Improved recognition of native-like protein structures using a family of designed sequences

Patrice Koehl* and Michael Levitt

Department of Structural Biology, Fairchild Building, Stanford University, Stanford, CA 94305

Edited by Peter G. Wolynes, University of California at San Diego, La Jolla, CA, and approved November 21, 2001 (received for review August 3, 2001)

The goal of the inverse protein folding problem is to identify amino acid sequences that stabilize a given target protein conformation. Methods that attempt to solve this problem have proven useful for protein sequence design. Here we show that the same methods can provide valuable information for protein fold recognition and for *ab initio* protein structure prediction. We present a measure of the compatibility of a test sequence with a target model structure, based on computational protein design. The model structure is used as input to design a family of low free energy sequences, and these sequences are compared with the test sequence by using a metric in sequence space based on nearest-neighbor connectivity. We find that this measure is able to recognize the native fold of a myoglobin sequence among different globin folds. It is also powerful enough to recognize near-native protein structures among nonnative models.

Knowing the structure of a protein is most useful for predicting, analyzing, and modifying its function. As it is not feasible to determine experimentally the structure of every protein, structure prediction has become central to the field of structural biology and more specifically to structural genomics. On the basis of their study of ribonuclease A (1), Anfinsen and coworkers provided the first clues that all of the information required for folding a protein is to be found in its sequence. Not long after this discovery, people took on the challenge of discovering the rules that allow the protein to fold. This problem is far from simple and has not yet been solved (2). Three major routes are usually considered paths to the solution: homology modeling, threading, and *ab initio* prediction. To study a protein with unknown conformation C, the first two methods follow the same scheme: a similar protein whose three-dimensional structure is known is identified, and this protein is used as a scaffold to generate a model for C. When the sequences of the two proteins are homologous (i.e., when they have an obvious common ancestry), sequence similarity is assumed to infer structural similarity (3, 4), and the method is then referred to as "homology modeling." When the two sequences show no obvious evolutionary relationship, the method is referred to as "fold recognition," which works by assessing the compatibility of the target sequence with each member of a library of known structures (5).

Ab initio structure prediction methods try to build a model for the target protein structure without using a specific template protein. Most of these methods proceed by first generating a large collection of possible conformations (decoys), which are then searched with a scoring function to identify native or, more realistically, near-native conformations (6–9). This second step resembles the fold recognition problem, with the major difference that the library of folds considered includes computer-generated models instead of naturally occurring protein folds. In this paper, we show how recent developments of threading techniques can be applied to the problem of the recognition of near-native conformations among nonnative models.

Fold recognition techniques can be divided into two groups. Structure-based techniques such as threading rely on fitting the sequence of the unknown protein to the known structure of a template protein. This is usually done by using a conformational energy calculation (10–15). Sequence-based techniques, in contrast, aim to detect similarity between the sequence of the unknown protein and the sequence of the template protein (see, for example, ref. 16). These techniques work well when the two sequences are homologous and become less sensitive when the sequences are distantly related. Recent work has focused on improving the sensitivity of sequence-based fold recognition techniques by using information coming from multiple sequence alignments. The sequences of the proteins belonging to the same family as the template protein are used to generate either position-specific substitution matrices [profiles (17)] or hidden Markov models [HMM (18)]. The sequence of the unknown protein is subsequently tested against these profiles or HMMs. The same idea can be extended to include sequences homologous to the sequence of the protein under study. In the latter case, fold recognition involves profile–profile alignments (19). These methods provide more sensitive recognition of similarity between protein sequences that have significantly diverged through evolution. Finally, it was found that combining the information contained in the sequence-based profiles with structure-based scores has significantly improved fold recognition (20).

Ab initio protein structure prediction provides a collection of computer-generated model structures, referred to as decoys, for the protein of interest. These decoy structures are tested against the sequence of that protein, in the hope of identifying near-native conformations. A wide variety of scoring energy functions has been developed for that purpose (6, 21–26). These functions are either based on physical principles or derived from known protein structures by using statistical methods.

In parallel to the recent developments of fold recognition techniques, new methods for *ab initio* protein structure prediction have been proposed that include multiple sequence information (27–30). On the basis of the observation that homologous proteins fold into similar structures, these methods simultaneously generate decoys for a family of homologous sequences, under the constraint that these decoys look alike. These procedures were shown to generate better decoys than methods that rely on only a single sequence (28–30). Unlike fold recognition techniques, however, there is no sequence information for the model proteins, because all these decoys are computationally generated with the same sequence. In this paper, we show that this can be overcome by designing sequences that would stabilize these decoys. These designed sequences can then be used to identify near-native conformations by using sequence-based fold recognition. We refer to this strategy as the "reverse design approach" or RDA.

Methods

Given a protein sequence and a collection of structural models, the aim of our method is to use the sequence information implicit

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

Abbreviations: RDA, reverse design approach; cRMS, coordinate root-mean-square distance; PDB, Protein Data Bank; RAPDF, residue-specific all-atom conditional probability discriminatory function.

*To whom reprint requests should be addressed. E-mail: koehl@csb.stanford.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

in each model to identify the model closest to the native structure of the protein of interest. It is divided into two steps:

(*i*) Protein Sequence Design. A complete description of the protein design procedure is given in our previous work (31). We list here the modifications pertinent to this work. The program starts with the backbone of the template protein structure and constructs all possible side chains (including all 20 types) for each residue. Conformations for the side chains are taken from the rotamer library of Tuffery *et al.* (32), which has been corrected for duplicates. Interactions among these side chains are precomputed and stored efficiently in a large energy matrix. A full atom representation is used, and the energy function includes van der Waals interactions, electrostatics, and an environment free energy to account for the solvent (33). Once the energy matrix is built, the sequence design procedure is initialized by generating randomly 100 sequences with a fixed amino acid composition (taken from the native sequence of the template structure). These sequences are chosen to cover as much sequence space as possible by imposing the requirement that the sequence identity between any two sequences is less than 20%. Each sequence is threaded on the target backbone, and a full-atom model is built by using a self-consistent mean field approach (34, 35). The energy of this model drives the Monte Carlo optimization in sequence space. Moves are defined as exchange of the amino acid types of two positions chosen at random in the sequence. After each move, the new sequence is used to generate a new structural model, whose energy is compared with the energy of the model structure before the move. The new sequence is consequently accepted or rejected by using the Metropolis scheme (36). These moves maintain the amino acid composition of the designed sequence constant and are expected to ensure specificity, in accordance with the random energy model (37–39). The optimizations of the 100 sequences are performed in parallel. Efficiency is reached through usage of the precomputed energy matrix.

(*ii*) A New Metric in Sequence Space. In the second step, the sequences designed for each structural model are compared with the native sequence of the target protein. Percentage Sequence Identity (I) is an intuitive distance measure in sequence space. It does have a serious limitation: high sequence identity is a reliable measure of homology, whereas low sequence identity ($<25\%$) is much less informative (40). We first transform the percentage sequence identity into a distance D defined as $100 - I$. This still suffers from being meaningless for $D > 75$. The distance D is also capped, i.e., the maximum distance between two sequences is 100. There is therefore a need for a distance measure that provides a better estimate of the distance between two far-away sequences. We introduce a new distance measure D_{loc}, which preserves sequence identity as a distance for neighboring sequences, and approximate the distance D_{loc} between two distant proteins as the addition of a series of short hops between neighboring sequences. The best series of hops is computed efficiently by finding the shortest paths in the graph with edges connecting neighboring points.

The algorithm that computes all distances D_{loc} for a set of N sequences has two steps. The first step determines which sequences are neighbors in sequence space. We use the nearest-neighbor approach: for each sequence, we first find the K closest sequences on the basis of sequence identity. K is a parameter of our approach. These neighborhood relations are then represented as a weighted graph G over the N sequences, with edges of weight $D(i, j)$ between two neighboring sequences i and j and infinite weight for all other pairs of sequences. In the second step, we estimate the distance D_{loc} between all pairs of sequences by computing their shortest path in the graph. We use Floyd's algorithm for finding this shortest path. A similar graph distance was recently introduced in the ISOMAP approach for solving dimensionality reduction problems (41).

(*iii*) Representation of the Protein Sequence Space. Protein sequences occupy a high-dimensional space, which is difficult to visualize. Here we circumvent this problem by using a technique for nonlinear dimensionality reduction. We apply classical distance geometry (42) to the matrix of graph distances D_{loc}, constructing an embedding of the data in an M-dimensional Euclidean space E that best preserves the intrinsic geometry of the sequence space. Briefly, the N sequences in the data set are represented by N points with coordinates $\mathbf{r}$ in E. For simplicity, the N points are considered centered at the origin of the coordinate system of E (i.e., $\Sigma_{i=1}^{N} \mathbf{r}_i = 0$). We then consider the embedding matrix G defined by:

$$G(i, j) = \mathbf{r}_i \cdot \mathbf{r}_j \qquad [1]$$

where $\cdot$ is the vector dot product.

The matrix G is derived from the matrix of graph distances D_{loc} by using:

$$G(i, j) = -0.5\left(D_{\text{loc}}(i, j)^2 - \frac{1}{N}\sum_{k=1}^{N} D_{\text{loc}}(k, j)^2 - \frac{1}{N}\sum_{l=1}^{N} D_{\text{loc}}(i, l)^2 + \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N} D_{\text{loc}}(k, l)^2 \right). \qquad [2]$$

The optimal embedding of the N points in E is obtained by first diagonalizing G ($G = P'DP$) and then setting the coordinates $\mathbf{r}_i$ of point i as (42)

$$r_i(j) = \sqrt{D(j, j)}P(i, j) \qquad [3]$$

The quality of the embedding is measured by using:

$$R = \frac{1}{N}\sqrt{\sum_{i,j}(d_E(i, j) - D_{\text{loc}}(i, j))^2}\,. \qquad [4]$$

The summation extends over all pairs of sequences (i, j), and d_E is the Euclidean distance in E. R is called the residual variance for the embedding.

Results

(*i*) Sequence-Based Fold Recognition. We first tested our reversed design approach on a well-characterized fold recognition problem based on globin sequences. Globins constitute a large family of proteins with similar fold (43). This family includes at least five subfamilies: myoglobins, hemoglobins, erythrocrourins, leghemoglobins, and plant phycocyanins. We have recently shown that we could design a sequence stable for the myoglobin fold that recognizes myoglobins but not hemoglobins when tested against a database of protein sequences (44). In this study, we plan to use the information contained in similarly designed sequences to help identify the correct fold of a globin sequence. We consider four different globins: two distantly related myoglobins [from sperm whale and yellowfin tuna, whose Protein Data Bank (PDB) codes are 5mbn and 1myt, respectively], one hemoglobin (from marine bloodworm, PDB code 2hbg), and one leghemoglobin (from yellow lupine, PDB code 2gdm). A detailed comparison of 5mbn with the three other globins is given in Table 1. Although all four globins are structurally similar, pairwise sequence comparison identifies only 1myt as homologous to 5mbn with an E value of 8×10^{-25}. We have designed 100 sequences for each of the four proteins by using the procedure

Table 1. Comparison of the sperm whale myoglobin (PDB code 5mbn) with the myoglobin of yellowfin tuna (PDB code 1myt), the hemoglobin of bloodworm (PDB code 2hbg), and the leghemoglobin of yellow lupine (PDB code 2gdm)

Master, 5mbn	1myt	2hbg	2gdm
Sequence identity (%)*	44.5	25	23
PSIBLAST *E*-value (1 cycle)†	8×10^{-25}	>1	>1
PSIBLAST *E*-value (5 cycles)†	4×10^{-25}	10^{-9}	10^{-15}
Structure similarity (cRMS, Å)‡	1.07	1.65	2.16

*The sequence identities are deduced from sequence alignments based on the BLOSUM 62 substitution matrix (50) with a penalty for gap opening of 12 and a penalty for gap extension of 1.

†The sequence of 5mbn was compared to the nonredundant SWISSPROT (45, 46) database of April 2001 (640,428 sequences) by using PSIBLAST (17) with and without five cycles of iteration.

‡We have used STRUCTAL (52) for protein structure superposition.

described above. The relative positions of these 400 sequences and of the corresponding four native sequences are shown in a two-dimensional projection of the sequence space whose metric is defined by the nearest-neighbor distance D_{loc} (Fig. 1*A*). It is noteworthy that the observed dimensionality of this sequence space is three (Fig. 1*B*). The two-dimensional embedding identifies four clusters of sequences, each corresponding to one of the four globins. These four clusters are connected with a graph based on a number of neighbors K equal to 40, indicating that they share a common region in sequence space. It should be noted that the two sequence clusters corresponding to the two myoglobin structures are already connected with $K = 2$. In each case, the native sequence of a protein is within, or very close to, the cluster of sequences that were designed for that protein. Interestingly, the two clusters corresponding to the two myoglobins overlap, whereas the clusters for the hemoglobin and leghemoglobin are further apart.

Recent fold recognition methods include multiple sequence alignment information. When the sequence of 5mbn is compared with the SWISSPROT (45, 46) nonredundant database by using PSIBLAST (17) with five iterations, all three globins (1myt, 2hbg, and 2gdm) are confidently detected as homologous to 5mbn (Table 1). PSIBLAST proceeds by building a profile at the end of each iteration, incorporating information on all sequences found to be similar to the test sequence. This profile is used at the next iteration, yielding more sensitive detection of distant homology (2). This is observed here in this study of myoglobin (Table 1 shows a comparison of PSIBLAST results obtained with one iteration and five iterations). Similarly, the sequences designed for each globin fold considered in the test case described above can be included in profiles, which would in turn be compared with test sequences (44). Here we find that a direct measure of the mean of the distances, D_{loc}, between the test sequence S and the sequences designed for the target fold F also provides a sensitive estimate of the compatibility between the test sequence and the fold. In the comparison of the native sequence of 5mbn with the four clusters of sequences designed from the structures of the four globins, 5mbn, 1myt, 2hbg, and 2gdm, the mean D_{loc} distances correlate well with the structural similarity between 5mbn and the target structures (Fig. 1*C*).

(*ii*) *Ab Initio* Structure Prediction: Identification of Near-Native Conformations by Using Sequence Design. Most *ab initio* protein structure prediction methods proceed in two steps. First, a large data set of model structures is built, either systematically (9) or

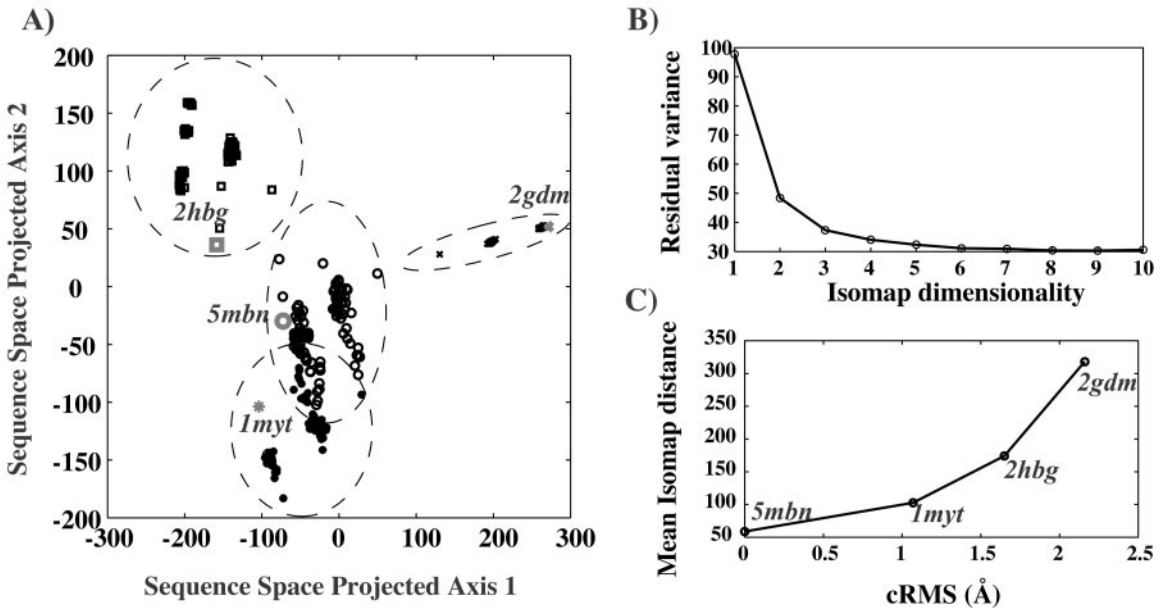

Fig. 1. A well-characterized fold recognition problem involves the globin family. The restricted library of folds consists of four globins: two myoglobins (5mbn and 1myt), one hemoglobin (2hbg), and one leghemoglobin (2gdm). One hundred sequences are optimized for stability and specificity for each of the four proteins. The corresponding 400 sequences are pooled with the native sequences and given as input to ISOMAP (41). The underlying distance for close neighbors in sequence space is $D = 100 - I$, where I is the percent sequence identity between the two sequences [computed on the basis of the structure alignment of their corresponding structures; we use STRUCTAL (51) for protein structure superposition]. The neighborhood of each sequence includes its 40 closest sequences (i.e., $K = 40$; see text). (*A*) A two-dimensional projection of the sequence space spanned by these 404 sequences is shown. The designed sequences are shown in black with small marks (○ for 5mbn, * for 1myt, □ for 2hbg, and x for 2gdm), whereas the native sequences are shown in gray with large marks. The sequences are found in four clusters, each corresponding to one of the globin structures. The native sequence of each protein is found within, or very close to, the sequences designed for that protein. (*B*) The residual variance of the ISOMAP embedding is plotted versus the dimensionality of the projection. The dimensionality of the sequence space covered by the four globins is estimated to be 2. (*C*) The mean ISOMAP distance (for $K = 40$) between the native sequence of 5mbn and each family of sequences designed for the four globin structures is plotted versus the cRMS between the corresponding structure and 5mbn.

through assembly of small parts (8). Second, near-native conformations are identified in this data set (26). Recent progress has concerned the generation of the decoys, in the hope of generating better models and enriching the set with near-native structures. Recognition of near-native conformations in this data set is presumably a harder problem than fold recognition, as it does not benefit from any additional sequence information for the model structures. RDA was designed to circumvent this difference by generating sequences by computer design experiments that would stabilize the model structures.

We first tested RDA on the identification of near-native conformations for 1ctf. 1ctf, the C-terminal domain of the L7/L12 ribosomal protein of *Escherichia coli*, is a small highly stable protein of 68 residues, which has been extensively used as a test protein for *ab initio* prediction methods. A data set of 1,000 model structures for 1ctf was downloaded from the web server (http://depts.washington.edu/bakerpg/) of David Baker's group (8). These models were generated by using ROSETTA, a program for *ab initio* protein structure prediction that assembles a protein fold from a database of small fragments (8). Each decoy comes as a PDB file containing the coordinates of the main-chain atoms (N, CA, C, and O) and CB for each residue. We generated a full-atom model for each of these decoys by adding the side chains using our self-consistent mean-field approach (33). The data set contains the native structure of 1ctf and 999 decoys whose root-mean-square distances (cRMS) computed over C^{α} compared with 1ctf are in the range 3.45–18.7 Å. All these models contain similar secondary structures. The compatibility of the native sequence of 1ctf with each model in the data set was measured by using a full-atom database-derived scoring function similar to the residue-specific all-atom conditional probability discriminatory function (RAPDF) potential introduced by Samudrala and Moult (25). Ten models, including the native structure of 1ctf, were selected on the basis of the criteria that they are poorly discriminated by the RAPDF scoring function. They correspond to near-native models with high energies and nonnative models with low energies, some with energies even lower than that of the native structure (Fig. 2*A*).

One hundred sequences were designed for each of the 10 models. All sequences designed for the native structure of 1ctf are very similar to the native sequence of 1ctf (average sequence identity of 37%, varying between 21 and 48%). In contrast, the sequences designed for the nonnative model at 11.4 Å from 1ctf show much-reduced sequence identity with the native sequence of 1ctf (average sequence identity 20%, varying from 13 to 30%). The 1,000 designed sequences together with the native sequence of 1ctf were pooled, and all distances D between any two sequences were computed and stored in a distance matrix. These distances were converted to local distances D_{loc} by using different values for the number K of neighbors for each sequence used to define the underlying graph (see *Methods*). Low values of K will restrict connectivity to sequences that are very similar to each other, whereas larger values of K allow detection of sequences that are more distantly related. For $K = 2$, only the 100 sequences generated for the native structure of 1ctf are found in the vicinity of the native sequence of 1ctf (Fig. 2*B*). For $K = 5$, the graph is extended to all 100 sequences designed for the near-native model at 3.9 Å from 1ctf (Fig. 2*C*). More near-native models are detected for $K = 11$ and $K = 15$ (Fig. 2 *D* and *E*). All three nonnative models with cRMS greater than 7 Å are not connected to the graph at $K = 15$ and become connected only when $K = 35$. By using the mean distance D_{mean} between the native sequence of 1ctf and the cluster of sequences generated for a model structure, the RDA is able to recover a good ranking (in terms of structure) of the different models for 1ctf (Fig. 3*A*). Fig. 2 shows that the different sequence clusters are connected sequentially to the native sequence of 1ctf with increasing values of K, in an order that reflects the structural difference between

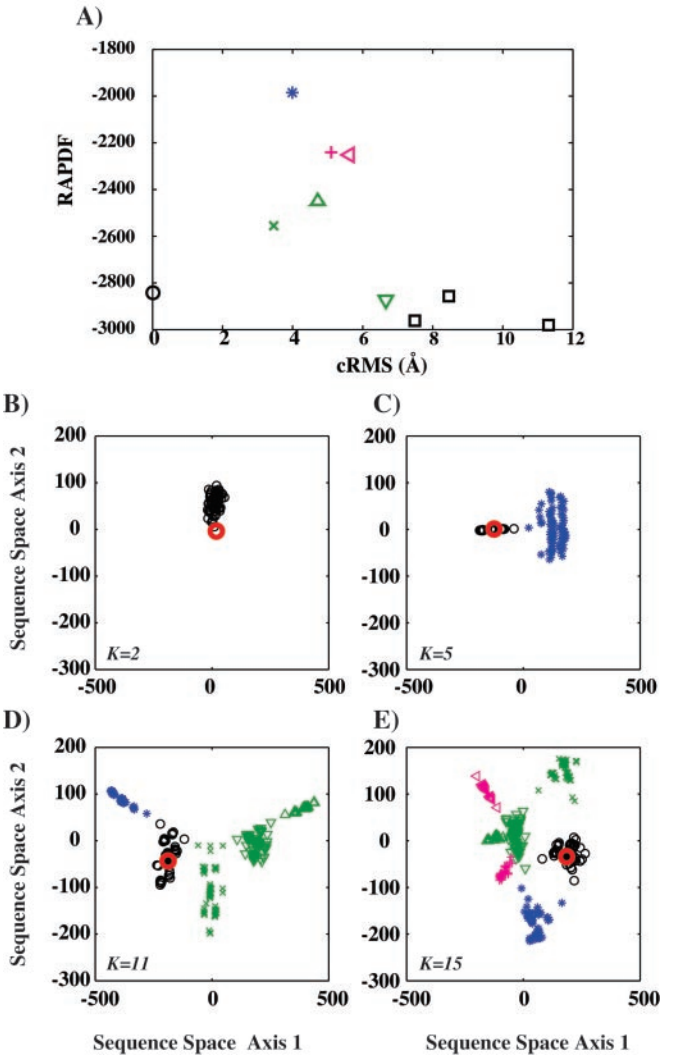

Fig. 2. Recognition of native-like structures for 1ctf among nonnative models. (*A*) Ten model structures (including the native structure) were selected from the data set of 1,000 decoys generated by the group of David Baker for 1ctf (8). The RAPDF score of these 10 models is plotted versus cRMS. These models were chosen such that the all-atom potential of mean force RAPDF (25) fails to distinguish near-native from nonnative conformations. (*B–D*) One hundred sequences were designed for each of the 10 models selected for 1ctf. The corresponding 1,000 sequences together with the native sequence of 1ctf are given as input to ISOMAP. Two-dimensional embeddings of the sequence space covered by these 1,001 sequences are shown for the increasing value of *K*, the number of sequences that defines the neighborhood of each sequence in the underlying graph. The marks used to identify each sequence cluster are consistent with *A*. The native sequence is shown in red as a big circle (○). Note that the sequences designed for each particular model structure cluster in sequence space.

the model used to generate the sequences and the native structure of 1ctf. On the basis of this observation, we computed for each sequence cluster K_{first} the value of K for which the first connection to the native sequence of 1ctf is observed. In Fig. 3*B*, we plot K_{first} against the cRMS between the model structure and 1ctf. K_{first} is found to correlate well with cRMS and can be used as an alternative ranking function to the mean distance D_{mean} used in Fig. 3*A*.

The best near-native conformation in the 1ctf decoy dataset of David Baker differs from the native structure of 1ctf by 3.45 Å. To test the ability of RDA to identify near-native structures closer to the true structure of the protein under study, we chose

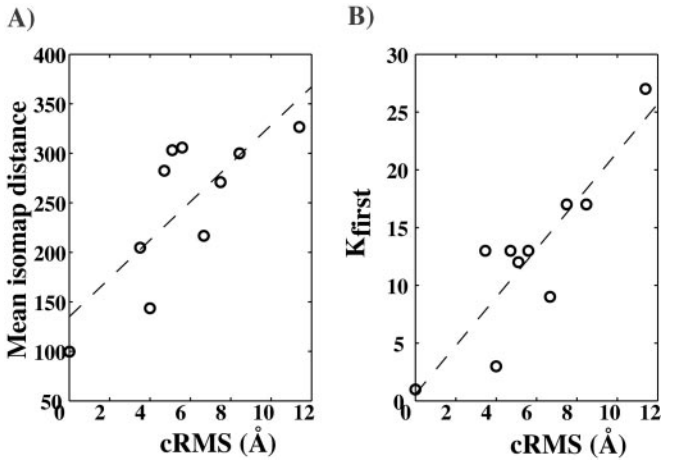

Fig. 3. Plots of score versus cRMS for the data set of 10 models selected for 1ctf (see legend of Fig. 2). (*A*) The mean ISOMAP distances (for $K = 35$) between the native sequence of 1ctf and the families of sequences designed for the model structures are plotted versus cRMS. A significant correlation of 0.77 is observed between these distances and cRMS for nonnative models. (*B*) The first value of K, K_{first}, for which a connection is observed between the family of sequences designed for a model structure, is plotted versus the cRMS of the model to the native structure of 1ctf. A significant correlation of 0.88 is observed between K_{first} and cRMS. The dotted lines in *A* and *B* show the best line fits to the data.

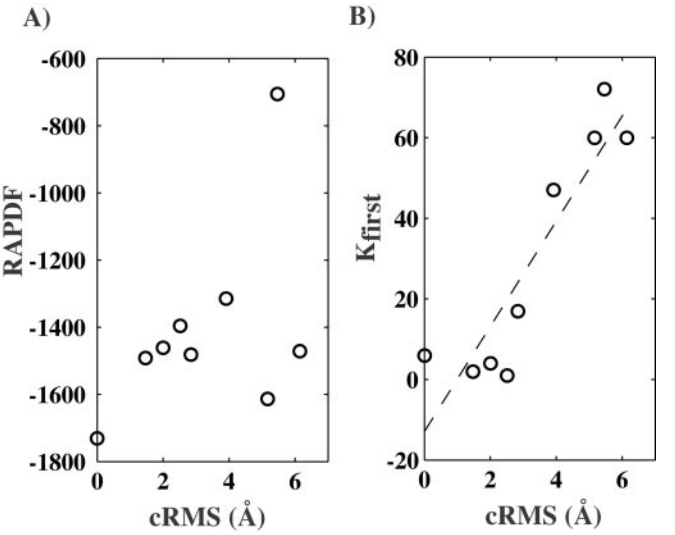

Fig. 4. Recognition of native-like structures for 4rxn among nonnative models. (*A*) Nine model structures (including the native structure) were selected from the data set of 638 decoys generated by B. Park and M.L. (6). The RAPDF score of these 10 models is plotted versus cRMS. No correlation between the RAPDF score and cRMS is observed. (*B*) The first value of K, K_{first}, for which a connection is observed between the family of sequences designed for a model structure is plotted versus the cRMS of the model to the native structure of 4rxn. A significant correlation of 0.91 is observed between K_{first} and cRMS. The dotted line shows the best line fit to the data.

as a second test set for our method the protein rubredoxin (PDB code 4rxn). We downloaded a data set of 678 decoys for 4rxn from the "Decoys 'R' Us" web site (http://dd.stanford.edu/). These models were generated by systematically varying a few structural degrees of freedom of the protein (6). Each decoy comes as a PDB file containing the coordinates of all atoms of the protein. The data set contains the native structure of 4rxn, and 677 decoys whose cRMS compared with 4rxn are in the range 1.35–9.57 Å. The compatibility of the native sequence of 4rxn with each model in the data set was measured by using the RAPDF scoring function (25). Nine models, including the native structure of 4rxn, were selected on the basis of the criteria that the RAPDF scoring function for these proteins is poorly correlated with cRMS (Fig. 4*A*). One hundred sequences were designed for each of the nine models. The 900 designed sequences together with the native sequence of 4rxn were pooled, and all distances D between any two sequences were computed and stored in a distance matrix. These distances were converted to local distances D_{loc} by using different values for the number K of neighbors used to define the underlying graph. For each cluster of 100 sequences, we compute K_{first}, the value of K for which the first connection with the native sequence of 4rxn is observed. These values are plotted against the cRMS between the model and the native structure of 4rxn in Fig. 4*B*. K_{first} correlates well with cRMS ($r = 0.91$), providing a ranking measure that can be used to distinguish near-native structures for 4rxn from nonnative models. Interestingly, the values of K for the first connection do not distinguish between the native structure of 4rxn and decoys that are very close to 4rxn (cRMS < 2; Fig. 4*B*). The mean ISOMAP distances (for $K = 80$) between the native sequence of 4rxn and the families of sequences designed for the model structures show a similar correlation with cRMS ($r = 0.79$; results not shown).

Discussion

The results shown above demonstrate the ability of our strategy for protein sequence design to provide useful information on the sequence space compatible with a given protein structure. We also show that the family of sequences designed for a protein contains enough information about its structure that it is able to identify its native structure.

Eisenberg and coworkers originally introduced the concept of projecting the structural information of a protein model into sequence space by defining three-dimension–one-dimension (3D-1D) profiles. These profiles were initially used for fold recognition experiments (12) and then applied to the problem of the assessment of protein models (47). In that method, each residue position in the protein model is characterized by its environment (which combines local secondary structure information, accessibility, and number of polar contacts) and is represented by a row of 20 numbers in the profile, corresponding to the 20 amino acid types. These numbers are the statistical preferences of each amino acid type for this environment. The 3D-1D profiles are based on the so-called frozen approximation, i.e., the hypothesis that residue environments are conserved in protein with similar folds. This hypothesis has been recently questioned (48, 49). The method described in this paper does not assume the frozen approximation (31, 44), in that it designs a sequence profile for each model structure.

Both fold recognition techniques and *ab initio* protein structure prediction require a method of comparing three-dimensional models of a particular sequence and determining which model is closer to the native structure. This is often accomplished by using a scoring scheme based on either a physical energy function or a knowledge-based potential, or even a combination of the two (for review, see ref. 26). These scoring functions are usually tested on standard test sets such as Prostar (http://prostar.carb.nist.gov/) and "Decoys 'R' Us" (http://dd.stanford.edu) for their ability to recover native and near-native structures among misfolded models. A "good" scoring function should provide a significant correlation between the score of a model and its cRMS to the corresponding native protein structure. Although significant progress has been made in designing more discriminative scoring functions, the latter remain the bottleneck in protein structure prediction: the two recent critical assessment of protein structure prediction method

experiments have shown that most predictors had better models in their decoy sets than the one they submitted as "best" predictions. In this paper, we have shown that protein sequence design can help recognize near-native folds among nonnative folds even when other techniques fail (Figs. 3 and 4). We are well aware that this problem remains difficult, and that it is very likely that a test case can be designed such that the procedure presented here fails. We believe, however, that it is the combination of several methods based on different principles that will help solve this problem.

The computing time required by our method represents its major limitation: the full design of 100 sequences to fit a single structure of a 68-residue protein such as 1ctf requires 25 h of central processing unit time on a Compaq DS20 computer. At this stage, it cannot be used to test a large collection of decoy structures and is limited to the testing of a small well selected data set of model structures, most probably as a second round of screening after the initial models have been evaluated with other scoring functions. We are currently working on new techniques that would reduce this computing time.

An interesting aspect of our procedure is the introduction of D_{loc}, a distance metric in sequence space, based on a neighborhood-connecting graph. We show that this distance identifies hierarchically sequence and structure similarities between proteins and should prove useful in studies of protein evolution.

This work was supported by grants to M.L. from the Department of Energy (DE-FG03-95ER62135) and the National Institutes of Health (GM 41455).

1. Anfinsen, C. (1973) *Science* **181,** 223–230.
2. Murzin, A. (2001) *Nat. Struct. Biol.* **8,** 110–112.
3. Chothia, C. & Lesk, A. (1986) *EMBO J.* **5,** 823–826.
4. Sander, C. & Schneider, R. (1991) *Proteins Struct. Funct. Genet.* **9,** 56–68.
5. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992) *Nature (London)* **358,** 86–89.
6. Park, B. H. & Levitt, M. (1996) *J. Mol. Biol.* **258,** 367–392.
7. Simons, K. T., Kooperberg, C., Huang, E. S. & Baker, D. (1997) *J. Mol. Biol.* **268,** 209–225.
8. Simons, K. T., Bonneau, R., Ruczinski, I. & Baker, D. (1999) *Proteins Struct. Funct. Genet.* **43,** 171–176.
9. Xia, Y., Huang, E., Levitt, M. & Samudrala, R. (2000) *J. Mol. Biol.* **300,** 171–185.
10. Bryant, S. H. & Amzel, L. M. (1987) *Int. J. Pept. Protein Res.* **29,** 46–52.
11. Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M. J. (1990) *J. Mol. Biol.* **216,** 167–180.
12. Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991) *Science* **253,** 164–170.
13. Godzik, A., Kolinski, A. & Skolnick, J. (1992) *J. Mol. Biol.* **227,** 227–238.
14. Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 9029–9033.
15. Mirny, L. A. & Shakhnovich, E. I. (1998) *J. Mol. Biol.* **283,** 507–526.
16. Jones, D. (1999) *J. Mol. Biol.* **287,** 797–815.
17. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
18. Karplus, K., Sjolander, K., Barrett, C., Cline, M., Haussler, D., Hughey, R., Holm, L. & Sander, C. (1997) *Proteins Struct. Funct. Genet.* **S1,** 134–139.
19. Rychlewski, L., Jaroszewski, L., Li, W. Z. & Godzik, A. (2000) *Protein Sci.* **9,** 232–241.
20. DiFrancesco, V., Geetha, V., Garnier, J. & Munson, P. (1997) *Proteins Struct. Funct. Genet.* **S1,** 123–128.
21. Halgren, T. A. (1995) *Curr. Opin. Struct. Biol.* **5,** 205–210.
22. Jones, D. T. & Thornton, J. M. (1996) *Curr. Opin. Struct. Biol.* **6,** 210–216.
23. Thornton, J. & Jones, D. (1996) *Curr. Opin. Struct. Biol.* 6.
24. Park, B. H., Huang, E. S. & Levitt, M. (1997) *J. Mol. Biol.* **266,** 831–846.
25. Samudrala, R. & Moult, J. (1998) *J. Mol. Biol.* **275,** 895–916.
26. Laziridis, T. & Karplus, M. (2000) *Curr. Opin. Struct. Biol.* **10,** 139–145.
27. Keasar, C., Elber, R. & Skolnick, J. (1997) *Folding Des.* **2,** 247–259.
28. Keasar, C., Tobi, D., Elber, R. & Skolnick, J. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 5880–5883.
29. Badretdinov, A. & Finkelstein, A. V. (1998) *J. Comp. Biol.* **5,** 369–376.
30. Bonneau, R., Strauss, C. E. M. & Baker, D. (2001) *Proteins Struct. Funct. Genet.* **43,** 1–11.
31. Koehl, P. & Levitt, M. (1999) *J. Mol. Biol.* **293,** 1161–1181.
32. Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. (1991) *J. Biomol. Struct. Dyn.* **8,** 1267–1289.
33. Koehl, P. & Delarue, M. (1994) *Proteins Struct. Funct. Genet.* **20,** 264–278.
34. Koehl, P. & Delarue, M. (1994) *J. Mol. Biol.* **239,** 249–275.
35. Koehl, P. & Delarue, M. (1996) *Curr. Opin. Struct. Biol.* **6,** 222–226.
36. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953) *J. Chem. Phys.* **21,** 1087–1092.
37. Shakhnovich, E. I. & Gutin, A. M. (1993) *Protein Eng.* **6,** 793–800.
38. Shakhnovich, E. I. & Gutin, A. M. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 7195–7199.
39. Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1997) *Biophys. J.* **73,** 3192–3210.
40. Rost, B. (1999) *Protein Eng.* **12,** 85–94.
41. Tenenbaum, J., deSilva, V. & Langford, J. (2000) *Science* **290,** 2319–2323.
42. Havel, T. F., Kuntz, I. D. & Crippen, G. M. (1983) *Bull. Math. Biol.* **45,** 665–720.
43. Bashford, D., Chothia, C. & Lesk, A. M. (1987) *J. Mol. Biol.* **196,** 199–216.
44. Koehl, P. & Levitt, M. (1999) *J. Mol. Biol.* **293,** 1183–1193.
45. Bairoch, A. & Böckman, B. (1991) *Nucleic Acids Res.* **19,** 2247–2249.
46. Bairoch, A. & Apweiler, R. (2000) *Nucleic Acids Res.* **28,** 45–48.
47. Lüthy, R., Bowie, J. U. & Eisenberg, D. (1992) *Nature (London)* **356,** 83–85.
48. Russel, R. B. & Barton, G. J. (1994) *J. Mol. Biol.* **244,** 332–350.
49. Rodionov, M. A. & Blundell, T. L. (1998) *Proteins Struct. Funct. Genet.* **33,** 358–366.
50. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 10915–10919.
51. Subbiah, S., Laurents, D. V. & Levitt, M. (1993) *Curr. Biol.* **3,** 141–148.