space of beta-sheet structures. Resampling is a useful strategy more generally for exploring very large spaces with a high degree of combinatorial complexity. It requires a method of assigning a set of feature values to each sample, and an efficient means of generating new samples with specified features. These feature sets need not be enumerated in advance—each round of resampling begins with a population of samples, and we can learn the feature set from the samples themselves (as is done here for beta-strand pairings). To shift the population toward the solution, it must be possible to bias the feature selection between rounds of resampling, either by using characteristics of the solution that are known in advance (e.g., topology constraints in beta-sandwich proteins) or by selecting features from a subset of the models (e.g., the low-energy models). Resampling allows intensified sampling in promising regions, and recombination of low-energy features to generate models with feature combinations that would be difficult to sample using the base method.

The premise underlying resampling approaches—that for any given feature with multiple possible states, the native state is on average lower in energy than nonnative states—has appeared in many guises in theoretical work on protein folding. It is closely related to the principle of minimum frustration in energy landscape theories of protein folding,[16] and very simple models with this property show that protein folding can take place rapidly despite the astronomically number of possible conformations.[17] Physically, the average energy bias toward native features arises from the significant gap in the total energy between the experimentally observed native structure and nonnative structures (which must be very much higher in energy as they are not observed), and the approximate decomposability of the overall energy of a protein into the sum of the energies of its parts.

ACKNOWLEDGMENTS

The authors thank Keith Laidig for flawless administration of their computing resources. This work was supported by the Howard Hughes Medical Institute.

REFERENCES

1. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP)-round V. Proteins 2003;53 (Suppl 6):334–339.
2. Bonneau R, Ruczinski I, Tsai J, Baker D. Contact order and ab initio protein structure prediction. Protein Sci 2002;11:1937–1944.
3. Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M, Hughey R. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. Proteins 2003;53 (Suppl 6):491–496.
4. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82–95.
5. Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim DE, Meiler J, Misura KM, Baker D. Free modeling with Rosetta in CASP6. Proteins 2005;61 (Suppl 7):128–134.
6. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
7. Kister AE, Finkelstein AV, Gelfand IM. Common features in structures and sequences of sandwich-like proteins. Proc Natl Acad Sci USA 2002;99:14137–14141.
8. Fokas AS, Papatheodorou TS, Kister AE, Gelfand IM. A geometric construction determines all permissible strand arrangements of sandwich proteins. Proc Natl Acad Sci USA 2005;102:15851–15853.
9. Bonneau R, Strauss CE, Baker D. Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. Proteins 2001;43:1–11.
10. Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science 2005;309:1868–1871.
11. Shortle D, Simons KT, Baker D. Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci USA 1998;95:11158–11162.
12. Siew N, Elofsson A, Rychlewski L, Fischer D. MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 2000;16:776–785.
13. Misura KMS, Baker D. Progress and challenges in high-resolution refinement of protein structure models. Proteins 2005;59: 15–29.
14. Kay MS, Baldwin RL. Packing interactions in the apomyglobin folding intermediate. Nat Struct Biol 1996;3:439–445.
15. Colon W, Elove GA, Wakem LP, Sherman F, Roder H. Side chain packing of the N- and C-terminal helices plays a critical role in the kinetics of cytochrome c folding. Biochemistry 1996;35:5538–5549.
16. Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of protein folding: the energy landscape perspective. Annu Rev Phys Chem 1997;48:545–600.
17. Zwanzig R, Szabo A, Bagchi B. Levinthal's paradox. Proc Natl Acad Sci USA 1992;89:20–22.