Protein Science (2000), *9*:812–819. Cambridge University Press. Printed in the USA.

Scoring functions in protein folding and design

RUXANDRA I. DIMA,[1] JAYANTH R. BANAVAR,[1] AND AMOS MARITAN[2]

[1] Department of Physics and Center for Materials Physics, 104 Davey Laboratory, The Pennsylvania State University, University Park, Pennsylvania 16802
[2] International School for Advanced Studies (S.I.S.S.A.),Via Beirut 2-4, 34014 Trieste, INFM and the Abdus Salam International Center for Theoretical Physics, Trieste, Italy

(RECEIVED August 16, 1999; FINAL REVISION November 15, 1999; ACCEPTED November 15, 1999)

Abstract

We present an analysis of the assumptions behind some of the most commonly used methods for evaluating the goodness of the fit between a sequence and a structure. Our studies on a lattice model show that methods based on statistical considerations are easy to use and can capture some of the features of protein-like sequences and their corresponding native states, but unfortunately are incapable of recognizing, with certainty, the native-like conformation of a sequence among a set of decoys. Meanwhile, an optimization method, entailing the determination of the parameters of an effective free energy of interaction, is much more reliable in recognizing the native state of a sequence. However, the statistical method is shown to perform quite well in tests of protein design.

Keywords: energy-based approach; Monte Carlo procedure; perceptron optimization procedure; statistical approach; three-dimensional lattice model

The problems of protein folding, protein design, and docking of ligands to protein structures are among the most interesting and challenging problems in molecular biology. They all refer to the relationship between sequences of amino acids and their corresponding native state conformation. A key ingredient for the solution of these problems entails the development of a scoring function that can identify the native-like fold of a given sequence of amino acids from a pool of decoy conformations. There exist at least two different types of approaches to this problem: one that uses a scoring scheme based on statistical considerations applied to a database of sequences and structures, and another that uses only energetic considerations to extract the quantities that make up the scoring function. In the first approach, to find the most likely structure (on statistical grounds) for a given protein sequence, one first determines the distribution of amino acids in various environments (Bowie et al., 1991; Simons et al., 1997, 1999) and/or the distribution of the contacts between the 20 types of amino acids in proteins with known tertiary structure (Tanaka & Scheraga, 1976; Miyazawa & Jernigan, 1985, 1996, 1999; Hendlich et al., 1990; Sippl, 1990; Jones et al., 1992; Sippl & Weitckus, 1992; Kolinski et al., 1993). Then, based on either the quasichemical approximation and Boltzmann statistics (Tanaka & Scheraga, 1976; Miyazawa & Jernigan, 1985, 1996, 1999; Hendlich et al., 1990; Sippl, 1990; Jones et al., 1992; Sippl & Weitckus, 1992) or on Bayes theorem (Simons et al., 1997, 1999), one converts these distributions into a scoring function. For a given sequence, the structure that corresponds to the best score is considered to be the most native-like conformation. This method has been used in a wide range of problems: to identify structures from a pool of decoys that can house a sequence of amino acids whose tertiary structure is a priori unknown (Bowie et al., 1991; Sippl & Weitckus, 1992; Simons et al., 1997, 1999), to judge the quality of protein structure models (Luthy et al., 1992; Wilmanns & Eisenberg, 1993; Simons et al., 1997), to predict docking of ligands to protein structures (Pellegrini & Doniach, 1993), to simulate the folding of a protein (Wilson & Doniach, 1989; Skolnick & Kolinski, 1990; Sun, 1993; Kolinski & Skolnick, 1994), and to identify the native fold of a protein sequence among many incorrect alternatives (Hendlich et al., 1990; Bryant & Lawrence, 1991; Jones et al., 1992; Miyazawa & Jernigan, 1996; Park & Levitt, 1996). Thomas and Dill (1996) presented a thorough analysis through a lattice model study of the degree of accuracy of statistical potentials extracted from protein structures based on Boltzmann statistics and on the quasi-chemical approximation (Tanaka & Scheraga, 1976; Miyazawa & Jernigan, 1985; Sippl, 1990; Jones et al., 1992; Kolinski et al., 1993). Their conclusion was that these potentials are not accurate enough to lead to good predictions regarding the folding of a sequence of amino acids because the method neglects the excluded volume in proteins and the use of the Boltzmann distribution to convert frequencies of contacts between various amino acids into energies of interaction is not firmly grounded.

The second method (Maiorov & Crippen, 1992; Clementi et al., 1998; Seno et al., 1998; von Mourik et al., 1999) starts from the idea that the interaction energies between amino acids parametrizing a coarse grained free energy must be such that the energy of a sequence in its own native state is lower than in any other alter-

Reprint request to: Ruxandra Dima, Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742; e-mail: dima@ipst.umd.edu.

native conformation. For each sequence in a data bank, assuming a simple contact free energy, one obtains a set of linear inequalities involving the unknown interaction parameters. These inequalities can then be solved to obtain the interaction potentials that give an energetic measure of the goodness of the fit between a sequence and a structure. This method is extremely powerful on lattice models (Seno et al., 1998; von Mourik et al., 1999). When applied to real proteins, there are difficulties in generating viable alternative conformations that compete significantly with the native structure in housing each of the sequences in the training set. However, using decoy structures obtained by simple gapless threading, the performance of the method is slightly superior to those of previously proposed strategies despite the fact that gapless threading does not produce sufficiently competitive alternatives. The purpose of our study is to evaluate the degree of performance of these approaches in an unbiased manner.

There have been a large number of studies that have considered the strengths and weaknesses of these two approaches (Goldstein et al., 1992a, 1992b; Rooman & Wodak, 1995; Hao & Scheraga, 1996a, 1996b, 1999; Koretke et al., 1996, 1998; Mirny & Shakhnovich, 1996; Skolnick et al., 1997; Chiu & Goldstein, 1998; Vendruscolo & Domany, 1998; Zhang, 1998; Zhang & Skolnick, 1998; Betancourt & Thirumalai, 1999). One of the main assumptions used in both of the above approaches is that a pairwise contact Hamiltonian is a good approximation to the real Hamiltonian in proteins. The study of Vendruscolo and Domany (1998) showed, using a contact map method to generate compact conformations that compete significantly with the native one in housing a given sequence, that this may not be firmly grounded. They concluded that higher order terms (at least three-body terms) must be considered for a sequence to recognize its native state structure from a pool of conformations. The same conclusion was drawn by Betancourt and Thirumalai (1999) based on the relatively high sensitivity of predicted native state structures to variations in two different (but well-correlated) sets of statistically determined interaction potentials between amino acids. The study of Mirny and Shakhnovich (1996), which used an optimization procedure based on energy considerations on both a lattice model and on real proteins, also reached the conclusion that a more accurate representation of the energy function (including multibody effects, local conformational preferences, etc.) is likely to be necessary to better discriminate the native state from a set of decoys.

In general, one might expect that a reliable set of parameters will be useful in both folding and inverse folding (design). It is commonly believed that the parameters extracted using the statistically based approach can be useful in folding studies by recognizing the native state conformation of a sequence from a set of structures, but their applicability in design problems had been questioned. In their study, Rooman and Wodak (1995) argued that these interaction potentials can also be useful for design, if one corrects them for many-body effects that dominate the frequencies of interaction between residues as compiled from a data bank of proteins. It has been known that these potentials are capable of providing a measure of energy-like quantities, such as hydrophobicities comparable with the values measured in experiments (Zhang et al., 1997). Zhang (1998) showed, using an HP lattice model, that the statistically extracted energies of interaction lie within the range of energies defined by the thermodynamical foldability and the kinetic accessibility of the proteins under study. One of the main conclusions of the study of Thomas and Dill (1996) was that the approximate choice of the reference state in the quasi-chemical approach with the chain connectivity being neglected is not entirely justifiable. Skolnick et al. (1997) addressed this issue through the extraction of the interaction potentials using the standard statistical approach, but with different choices of a reference state. Working on a set of real proteins, they showed that the results given by the quasi-chemical approach correlate well with the results obtained when using a reference state that took into consideration the constraints of the chain connectivity, the protein chain compactness, and the presence of regular secondary structure elements. Other studies (Goldstein et al., 1992a, 1992b; Hao & Scheraga, 1996a, 1996b, 1999; Koretke et al., 1996, 1998; Chiu & Goldstein, 1998; Zhang & Skolnick, 1998) have shown that the optimization of the knowledge-based interaction potentials for protein folding so as to maximize the Z-scores (Luthy et al., 1992) of the proteins in a given data bank usually leads to an improved performance in the recognition of the native state conformations of a set of sequences.

In this study, we address a different question from those considered previously. Almost all of the above approaches implicitly assume that the conformation corresponding to the best statistical score is the native state structure of the sequence under study is true. Here we concentrate on presenting a direct check of this fundamental assumption, rather than on uncovering the implications of other assumptions such as chain connectivity, excluded volume, and the regularity of secondary structure on the determination of accurate effective potentials (Mirny & Shakhnovich, 1996; Thomas & Dill, 1996; Skolnick et al., 1997; Zhang & Skolnick, 1998).

To illustrate the main points of the statistical method, we will follow the approach of Baker's group (Simons et al., 1997, 1999). This approach leads to practically the same scoring function as the other statistical methods and has the advantage that the derivation does not rely on the assumption that a database of protein sequences with their native state conformations obeys some sort of Boltzmann statistics nor does it require the knowledge of a reference state. Bayes theorem states that the probability that the sequence folds into the structure is

$$P(\text{structure}|\text{sequence}) = P(\text{structure})\,\frac{P(\text{sequence}|\text{structure})}{P(\text{sequence})}. \tag{1}$$

Here, $P(\text{structure})$ is the probability to find the structure (independent of the sequence) as, for example, characterized by its type (alpha and/or beta) and its degree of compactness. $P(\text{sequence}|\text{structure})$ is the probability that the sequence has the structure as its native state, and $P(\text{sequence})$ is the probability to find the sequence (independent of any structure). As long as one is interested in folding a given sequence using a pool of conformations, $P(\text{sequence})$ is the same for all possible structures and, therefore, it plays no role in the analysis. The expression for $P(\text{sequence}|\text{structure})$ used in Simons et al. (1997, 1999) is a combination of two terms:

$$P(\text{sequence}|\text{structure}) = P_{env}P_{\text{pair}} \tag{2}$$

where

$$P_{env} = \prod_i P(aa_i|E_i) \tag{3}$$

is the product, over all residues along the chain, of the probability to find each residue aa_i from a sequence in a certain structural environment E_i and represents a measure of the goodness of the alignment between a sequence and a structure, in the spirit of the profile method of Bowie et al. (1991).

$$P_{\text{pair}} = \prod_{i \leq j} \frac{P(r_{ij} \,|\, aa_i, aa_j)}{P(r_{ij})} \tag{4}$$

is inspired by the method of Sippl (1990) for the extraction of the potentials of mean force and is a product over all possible pairs of amino acids of the frequency of finding a given amino acid pair (aa_i, aa_j) in contact at a certain distance r_{ij} along the sequence normalized to the probability of finding residues of a structure at the same distance irrespective of the identity of the amino acid pairs.

The scoring function used in Simons et al. (1997, 1999) is just the negative logarithm of the expression in Equation 1. The procedure is to consider, for a given amino acid sequence, its score on each target conformation, and to select as the native-like structure the one that corresponds to the lowest score. This procedure relies on the assumption that the structure related to the lowest score (i.e., the one having the highest statistical probability) is likely to be the native conformation of the sequence under study. It is this assumption that we are investigating here. The question we try to answer is: how well does the structure with the best statistical score correlate with the true native state conformation of the sequence? The answer for real proteins is obviously unknown, due to an imperfect knowledge of the interaction energy. To circumvent this problem, we used a lattice model that has been shown to possess many of the characteristics of real proteins (Lau & Dill, 1989; Chan & Dill, 1993; Li et al., 1996), and that has the big advantage that we can exactly enumerate all the relevant conformations for a sequence. We built a database containing the same type of information as the Protein Data Bank (PDB) (Bernstein et al., 1977)—a set of sequences along with their unique native state conformations. Then we extracted the statistical parameters in Equation 2 and we tested their ability to reproduce the correct folds of a set of sequences.

The model we used is similar to the HP model of Lau and Dill (1989). It consists of sequences of length 27 made up of 20 types of amino acids, while the space of conformations consists of all 103,346 maximally compact self-avoiding walks that fit on a $3 \times 3 \times 3$ cubic lattice. The restriction to maximally compact conformations is valid in this coarse-grained model of a protein when there is sufficiently strong overall negative (attractive) shift of all the interactions between the amino acids. We also assumed that all conformations appear with equal probability, making the factor $P(\text{structure})$ in Equation 1 merely an additive constant in the scoring function, which could be neglected in subsequent calculations. In such a model, two residues are considered to be in contact if they are neighbors in the lattice but are not next to each other along the chain. We chose, as in any HP model (Lau & Dill, 1989; Li et al., 1996) to represent the energy of a sequence S on a conformation Γ by an effective pairwise contact Hamiltonian:

$$H(S, \Gamma) = \sum_{i \leq j} w_{ij} e_{ij} \tag{5}$$

where i and j represent the types of amino acids in the sequence, w_{ij} is the total number of $i - j$ contacts, and e_{ij} is the energy of interaction between the amino acids type i and j. All sequences were chosen such that each location of the chain was assigned an amino acid generated at random according to its frequency of occurrence in nature (Creighton, 1993), and the 210 interaction energies between the amino acids were taken from Table III of Kolinski et al. (1993) with an assumed large negative shift in each of these values. These energy parameters were obtained as the negative logarithm of the observed frequency of the particular pair in a data bank of proteins collected from the PDB relative to the random frequency of the corresponding pair calculated employing the Bragg–Williams approximation.

Using this energy function, we generated a data bank of 748 sequences with unique ground state conformations and with energy gaps between the lowest energy state and the first excited state larger than 1.4 (which is a characteristic of less than 2% of the total number of sequences). This cutoff value for the gap was chosen to ensure that these sequences are thermodynamically stable in their native conformations.

The next step was to determine the quantities that go into Equation 3 and 4. The parameters from Equation 3 can be easily extracted from a set of sequences with their native states from the knowledge of the various environments present in a typical structure. In our lattice model, the number of possible environments is 8, and because there are 20 types of amino acids, the total number of parameters that enter into Equation 3 is 160. The pair probabilities that go into Equation 4 can be implemented in a variety of ways. We will use here the two most common ones.

Statistical procedure 1 (SP1)

Following the approach of Sippl (1990), we discriminated among the pairs of amino acids on the basis of the identity of the residues in the pair, the sequence separation between the members of the pair and the order of the residues in the pair. This led to a total of 5,200 parameters that we extracted from a set of sequences and their corresponding native conformations. We then determined the most probable structure for a sequence based on the probability from Equation 1. The parameters we used in Equation 1 were these 5,200, and the 160 parameters from above that refer to the probability of finding an amino acid in a certain environment.

Statistical procedure 2 (SP2)

Following the approach of Miyazawa and Jernigan (1985), we discriminated among the pairs of amino acids only on the basis of the identity of the residues in the pair. This approach, which is a coarse grained version of SP1, led to a total of 210 parameters that we extracted from a set of sequences and their corresponding native conformations. We then determined the most probable structure for a sequence based on the quantity in Equation 1 for which we used these parameters and the 160 parameters from above that refer to the probability of finding an amino acid in a certain environment.

Optimization procedure (OP)

As mentioned before, the scoring function given by the second type of method is based on the 210 interaction energies between the various types of amino acids present in proteins. To extract these parameters, we used an optimization procedure. For each sequence in the data bank, we built a set of linear inequalities involving the unknown interaction potentials. For the Hamiltonian in Equation 5, such an inequality has the form

$$H(S,\Gamma_{\text{native}}) < H(S,\Gamma_d)$$

or

$$\sum_{i\le j}(w_{ij}^{(d)} - w_{ij}^{(\text{native})})e_{ij} > 0 \qquad (6)$$

where $H(S,\Gamma_{\text{native}})$ is the energy of the sequence S in its native state and $H(S,\Gamma_d)$ is the energy of the sequence in a decoy conformation Γ_d. We obtained the interaction energies between the amino acids by enforcing these types of inequalities for a set of sequences with known native conformations and competing structures and by chosing the e_{ij}s to maximize the smallest of these inequalities using a perceptron procedure (Krauth & Mezard, 1987). For each sequence in the training database, we considered as competing structures all conformations present in the database with the exception of its own native conformation.

When one attempts to extract statistical information from real proteins, one severe constraint is imposed by the relatively small number of protein sequences with known native conformations. To assess the extent of this constraint on the accuracy of statistical parameters, we extracted the parameters for SP1, SP2, and OP using data banks of different sizes. More precisely, we used as training sets the data banks made of the first 200, 250, or 350 sequences (with their corresponding native states) out of the total of 748. The remaining sequences and conformations in each case were used to estimate the accuracy of the extracted parameters.

The details of all these procedures may be found in Methods.

Results

The results of the folding of the 748 sequences present in our data bank, using each of the three procedures presented above, are summarized in Table 1. One of the main assumptions of any statistical approach is that there exists a reasonably good correlation between the native state of a sequence and the structure with the best statistical score. One would then be able to extract parameters from a set of sequences and then to use those parameters to predict the behavior of other sequences The message of the results summarized in Table 1 is that this assumption is not true in general.

Another hope underlying a statistical analysis is that even if the native-like conformation is not the one corresponding to the best statistical score, it can nevertheless be found among the top scoring structures. We investigated this by looking at the rank of the native state energy as given by the order of the scores of a sequence on decoy structures, and we found this to be only marginally correct. The training database that we used for each of the above procedures was the one containing the first 350 sequences from the total of 748, but the results were substantially similar for other training sets. The histograms of the ranks of the native state energy for the 398 sequences that were not part of the training set and whose native states were not the best statistical scorers are shown in Figures 1–3.

One possible idea for improving the performance of these statistical scores is to increase the total number of parameters that go into Equation 2. Unfortunately, as shown by the results from Table 1 and from the three histograms, this is not a real option. The striking fact that emerges from these results is that the increase in the number of statistical parameters (on using SP1 instead of SP2) is able to improve the prediction for the sequences in the training set but only at the cost of a clear decrease in the performance on the remaining sequences. The reason why this happens is simple. The parameters from SP1, which contain information about the sequence separation between the two residues in contact and about their order in the pair, obviously encode a lot more of the features of the sequences and conformations from the training set (they are essentially tailor-made for this set) than the parameters from SP2

Table 1. *The results of the folding of all 748 sequences in the data bank using all 103,346 maximally compact structures as decoy conformations for each sequence*[a]

	SP1	SP2	OP
Training set = 200 sequences			
Fold 200 seq.	59%	55%	91%
Fold 548 seq.	5%	18%	80%
Training set = 250 sequences			
Fold 250 seq.	52%	36%	92%
Fold 498 seq.	5%	24%	85%
Training set = 350 sequences			
Fold 350 seq.	44%	33%	92%
Fold 398 seq.	8%	26%	90%

[a] The parameters used in each of the three procedures mentioned in the text (SP1, SP2, and OP) were extracted from the sets comprised of the first 200, 250, and 350 sequences, respectively, and conformations (out of the total of 748). The numbers represent the percent of sequences for which the native state corresponds to the best score. The first row shows the results of a test of folding of the sequences in the training set among all maximally compact conformations.

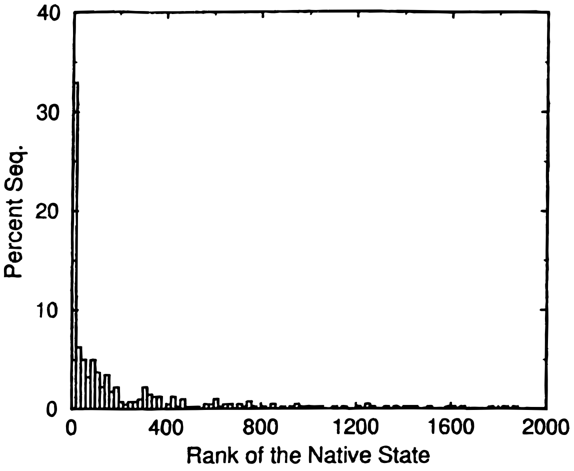

Fig. 1. Histogram of the rank of the native state energy for SP1. The corresponding parameters were extracted from the set of the first 350 sequences and conformations (out of the total of 748). The sequences whose ranks appear here are from the set of 398 sequences, excluding those in the training set, whose native conformations did not correspond to the best score. The *y*-axis shows the percent of sequences having a given rank (2 and above) and each bin on the *x*-axis has a width of 20.

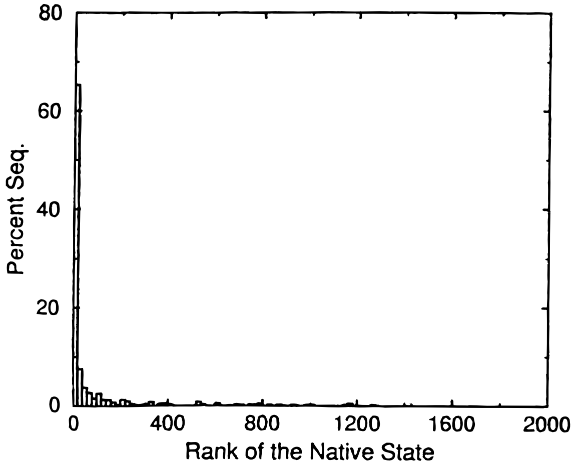

Fig. 2. Same as in Figure 1, except for the SP2 method.

and, therefore, are able to reproduce these characteristics quite well during folding. Now, when these same parameters are tested on another set of sequences, which in general are very different from those in the training set, it is to be expected that the SP1 approach will only have a modest success in folding them exactly, because the characteristics of these sequences could be somewhat different from those of the sequences in the training set. This conclusion is in agreement with the result of Mirny and Shakhnovich (1996) regarding the over-determination or the under-determination of the problem of finding the parameters.

In the above analysis, we used all 103,346 available conformations as target structures. The study of Li et al. (1996) pointed out that only a small fraction of these conformations is designable (i.e., is able to house a large number of sequences). This is exactly what also happens in the case of real proteins where the general belief is that the number of three-dimensional conformations that can house all the protein-like sequences is much smaller than the total number of sequences. Many sequences fold into the same conformation, and many of the other structures are not viable (they do not house any protein-like sequence). Therefore, it is worth studying the effect that a reduction in the number of target conformations (based on their capacity of housing sequences) would have on the performance of the statistical parameters. To test this, we extracted the parameters of SP1, SP2, and OP using only the database of the first 350 sequences and the corresponding structures. Next, we generated a set of 107 sequences that were different from the 350 in the training set, but that had their unique native state conformations among the same 350 structures. We then folded all 457 sequences, with each of the three procedures, using as decoy conformations for each sequence only the 350 structures from the training data bank. This was done because these 350 structures are surely designable ones. The results are presented in Table 2 and are consistent with the data in Table 1 and Figures 1–3. Clearly, there is the expected improvement in the folding performance of the statistical parameters, but even the best success rate of 76% is far from the nearly perfect success obtained when using the parameters derived from the optimization procedure. This underscores the result that even the most positive scenario for statistical scores may not lead to complete success in predicting native-like folds of sequences of amino acids.

As noted before, Rooman and Wodak (1995) have shown that the statistical scores might be useful not only for folding, but also for design. We now proceed to an evaluation of the ability of the statistical approach to predict a sequence or a set of sequences that can fold rapidly and reproducibly into a given native target conformation. For this, we first extracted the parameters of SP2 using only the database of the first 350 sequences and structures. Next, to build a data bank of highly designable conformations, we selected 24 sequences, characterized by large energy gaps (i.e., larger than 2.0), and their native conformations from the remaining 398 sequences and structures. The goal was to design, for each of these 24 conformations, sequences that could fold into them. If, indeed, the score is an indicator of the fit between a sequence and a structure, the idea was that a good sequence for design could be found among those having very low scores on the target conformation. Here, an additional wrinkle, compared to the case of folding, is that $P(\text{sequence})$ from Equation 1, which is a measure of the free energy of the sequence, can no longer be disregarded, because we are dealing with different sequences. To keep $P(\text{sequence})$ approximately constant, we exploited the observation of Micheletti et al. (1998) that the zero temperature free energies of sequences

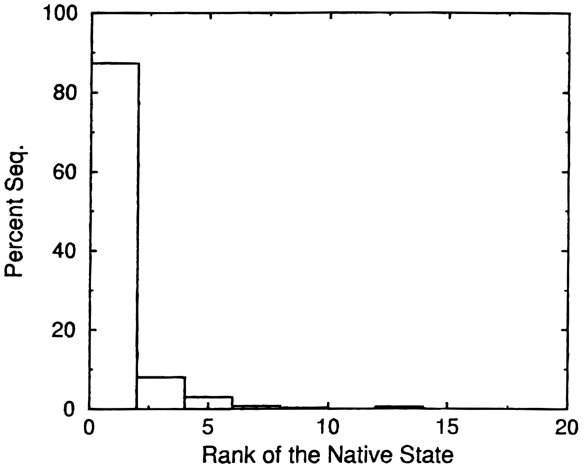

Fig. 3. Same as in Figure 1, except for the OP method. Note that the bin width on the *x*-axis is now 2.

Table 2. *The results of the folding of the 457 sequences using just the first 350 conformations (out of the total of 748 present in the data bank) as decoy conformations for each sequence*[a]

	SP1	SP2	OP
Training set = 350 sequences			
Fold 350 seq.	91%	86%	100%
Fold 107 seq.	62%	76%	97%

[a] The parameters used in each of the three procedures (SP1, SP2, and OP) were extracted from the set comprised of the first 350 sequences and conformations (out of the total of 748). The numbers represent the percent of sequences for which the native state corresponds to the best score. The first row shows the results of a test of folding of the sequences in the training set among all maximally compact conformations.

with the same composition are substantially similar. For each target conformation, the amino acid composition of the trial sequences generated was held fixed to that of the original sequence. As a result, for each of the 24 conformations, we applied the following procedure: we started with the original sequence, we evaluated its score on the target conformation; then, along the lines of Shakhnovich et al. (1996) (whose design strategy is the energy-based analog of the procedure we were using), we performed a Metropolis Monte Carlo optimization procedure at $T = 0$ in the space of sequences with the same sublattice composition as the original one; i.e., at each step we generated a new sequence by exchanging the residues between two positions in the old sequence that were chosen at random (but from the same sublattice), and we calculated the score of the new sequence on the target conformation. The new sequence was accepted only if its score was lower than that of the previous sequence. The procedure stopped when there was no improvement in the score over the last 10^6 steps. Then we checked, using the true interaction potentials and all 103,346 available conformations as the decoy set, whether the target conformation was indeed the native state of the sequence corresponding to the lowest score. This test was successful in 23 of the 24 cases but, when we considered the sequences corresponding to the two lowest scores, the success of the design procedure was complete. It is very likely that our procedure leads to a local minimum of the score. To obtain more low-lying minima, we also performed a simulated annealing procedure for each of the mentioned 24 conformations. We started from a relatively high value of the fictitious optimization temperature, and then we reduced the temperature exponentially ($T_{i+1} = 0.95T_i$). We ran the program 20,000 steps at each temperature, and we stopped the procedure when it could no longer improve the score over the last 10^5 steps. As expected, this procedure led to lower final scores than before on each conformation. We again obtained very good success in the design procedure: for 23 of the 24 conformations the sequence corresponding to the lowest selected score had its ground state in the target conformation (surprisingly, the one conformation for which the procedure failed was different from the previous failure with the local minima search), proving that, indeed, the statistical approach might be quite valuable for design.

Methods

To extract the parameters in Equation 3, we needed to define structural environments on the three-dimensional lattice. We started from the fact that in the cubic lattice there are four distinct positions in which one finds a residue: in the center of the $3 \times 3 \times 3$ cube, in the middle of a face, in the middle of an edge, or in a corner. Each of these four positions is characterized by a different number of nearest neighbors. This translates, for each type of amino acid, into eight possible environments [in the spirit of the Bowie et al. (1991) profile method] resulting in a total of 160 environment parameters for Equation 3: (1) the amino acid is at the first position in the chain and has four neighbors; (2) the amino acid is in the middle of the chain and has four neighbors; (3) the amino acid is at the last position in the chain and has four neighbors; (4) the amino acid is in the middle of the chain and has three neighbors; (5) the amino acid is at the first position in the chain and has two neighbors; (6) the amino acid is in the middle of the chain and has two neighbors; (7) the amino acid is at the last position in the chain and has two neighbors; and (8) the amino acid is in the middle of the chain and has one neighbor. Starting from this classification, we then looked at sequences in their native state conformations and we determined how many times each type of amino acid appears in each of the above environments. The probability of finding a residue aa_i in an environment E_j, $P(aa_i|E_j)$, is given by

$$P(aa_i|E_j) = \frac{n(aa_i : E_j)}{\sum_{k=1}^{8} n(aa_i : E_k)} \tag{7}$$

where $n(aa_i : E_j)$ is the number of times the amino acid type i is found in environment E_j.

For the parameters in Equation 4, one needs to account for amino acid pairs in contact and their corresponding sequence separations. We considered each case separately. For SP1, besides the identity of the amino acids in the pair, we took into account the sequence separation between the residues in the pair and we considered the pairs (aa_i, aa_j) and (aa_j, aa_i) to be nonequivalent. We then looked at a data bank of sequences with their native state conformations and we counted the number of pairs (aa_i, aa_j) at each sequence separation k, $n(aa_i, aa_j : k)$. From these, the probability of finding a pair (aa_i, aa_j) at a given sequence separation k, $P(aa_i, aa_j : k)$, is given by

$$P(aa_i, aa_j : k) = \frac{n(aa_i, aa_j : k)}{\sum_{aa_i=1}^{20} \sum_{aa_j=1}^{20} n(aa_i, aa_j : k)}. \tag{8}$$

In this model, for each sequence of 27 beads, there are 13 possible sequence separations between residues in contact (all odd numbers between 1 and 25), and there are 400 pairs of amino acids so the total number of parameters for Equation 4 is 5,200. This very large number suggests, as pointed out by Sippl (1990), that we might be faced with the problem of small data sets when attempting to extract these parameters. To eliminate this bias, we performed a correction for small data sets along the line of Sippl (1990). For each pair of amino acids (aa_i, aa_j), we evaluated the number of times it appears in the database of sequences and conformations independent of the sequence separation between the two residues, $m(aa_i, aa_j)$, and we modified Equation 7 to

$$P(aa_i, aa_j : k) = \frac{1}{1 + m(aa_i, aa_j)\sigma} + \frac{m(aa_i, aa_j)\sigma}{1 + m(aa_i, aa_j)\sigma} \frac{400n(aa_i, aa_j : k)}{\sum_{aa_i} \sum_{aa_j} n(aa_i, aa_j : k)} \tag{9}$$

where σ is a small number that weights the contribution of each measurement of a pair (aa_i, aa_j) to the above probability. In all our calculations σ was set to $1/20$ meaning that if a pair is seen 20 times, the two terms in Equation 8 have equal weight. For SP2, we disregarded the sequence separation between the residues in a pair and also the order of the amino acids in the pair, keeping only the distinction between pairs. The direct implication of these considerations is a drastic reduction in the total number of parameters in Equation 4, from 5,200 to 210. The probability of having a pair (aa_i, aa_j) is determined by normalizing the number of (aa_i, aa_j)

pairs found in a data bank of sequences and structures, $n(aa_i, aa_j)$, to the total number of amino acid pairs in the data bank:

$$P(aa_i, aa_j) = \frac{n(aa_i, aa_j)}{\sum_{aa_i=1}^{20} \sum_{aa_j=aa_i}^{20} n(aa_i, aa_j)}. \tag{10}$$

Conclusions

We have presented a lattice study of the performance of two of the most commonly used methods for evaluating the goodness of the fit between a sequence and a structure. We find that the main assumption of the method, based on statistical considerations, that the most statistically probable structure for a given sequence is likely to be its native conformation is not firmly grounded. All our attempts to improve the performance of this kind of method were unable to lead to complete success in folding a set of sequences. Another possible shortcoming of the statistical approach comes from the fact that it assumes, in the way its scoring function is built, that sequences that are highly similar should fold into related structures, and this assumption has increasingly been called into question (Dalal et al., 1997; Cordes et al., 1999). Even so, we were able to show that this approach leads to very good success in design studies. This is pleasing because the probability from Equation 2 is a direct measure of the goodness of the fit between a sequence and a structure.

The other approach, based on the knowledge of the effective energy parameters of interaction between amino acids, showed very good success in folding various sequences. The problem here, in the case of real proteins, is the difficulty in finding viable alternative conformations that compete significantly with the native one in housing the sequences in the training set—a requirement for determining the optimal parameters.

The model that we have used and the procedure, due to their simplicity, may be called into question. One possible objection is that our approach is not completely unbiased because, in the optimization procedure, we use the same functional form for the Hamiltonian as the true one. As a result, it may not be entirely unexpected that the corresponding extracted energy parameters perform well in the recognition of the native state of the various sequences. A subject worthy of study is the degree to which a pairwise contact Hamiltonian is able to approximate more complex functional forms of the actual interaction potential. Another possible objection is that the lattice model, which we have used, considers only maximally compact conformations of a short chain and may not be a good approximation of real conformations. It is true that, if a method works on a model, it does not necessarily follow that it is going to be successful when applied to the case of real proteins. Still, it is common sense that if the method fails on the model, its chances to be successful on real proteins cannot be too good. In addition, our very good success in designing sequences on the model, using the statistical approach, corroborates well with the results reported by Bowie et al. (1991), which gives us confidence that the results of our study are not very likely to be just artifacts of the model.

Acknowledgments

We thank Cristian Micheletti, Flavio Seno, and Gianni Settanni for helpful discussions. This work was supported by INFN, INFM, NASA, NATO, The Petroleum Research Fund administered by the American Chemical Society, and by the Center for Academic Computing at Penn State.

References

Bernstein FC, Koetzle TF, Williams GJB, Meyer EE, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 122*:535–542.

Betancourt M, Thirumalai D. 1999. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci 8*:361–369.

Bowie JU, Luthy R, Eisenberg DA. 1991. Method to identify protein sequences that fold into a known three-dimensional structure. *Science 253*:164–170.

Bryant SH, Lawrence CE. 1991. An empirical energy function for threading protein sequence through the folding motif. *Proteins 16*:92–112.

Chan HS, Dill KA. 1993. The protein folding problem. *Phys Today Feb.*:24–32.

Chiu TL, Goldstein RA. 1998. Optimizing energy potentials for success in protein tertiary structure prediction. *Fold Des 3*:223–228.

Clementi C, Maritan A, Banavar JR. 1998. Folding, design, and determination of interaction potentials using off-lattice dynamics of model heteropolymers. *Phys Rev Lett 81*:3287–3290.

Cordes MHJ, Walsh NP, McKnight CJ, Sauer RT. 1999. Evolution of a protein fold in vitro. *Science 284*:325–327.

Creighton TE. 1993. *Proteins: Structures and molecular properties*. New York: W.H. Freeman and Company.

Dalal S, Balasubramanian S, Regan L. 1997. Protein alchemy: Changing beta-sheet into alpha-helix. *Nat Struct Biol 4*:548–550.

Goldstein R, Luthey-Schulten ZA, Wolynes PG. 1992a. Optimal protein-folding codes from spin-glass theory. *Proc Natl Acad Sci USA 89*:4918–4922.

Goldstein R, Luthey-Schulten ZA, Wolynes PG. 1992b. Protein tertiary structure recognition using optimized hamiltonians with local interactions. *Proc Natl Acad Sci USA 89*:9029–9033.

Hao MH, Scheraga HA. 1996a. How optimization of potential functions affects protein folding. *Proc Natl Acad Sci USA 93*:4984–4989.

Hao MH, Scheraga HA. 1996b. Optimizing potential functions for protein folding. *J Phys Chem 100*:14540–14548.

Hao MH, Scheraga HA. 1999. Designing potential energy functions for protein folding. *Curr Opin Struct Biol 9*:184–188.

Hendlich M, Lackner S, Witckus H, Floechner H, Froschauer R, Gottsbachner K, Casari G, Sippl MJ. 1990. Identification of native protein folds amongst a large number of incorrect models; The calculation of low energy conformations from potentials of mean force. *J Mol Biol 216*:167–180.

Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. *Nature 358*:86–89.

Kolinski A, Godzik A, Skolnick J. 1993. A general method for the prediction of the three-dimensional structure and folding pathway of globular proteins: Application to designed helical proteins. *J Chem Phys 98*:7400–7433.

Kolinski A, Skolnick J. 1994. Monte-Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins 18*:338–352.

Koretke KK, Luthey-Schulten Z, Wolynes PG. 1996. Self-consistently optimized statistical mechanical energy functions for sequence structure alignment. *Protein Sci 5*:1043–1059.

Koretke KK, Luthey-Schulten Z, Wolynes PG. 1998. Self-consistently optimized energy functions for protein structure prediction by molecular dynamics. *Proc Natl Acad Sci USA 95*:2932–2937.

Krauth W, Mezard M. 1987. Learning algorithms with optimal stability in neural networks. *J Phys A 20*:L745–L752.

Lau KF, Dill KA. 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules 22*:3986–3997.

Li H, Helling R, Tang C, Wingreen N. 1996. Emergence of preferred structures in a simple model of protein folding. *Science 273*:666–669.

Luthy R, Bowie JU, Eisenberg D. 1992. Assessment of protein models with three-dimensional profiles. *Nature 356*:83–85.

Maiorov VN, Crippen GM. 1992. Contact potential that recognizes the correct folding of globular proteins. *J Mol Biol 227*:876–888.

Micheletti C, Seno F, Maritan A, Banavar JR. 1998. Protein design in a lattice model of hydrophobic and polar amino acids. *Phys Rev Lett 80*:2237–2240.

Mirny LA, Shakhnovich EI. 1996. How to derive a protein folding potential? A new approach to an old problem. *J Mol Biol 264*:1164–1179.

Miyazawa S, Jernigan RL. 1985. Estimation of effective inter-residue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules 18*:534–552.

Miyazawa S, Jernigan RL. 1996. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol 256*:623–644.

Miyazawa S, Jernigan RL. 1999. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins 34*:49–68.

Park B, Levitt M. 1996. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. *J Mol Biol 258*:367–392.

Pellegrini M, Doniach S. 1993. Computer simulation of antibody binding specificity. *Proteins 15*:436–444.

Rooman MJ, Wodak SJ. 1995. Are database-derived potentials valid for scoring both forward and inverted protein folding. *Protein Eng 8*:849–858.

Seno F, Maritan A, Banavar JR. 1998. Interaction potentials for protein folding. *Proteins 30*:244–248.

Shakhnovich EI, Abkevich V, Ptitsyn O. 1996. Conserved residues and the mechanism of protein folding. *Nature 379*:96–98.

Simons KT, Kooperberg C, Huang E, Baker D. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol 268*:209–225.

Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. 1999. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins 34*:82–95.

Sippl MJ. 1990. Calculation of conformational ensembles from potentials of mean force. *J Mol Biol 213*:859–883.

Sippl MJ, Weitckus S. 1992. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations. *Proteins 13*:258–271.

Skolnick J, Jaroszewski L, Kolinski A, Godzik A. 1997. Derivation and testing of pair potentials for protein folding. When is the quasi–chemical approximation correct? *Protein Sci 6*:676–688.

Skolnick J, Kolinski A. 1990. Simulations of folding of globular proteins. *Science 250*:1121–1125.

Sun S. 1993. Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. *Science 2*:762–785.

Tanaka S, Scheraga HA. 1976. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules 9*:945–950.

Thomas DP, Dill KA. 1996. An iterative method for extracting energy-like quantities from protein structures. *J Mol Biol 257*:457–469.

Vendruscolo M, Domany E. 1998. Pairwise contact potentials are unsuitable for protein folding. *J Chem Phys 109*:11101–11108.

von Mourik J, Clementi C, Maritan A, Seno F, Banavar JR. 1999. Determination of interaction potentials of amino acids from native protein structures: Test on simple lattice models. *J Chem Phys 110*:10123–10133.

Wilmanns M, Eisenberg D. 1993. 3-Dimensional profiles from residue-pair preferences—Identification of sequences with beta/alpha-barrel fold. *Proc Natl Acad Sci USA 90*:1379–1383.

Wilson C, Doniach S. 1989. A computer model to dynamically simulate protein folding: Studies with crambin. *Proteins 6*:193–209.

Zhang C. 1998. Extracting contact energies from protein structures: A study using a simplified model. *Proteins 3*:299–308.

Zhang C, Vasmatzis G, Cornette JL, DeLisi C. 1997. Determination of atomic desolvation energies from structures of crystallized proteins. *J Mol Biol 267*:707–726.

Zhang L, Skolnick J. 1998. How do potentials derived from structural databases relate to "true" potentials? *Protein Sci 7*:112–122.