PROTEINS: Structure, Function, and Genetics 44:79–96 (2001)

How to Guarantee Optimal Stability for Most Representative Structures in the Protein Data Bank

Ugo Bastolla,[1,3] Jochen Farwer,[1] Ernst Walter Knapp,[1] and Michele Vendruscolo[2]
[1]*Free University of Berlin, Department of Biology, Chemistry and Pharmacy, Institute of Chemistry, Berlin, Germany*
[2]*Oxford Centre for Molecular Sciences, New Chemistry Laboratory, Oxford, United Kingdom*
[3]*Max-Planck-Institut for Colloids and Interfaces, Potsdam, Germany*

ABSTRACT **We proposed recently an optimization method to derive energy parameters for simplified models of protein folding. The method is based on the maximization of the thermodynamic average of the overlap between protein native structures and a Boltzmann ensemble of alternative structures. Such a condition enforces protein models whose ground states are most similar to the corresponding native states. We present here an extensive testing of the method for a simple residue-residue contact energy function and for alternative structures generated by threading. The optimized energy function guarantees high stability and a well-correlated energy landscape to most representative structures in the PDB database. Failures in the recognition of the native structure can be attributed to the neglect of interactions between different chains in oligomeric proteins or with cofactors. When these are taken into account, only very few X-ray structures are not recognized. Most of them are short inhibitors or fragments and one is a structure that presents serious inconsistencies. Finally, we discuss the reasons that make NMR structures more difficult to recognize. Proteins 2001;44:79–96.**

Key words: protein folding; contact maps; contact energy; threading

INTRODUCTION

Several methods to derive empirical energy functions for protein folding have been presented recently (for a review see Hao and Scheraga.[1] Such energy functions are a necessary ingredient in simplified models of proteins. They are referred to as *knowledge-based* potentials, since, instead of being derived from first principles, they are obtained from the ever-growing information contained in the Protein Data Bank (PDB).

Two main classes of methods are presently available to derive energy parameters: quasi-chemical methods, which are based on the frequency of structural motifs in the database of proteins[2–5] and optimization methods, the aim of which is to build models such that the experimentally known native structures have maximal stability among an ensemble of alternative structures.[6–16]

In Bastolla et al.[17] we proposed the overlap method, an optimization procedure based on the maximization of the average overlap between the native structure and the ensemble of alternative conformations. As for other optimization methods, a set of proteins (the "training set") whose native structures are known is required to tune the energy parameters. The optimized energy function is then applied to a second, independent set of proteins (the "test set"). We use the contact map representation of protein structures and a simple form of the energy function, based on pairwise contact interactions between residues. An important feature of the present method is that it can be applied to more complex protein models and energy functions. The contact pairwise energy function that we obtain ensures both high thermodynamic stability and correlated energy landscapes for the native structures of most proteins in a test set of about 1,200 proteins, which comprises most of the representative structures in the PDB.

An important aspect of any optimization method of energy functions for protein folding is the generation of an ensemble of alternative conformations. We generate them by gapless threading using protein structures from the PDB. Alternative conformations generated in this way are usually considered an easy test for energy functions, because they are few and most of them are not suitable to represent low-energy conformations of a given protein sequence for reasonable energy functions. However, it has been shown recently[9] that when a large database of proteins is used, there is no energy function of the simple form that we use that can ensure thermodynamic stability to the native state of all proteins simultaneously. In accordance with these results, in the present study we found that several native states are not recognized, i.e., there are alternative conformations with energy lower than the native one. In nearly all cases where recognition of the native state fails, we are able to identify the physical reason for it in the fact that we neglected some interactions: either interactions with neighboring chains in the case of oligomeric proteins or interactions with cofactors. The importance of such interactions in the test of energy functions has already been stressed by Maiorov and Crippen.[7] The neglected interaction energy can be calculated

Grant sponsor: Deutsche Forschungsgemeinschaft; Grant number: SFB 498, Project A5, Grant sponsor: EMBO.

*Correspondence to: Ugo Bastolla, Centro de Astrobiología (INTA), Ctra. de Ajalvir km.4, 28850 Torrejón de Ardoz (Madrid) Spain. E-mail: ugo@chemie.fu-berlin.de

Received 13 June 2000; Accepted 30 March 2001

for most oligomeric proteins. Interestingly, it turns out that it is well correlated to the difference in energy between the lowest energy decoy and the native state. When it is added to the native energy, the native state becomes stable. Therefore the use of decoys generated by gapless threading indicates when some important interactions are neglected, and it even allows for a rough estimate of the amount of neglected energy. Gapped threading is a more powerful method, since, by allowing gaps, it can produce a number of additional structures exponentially large in the length of the sequence. However, it is computationally more demanding and one has still to solve the problem of finding a good scoring function to compare structures of different length. On the other hand, gapless threading, despite its simplicity, is already sufficient to make the optimization problematic and to identify instances where neglected interactions make the recognition impossible. These considerations justify the use in the present study of gapless instead of gapped threading to generate alternative structures.

From the set of more than 1,000 X-ray structures that we considered, the lack of recognition remains unexplained only for nine proteins. Three of them have cofactors covalently bound that could explain the energy difference between the native state and the lowest energy state of the model. Another five are fragments or small inhibitors, for which we suggest that interactions with neighboring crystal cells can play a role for stability.[18] The most problematic structure, interferon γ (PDB code 1rfbA), when checked with the program WHAT-CHECK (see Hooft et al.[37]) reveals several structural inconsistencies (see below).

For NMR structures, we found that better results are obtained when an energy function is optimized separately: we suggest that the small physical differences between protein structures in a crystal and in solution do not allow us to use reliably an energy function valid for crystal structures to recognize NMR structures. Moreover, NMR structures are often peculiar, which makes their simultaneous stabilization with more typical proteins more difficult.

The energy function derived in this work generates well-correlated energy landscapes for most proteins in the training set. This is a prerequisite for structure prediction. Gapless threading, however, is not an adequate method to generate candidate structures similar to the real native one. If structures above a given threshold of similarity with the native one are present in the database, the most similar structure is recognized by the present energy function, but, if the similarity is too low, the predicted structure is almost unrelated to the native one.

We first define energy functions for coarse-grained models of proteins and describe the simplified protein model and the approximation on the energy function that we adopt. The structural overlap, on which the present method is based, and the stability parameter characterizing correlations in the energy landscape, are then defined. Next, the overlap method is introduced and compared to other existing optimization methods, the inequality method, and the Z score methods. We first test the overlap method on a lattice model. Then, we present results obtained for more than 1,000 protein structures, starting from a preliminary study of the number of contacts per residue, the overlap and its relation to the root mean square deviation (RMSD). Derivation and test of energy parameters are then described, distinguishing between monomeric and oligomeric proteins, and between structures determined by X-ray crystallography and structures determined by NMR spectroscopy. We discuss the reasons why in specific cases we fail to stabilize protein fragments and NMR structures. Before concluding the paper with an overall discussion, we address the performance of the present energy function for structure prediction.

MATERIALS AND METHODS

Coarse-Grained Protein Model

We denote by Γ the microscopic configuration of a protein, specified by the coordinates of all the atoms of the protein and of the solvent, and by $\mathbf{C} = f(\Gamma)$ its simplified version, as obtained from the coarse-graining function f. In this article, we adopt the contact map representation of the structure Γ, but the method described here can be generalized to other reduced representations.

The contact map of a structure Γ of N amino acids is the $N \times N$ matrix $\mathbf{C}$ with elements:

$$C_{ij} = \begin{cases} 1 & \text{if residues } i \text{ and } j \text{ are in contact,} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

We found the best results using the following definition: two residues are defined to be in contact if any pair of their heavy atoms is closer than the threshold distance $R_c = 4.5$ Å.[9] Contacts among residues separated by less than three positions along the sequence are not considered.

Given a particular structure, it is always possible to construct a unique contact map, but such correspondence is not invertible: given a contact map, there are always several structures corresponding to it. The entropy of a contact map measures the corresponding number of compatible structures, and it can be formally computed as

$$S(\mathbf{C}) = k_B \ln\left(\int d\Gamma \chi(\mathbf{C}, \Gamma) \right), \quad (2)$$

where $\chi(\mathbf{C}, \Gamma)$ is one if $\mathbf{C} = f(\Gamma)$ and zero otherwise, and Γ is a complete configuration of the system.

We denote the amino acid sequence of the protein by $\mathbf{S} = \{S_1, \cdots, S_N\}$. Assuming that the protein is in thermodynamic equilibrium at temperature T, we can define the effective energy of a chain with sequence $\mathbf{S}$ and contact map $\mathbf{C}$. This quantity depends on temperature, and it is given by

$$E_T(\mathbf{C}, \mathbf{S}) = -k_B T \ln\left(\frac{\int d\Gamma \chi(\mathbf{C}, \Gamma) e^{-\varepsilon(\Gamma, \mathbf{S})/k_B T}}{\int d\Gamma \chi(\mathbf{C}, \Gamma)} \right), \quad (3)$$

where $\varepsilon(\Gamma, \mathbf{S})$ is the microscopic energy of the sequence $\mathbf{S}$ in configuration Γ. The free energy of the contact map $\mathbf{C}$ is thus given by $G_T(\mathbf{C}, \mathbf{S}) = E_T(\mathbf{C}, \mathbf{S}) - TS(\mathbf{C})$.

Since it is unfeasible to compute this expression, one resorts to approximations. Here we assume that the effective energy function $E_T(\mathbf{C}, \mathbf{S})$ has the simple form of sum of pairwise contact contributions:

$$E(\mathbf{C}, \mathbf{S}, \mathbf{U}) = \sum_{ij} C_{ij} U(S_i, S_j), \tag{4}$$

where $U(a, b)$ is a 20×20 symmetric matrix, thus containing 210 interaction parameters that represent the free energy gain when amino acids of types a and b are brought in contact at a given temperature T.* The aim of this work is to determine these energy parameters and to test the corresponding energy function. It should be remembered, however, that effective interactions depend on the given thermodynamic conditions, and that they express a free energy rather than an energy.

An important aspect of every optimization method is the choice of the set of structures to be compared to the native. Let's denote by $\Omega(\mathbf{S})$ the set of conformations available to sequence $\mathbf{S}$. There are two possible strategies to generate such a set: one can either run off-lattice simulations (Monte Carlo or Molecular Dynamics) or retrieve protein structures from a database. The former strategy would be most suitable for structure prediction. However, it is very difficult to apply, both because of the computational limitation and because it is very difficult to impose via an effective energy function appropriate secondary and tertiary structures, as well as dihedral angles, bond angles, and bond lengths. These problems are less relevant when structures extracted from the PDB are used, because these structures are determined directly from experimental data. These decoys are, however, of poor quality because most of them tend to have high energies, and, therefore, negligible weights in the thermodynamic ensemble, for reasonable choices of the interaction parameters.

The easiest way to generate structures from the PDB is by gapless threading[20]: one considers a sequence $\mathbf{S}$ of length N and a structure Γ' of length $N' > N$ and takes from it all possible substructures Γ corresponding to N consecutive residues. In general, these substructures correspond to subsequences different from $\mathbf{S}$. Thus, steric constraints might be violated, because the original sequence is replaced and the side-chains are consequently changed. This has mostly the effect to lower the effective energy with respect to the "true" one, and thus it should make recognition of the native structure more difficult. We perform gapless threading in contact map space: we first compute the contact map of the structure Γ' and then we extract from it all possible submaps corresponding to N consecutive residues.

*In the following, we use units of energy such that $k_B T = 1$. Thus by $\mathbf{U} = U(a, b)$ we actually mean the dimensionless quantities $U(a, b)/k_B T$.

Overlap and Energy Landscapes

The basic quantity in the present method is the overlap $q(\mathbf{C}, \mathbf{C}')$, which measures the degree of pairwise similarity between the two contact maps $\mathbf{C}$ and $\mathbf{C}'$. In this paper, $q(\mathbf{C}, \mathbf{C}')$ is defined as

$$q(\mathbf{C}, \mathbf{C}') = \frac{\sum_{ij} C_{ij} C'_{ij}}{\max\left(\sum_{ij} C_{ij}, \sum_{ij} C'_{ij}\right)}. \tag{5}$$

In other words, the overlap between two contact maps is equal to the number of contacts that they have in common, divided by the maximum number of contacts of the two maps. With this normalization, $0 \leq q(C, C') \leq 1$ and $q(C, C') = 1$ if and only if the two contact maps are equal.

Since the energy function that we use is based only on contact maps, all conformations with $q = 1$ are degenerate in energy, and conformations with $q \approx 1$ have a very similar energy. Let us suppose that the effective energy function $E_T(\mathbf{C}, \mathbf{S})$ is known, and that, at appropriate temperature, the contact map of lowest effective energy, $\mathbf{C}_0$, coincides with the contact map of the native state, $\mathbf{C}_{\mathrm{nat}}$. Contact maps with large overlap with $\mathbf{C}_0$ have on average low energy, but their number is much smaller than that of contact maps with small overlap. Two scenarios are then possible. In the first one, all structures of low energy are similar to the ground state. In this case, we say that the energy landscape is well correlated. In the other scenario, typical for random heteropolymers, structures of low energy might also be found among those with small overlap. This might, for example, be the case of large proteins, where folding must be supported by chaperones.[23] Although this second scenario is the one to be avoided in constructing energy functions for protein fold prediction, it is also at present the most common one (see, e.g., Fig. 1 in ref. 21 and Fig. 4 in ref. 9).

The correlation of the energy landscape, embodied in the funnel image proposed by Bryngelson and Wolynes,[22] is a key issue of protein folding, because it is related to three important problems: the stability of the native state, its kinetic accessibility, and the possibility to perform structure prediction.

Concerning thermodynamic stability, if the alternative low-energy states have small overlap with the ground state, then the ground state dominates the Boltzmann ensemble only at very low temperatures and configurational fluctuations can be very large. If, on the other side, the energy landscape is well correlated, a cooperative folding transition can take place at much higher temperatures. This consequence of correlated energy landscapes has been observed in several lattice models.[24,25]

Concerning folding kinetics, the correlation in the energy landscape drives the system towards its ground state for any reasonable dynamics, whereas in a polymer with an uncorrelated energy landscape several almost equivalent energy minima are present and their attraction basins are very small; thus, the system has to be very close to the ground state in order to recognize it. All lattice models that

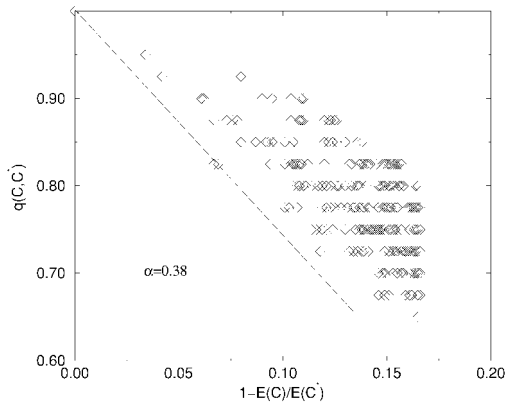

Fig. 1. Plot of the overlap vs. the normalized energy gap.

show good properties with respect to folding dynamics also show correlated energy landscapes.[24,25]

The lesson from lattice model studies is that it is possible to construct well-correlated energy landscapes. Therefore, it is worthwhile to attempt an extension of this result to more realistic models of protein structure, at least for small globular proteins. The realization of this plan would lead to optimism about protein structure prediction. For any approximated model, a structure found by energy minimization will, in general, be different from the true native state. However, if the energy landscape is well correlated, the lowest energy structure will also be most similar to the native state among those generated. On the contrary, for poorly correlated ("rugged") energy landscapes, the lowest energy structure found will, in general, be very different from the native state.

In the present model, if the maximal similarity between alternative structures and the native state is above a given threshold ($q > 0.5$ in the present case, but possibly larger if better algorithms are used to generate alternative structures), with high probability the most similar structure also has the lowest energy. Below such threshold, the lowest energy structure is only very weakly related to the most similar one.

A possible way to characterize the correlations of the energy landscape for a sequence $\mathbf{S}$, is to plot the energy gap between all structures $\mathbf{C}$ and the ground state $\mathbf{C}_0(\mathbf{S})$ as a function of their overlap (see Fig. 1). We define the dimensionless parameter $\alpha(\mathbf{S})$ by requiring that for every structure $\mathbf{C}$ the following inequality is fulfilled:

$$\frac{E(\mathbf{C}, \mathbf{S}) - E(\mathbf{C}_0, \mathbf{S})}{|E(\mathbf{C}_0, \mathbf{S})|} \geq \alpha(\mathbf{S})(1 - q(\mathbf{C}, \mathbf{C}_0)). \tag{6}$$

A similar parameter has been defined by Gutin et al.[26] The larger $\alpha(\mathbf{S})$ is, the larger is the energy gap characterizing structures that are weakly related to the ground state. For poorly correlated energy landscapes, $\alpha(\mathbf{S})$ is almost zero. Note that, in general, the stability parameter $\alpha(\mathbf{S})$ depends on the sequence. This parameter is more specific than the energy gap defined by Dinner et al.[24] and it is also related to the Z score.[20] The difference is that, instead of considering the energy of an average structure, it takes into account the structures with the lowest energy for a given overlap.

Overlap Method

In any meaningful protein model, the ground state of sequence $\mathbf{S}$, indicated as $\mathbf{C}_0(\mathbf{S}, \mathbf{U})$, where $\mathbf{U}$ are the interaction parameters, should coincide with the experimentally observed native state $\mathbf{C}_{\text{nat}}(\mathbf{S})$. Therefore, their overlap q_0, defined as

$$q_0 = q(\mathbf{C}_0(\mathbf{S}, \mathbf{U}), \mathbf{C}_{\text{nat}}(\mathbf{S})), \tag{7}$$

should be equal to 1. This condition is, however, not sufficient to guarantee foldability: folding is viable only for a model with a correlated energy landscape.

In order to optimize energy parameters so that these conditions are met, we introduce the most important quantity of the overlap method, the Boltzmann averaged native overlap $Q(\mathbf{S}, \mathbf{U})$:

$$Q(\mathbf{S}, \mathbf{U}) = \frac{\sum_{\Gamma} q(\mathbf{C}, \mathbf{C}_{\text{nat}}) e^{-E(\mathbf{S},\mathbf{C},\mathbf{U})/k_B T}}{\sum_{\Gamma} e^{-E(\mathbf{S},\mathbf{C},\mathbf{U})/k_B T}}, \tag{8}$$

where $\mathbf{C}$ is the contact map of configuration Γ. If the average overlap $Q(\mathbf{S}, \mathbf{U})$ with the native state is close to one, then either the ground state of S coincides with the native state or it is very close to it. When the temperature T is not very low, a high value of Q implies that structures very different from the native one have high energy; in other words, we can expect that the energy landscape is well correlated.

Following such considerations, we propose to optimize the energy parameters imposing that $Q(\mathbf{S}, \mathbf{U})$ is maximal for a training set of proteins. We maximize the sum of the overlap for the proteins in the training set:

$$\bar{Q}(\mathbf{U}) = \sum_{\mathbf{S}} Q(\mathbf{S}, \mathbf{U}). \tag{9}$$

The optimization is performed at a constant scale of the interaction parameters U, defined through

$$\sum_{a,b} U^2(a, b) = U^2. \tag{10}$$

Multiplying all parameters times a scalar U is similar to considering a temperature rescaled by a factor $1/U$. Without the above constraint, the system would tend to increase the scale U of the interactions in the course of the optimization, decreasing the effective temperature and making the ground states more stable.

The "effective temperature" $1/U$ deeply influences the characteristics of the optimization. Performing the optimization at a higher effective temperature (small U) produces larger values of the parameter α (more stable chains), provided that the native state has the lowest energy and that the system is still below the folding temperature. This is easily understood, since at a higher temperature more configurations are relevant in the Boltzmann sum defining the average overlap. With a lower

effective temperature, the resulting energy function confers less stability to native states, i.e., the values of $\alpha(\mathbf{S})$ are generally smaller but, since the optimization is less demanding, the number of chains for which the native state has lower energy increases.

Comparison to Existing Methods

We now compare the present optimization method to other existing methods used to derive energy functions, namely the inequalities method and the Z score methods.

In the inequalities method[6–9,15,16] one looks for an energy function such that the native energy is lower than the energy of all alternative structures for proteins in the training set. This condition is imposed by solving a system of linear inequalities. If the training set includes most known protein chains, there is no solution for such a system.[9] This is also the result that we found: in a large training set there are often proteins for which the native state does not coincide with the ground state. In most cases, there is a clear physical ground for this lack of recognition: for instance, interactions with neighboring chains and with cofactors are neglected in the model, but can be essential for the stability of some proteins (see also ref. 7). The same effect may be due to protein-protein interactions in the crystal that may be important for very small proteins. It is, however, very difficult to control all possible sources of errors in a large database. The overlap method singles out problematic proteins in the training set, attributing them a small value of $Q(\mathbf{S}, \mathbf{U})$, while for all other proteins $Q(\mathbf{S}, \mathbf{U})$ is close to unity. Thus, even if some of the training proteins are for various reasons unsuitable, the overlap method generates a good energy function.

Another advantage of the overlap method is that it does not consider only the ground state, but all structures of low energy, and thus imposes that the energy landscape is well correlated. This can indeed also be achieved with the perceptron method, by imposing[6,16]

$$E(\mathbf{C}, \mathbf{S}) - E(\mathbf{C}_{\mathrm{nat}}, \mathbf{S}) > \alpha(1 - q(\mathbf{C}, \mathbf{C}_{\mathrm{nat}})). \qquad (11)$$

An implementation of the perceptron method (the perceptron of maximal stability[9,16]) is more similar to the overlap method, since it allows choosing the largest possible value of α. There is, however, an important difference. With the overlap method, inequalities of the form[11] are obtained, but with stability coefficients $\alpha(\mathbf{S})$ that depend on the sequence considered. More stable proteins have larger values of $\alpha(\mathbf{S})$. In particular, $\alpha(\mathbf{S})$ is also correlated to the length of chain $\mathbf{S}$, and it appears to increase as $\sqrt{N}$. The maximization of the overlap is particularly efficient in imposing the above inequalities with sequence specific coefficients $\alpha(\mathbf{S})$ for different proteins, and it also leaves open the possibility that $\alpha(\mathbf{S})$ is negative if the native state is unstable like in the cases mentioned above.

Other optimization methods are based on the maximization of the stability gap between the native state energy and the average energy of the system.[10–14] This is achieved by minimizing the native Z score[20]:

$$Z(\mathbf{C}_{\mathrm{nat}}, \mathbf{S}, \mathbf{U}) = \frac{E(\mathbf{C}_{\mathrm{nat}}, \mathbf{S}, \mathbf{U}) - [E(\mathbf{C}, \mathbf{S}, \mathbf{U})]_\Omega}{\sqrt{[E(\mathbf{C}, \mathbf{S}, \mathbf{U})^2]_\Omega - [E(\mathbf{C}, \mathbf{S}, \mathbf{U})]_\Omega^2}}, \qquad (12)$$

where $[\,\cdot\,]_\Omega$ denotes average over all alternative structures in Ω.

The optimization of the Z score provides a well-correlated energy landscape for the resulting protein models. It is interesting to see whether the overlap method could give rise to stronger correlations, since the Z score only takes into account the average energy of high-energy structures while the average overlap is mostly influenced by low energy structures. When $Q = 1$, the ground state and the native state are guaranteed to coincide. On the other side, there is no clearly defined threshold value of the Z score that provides the same result. In particular, the typical value of Z for sequences folding to $\mathbf{C}_{\mathrm{nat}}$ depends on the chain length and on the amino acid sequence.

We note that the effective temperature $1/U$ allows us to interpolate the overlap method between two regimes, one similar to the inequalities method and the other similar to Z score methods. The first regime is obtained for large values of U, where we impose the conditions that most native states have lowest energy but do not attempt to optimize their stability. The second regime is obtained for small values of U, when the native structures that have lowest energy are also very stable, but their number is smaller. The choice of an appropriate intermediate value of the effective temperature is, of course, a subtle point of the present method.

Implementation

We optimize $\bar{Q}$ by a gradient descent method. The choice of this method is governed by simplicity and the fact that its performances are good, making more complicated methods unnecessary. We verified that simple gradient descent is able to reach the true optimum with good accuracy. For optimizations starting from random sets of parameters, the correlation coefficient between the end points of the optimization is larger than 0.90, and the value of $\bar{Q}$ in the runs is the same up to three digits. Convergence is reached with a small number of iterations, of the order of a few tens.

We start from a set $\mathbf{U}^0$ of interaction parameters (randomly chosen or obtained through a previous optimization) and apply the following iterative algorithm:

1. We compute the gradient of $\bar{Q}$ using the old set of interaction parameters.
2. We update the parameters according to the rule:

$$\tilde{\mathbf{U}}^{(t+1)} = \mathbf{U}^{(t)} + \delta \sum_{\mathbf{S}} \nabla_{\mathbf{U}} Q(\mathbf{S}, \mathbf{U}) - \delta\gamma \sum_{\mathbf{S}} \left(\frac{1 - q_0(\mathbf{S})}{q_0(\mathbf{S})} \right) \nabla_{\mathbf{U}} E(\mathbf{C}_{\mathrm{nat}}, \mathbf{S}, \mathbf{U}), \qquad (13)$$

where $q_0(\mathbf{S})$ (Eq. 7) is the overlap between the ground state and the native state of sequence $\mathbf{S}$ and depends on the energy function. When the optimization converges,

for most proteins $q_0(\mathbf{S}) = 1$ and the last term disappears. This term is important in the initial phase of the optimization, to avoid that the system is trapped in a suboptimal local maximum of $Q(\mathbf{S}, \mathbf{U})$.

3. We rescale the parameters: $\mathbf{U}^{(t+1)} = 1/\tau\tilde{\mathbf{U}}^{(t+1)}$, to keep constant the quantity U^2, Eq. (10).

Protein Sets

We used two sets of protein structures. The first set (database A) is the WHATIF database,[37] containing 456 crystallographic structures. The second set (database B) is a subset of the latest release of PDB select,[38] containing all protein chains of known structure with only one representative for every class of sequence homology. We exclude from the 1,012 structures in PDB select the chains that have more than 90% sequence homology with chains in database A and are left with 713 chains in database B. The third set (database C) is obtained by joining the two. It contains 1,169 chains, with one representative of every class of structures experimentally known and some homologous sequences, which allow reproducing structures of high overlap. We used all 1,169 structures to generate alternative structures, but only 1,074 structures of length $N \leq 455$ were used to test the energy function: 153 structures determined by NMR spectroscopy and 921 crystallographic structures. From these, 406 are monomeric proteins and 515 are chains that are part of oligomeric proteins.

RESULTS

Lattice Model

We performed the first test of the overlap method on a lattice model. Testing methods for deriving energy parameters on lattice models is an instructive procedure (see, for example, ref. 5), which is particularly useful to compare different methods (for recent results, see refs. 34, 35). However, since for lattice models the functional form of the energy, Eq. 4, is exact, and not approximate as for real proteins, lattice models cannot tell us much on the suitability of an energy function for fold recognition. Here, the task in studying the lattice model is to test the convergence of our optimization algorithm, and to verify that the overlap method produces well-correlated energy landscapes. We present the test as an illustration of these points.

Using the set $\mathbf{U}^{MJ}$ of contact interactions derived by Miyazawa and Jernigan,[2] Abkevich et al.[27] designed a sequence $\mathbf{S}^*$ with $N = 36$ "amino acids" such that its ground state coincided with a target contact map $\mathbf{C}^*$, representing the contact map of a maximally compact self-avoiding walk on the cubic lattice. The sequence was designed by minimizing the energy of the contact map $\mathbf{C}^*$ at fixed amino acid composition. This is equivalent to minimizing the Z score, and, in fact, the resulting sequence has a very negative $Z(\mathbf{C}^*, \mathbf{S}^*)$. The sequence $\mathbf{S}^*$ has a well-correlated energy landscape, and C^* is both very stable and fast folding.[27]

We applied the present method to the sequence-structure pair $\mathbf{S}^*$-$\mathbf{C}^*$, using a random interaction matrix as a starting point. After few iterations, we obtained energy parameters such that (1) $\mathbf{C}^*$ was the ground state of sequence $\mathbf{S}^*$ and (2) the energy landscape was well correlated. We show the latter feature in Figure 1, by plotting the overlap with the ground state, $q(\mathbf{C}, \mathbf{C}^*)$, vs. the normalized energy gap, $1 - E(C, \mathbf{S}^*, \mathbf{U})/E(C^*, \mathbf{S}^*, \mathbf{U})$, where the energies are calculated with the optimized energy parameters. Configurations are generated through a Monte Carlo algorithm. As such, we used the pruned-enriched Rosenbluth method (PERM),[30–32] a chain growth algorithm that is particularly efficient in sampling low-energy configurations for random models of polymers.

The parameter $\alpha(\mathbf{S}^*)$ defined in eq. 6 characterizes the correlations in the energy landscape. With the original parameter set $\mathbf{U}^{MJ}$, the sequence $\mathbf{S}^*$ has $\alpha(\mathbf{S}^*) = 0.23$. This value corresponds to a remarkable stability both with respect to thermodynamic fluctuations and with respect to mutations.[28,29] With the parameter sets determined with the present algorithm the value of α becomes even larger. We performed two optimization runs with different effective temperatures $1/U$, finding $\alpha = 0.34$ for $U^2 = 1.5$ and $\alpha = 0.38$ for $U^2 = 0.75$. As expected, using a higher effective temperature we obtain a more correlated energy landscape. As observed before, in fact, at higher effective temperature more structures give a relevant contribution to the Boltzmann ensemble.

The algorithm was remarkably efficient: after only two iterations we obtained a parameter set such that $\mathbf{C}^*$ was the ground state, and after six iterations we obtained the parameter set from which Figure 1 was drawn. The optimization needed less than 2 h on a Silicon Graphics workstation with a R4000 CPU.

Protein Structures

We will discuss the search for pairwise contact interactions that stabilize protein native states. There is no guarantee that this energy function exists, since the functional form Eq. (4) is only a phenomenological approximation, and it has been shown that this form of the energy function is not sufficient for fold recognition when alternative structures are generated by Monte Carlo.[8,33]

We considered here more than 1,000 structures, which are representative of all protein structures known to date. We derived energy parameters capable (1) of stabilizing most proteins against decoys produced by threading and (2) of identifying the "troublemakers," i.e., the sequences that, for different physical reasons, are impossible to stabilize. We found that most of such troublemakers are characterized by the neglect of important interactions, such as interactions with other chains forming the protein or interactions with cofactors (see also ref. 7).

We distinguish three classes of proteins. Monomeric proteins are the least problematic. For oligomeric proteins, it is possible to generate alternative structures by threading the sequences on the structure of a very long protein (longer than the sum of the chains composing it),[7] but only a small number of decoys can be obtained in this way. Thus, we decided to consider only one chain of oligomeric proteins at a time, treating it as a monomeric protein. In

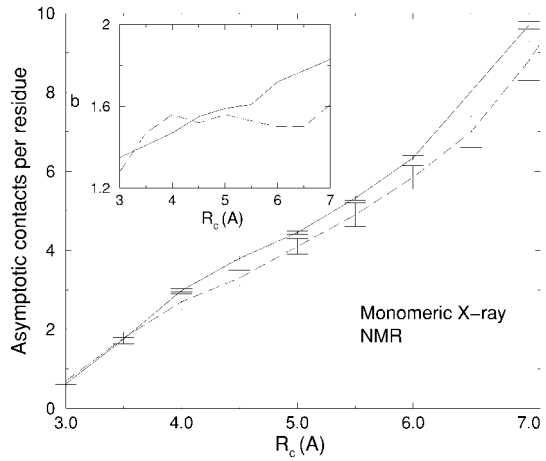

Fig. 2. Asymptotic number of contacts per residue c and surface parameter b (**inset**) as a function of the threshold distance for contacts.

doing this, we only take into account intrachain contacts and neglect interactions between different chains. Thus, it can be expected that the resulting model is less stable than for the case of oligomeric proteins.[7] This is exactly what happens, and many oligomeric proteins are not stabilized by the present potential, as expected. It would be advisable to eliminate such proteins from the training set,[6,7] and we did it in one of the present optimization runs. Notice, however, that in automatically reading a large database of PDB files, sometimes some oligomeric proteins can be misinterpreted as monomeric. It is, thus, important to verify that our method is rather robust with respect to insertions of "wrong examples" in the training set. The third class is constituted of structures determined by NMR. As we shall see, several observations suggest that they have to be considered separately.

For monomeric globular proteins, the average number of contacts per residue should be given by an asymptotic value c for infinitely long chains minus surface corrections:

$$N_c/N \approx c(1 - bN^{-1/3}). \tag{14}$$

Surface scaling predicts the value $b = 1.5$ for the surface parameter. The vicinity of the measured value of b to the predicted one is an assessment of the quality of the threshold distance R_c used in the definition of contacts. For small threshold distances, the number of contacts per residue is nearly independent of N and b is small. When the threshold is comparable to the size of small chains, the contacts per residues are much reduced for these chains and the estimated value of b as a result increases. The asymptotic number of contacts per residue, c, and the surface parameter b are shown in Figure 2 as a function of the threshold distance R_c for structures of monomeric proteins determined either by X-ray or by NMR spectroscopy.

For monomeric proteins determined from crystals and for the threshold value $R_c = 4.5$ Å, a least square fit yields the values $b = 1.55 \pm 0.1$ for the surface parameter and $c = 3.9 \pm 0.1$ for the asymptotic number of contacts per residue in the interior of the globule. Thus, an average residue in the core of the protein has nearly 12 geometrical contacts, two in each Cartesian direction: four contacts times two (because contacts are shared between two residues) plus four geometrical contacts with neighboring residues along the chain, which are not counted in N_c (remember that we only consider contacts at a distance larger than two along the chain). These results and the results of Vendruscolo et al.[9] make us confident that the threshold distance for contacts is chosen in a reasonable range.

As mentioned above, we consider in this work only contacts between amino acids belonging to the same chain (intra-chain contacts) while interchain contacts between different chains are not considered, even if they are important for the stability of the native state. The consequence is that in the present treatment, chains forming oligomeric proteins have on the average less contacts per residue than monomeric proteins. Since the native state of these chains is the structure observed in the complete protein, its stability in the model should be expected to be lower than for monomeric proteins,[7] and, in fact, for several of them the ground state in the model does not coincide with the experimental native state.

Structures determined by NMR are also studied separately. Their number of contacts is typically smaller than for X-ray structures of the same length (see Fig. 2). We find $c = 3.3 \pm 0.2$ for $R_c = 4.5$ Å. The number of contacts are also influenced by the residue type. In this context, hydrophobicity and the size of the side chain play a major role.

To conclude section, we observe that the number of alternative structures generated by threading is a rapidly decreasing function of the chain length N. Using the largest database (database C), for chains shorter than $N = 100$, we generated more than 100,000 structures. The number of alternative structures generated by using only NMR structures is smaller by a significant factor.

Overlap and RMSD

In this section, we discuss the relationship between the overlap q and the root mean square deviation (RMSD), traditionally used as measures of pairwise similarity between protein structures.

Since a contact map does not define a unique structure, protein structures corresponding to the same contact map (i.e., having $q = 1$) may have a RMSD different from zero, but typically smaller than one Å. As q decreases from one, the RMSD typically increases and finally reaches the value corresponding to random pairs of structures.

For a given overlap, the RMSD is influenced mainly by three factors: (1) the number of contacts per residue, $c = N_c/N$, (2) the fraction ϕ of non-local contacts, and (3) the chain length.[36] In the following, non-local contacts are defined as contacts between amino acids separated by more than 10 residues along the sequence. For fixed q, the larger are c and ϕ, the smaller is the RMSD (on the average). Consider, for example, the case of two structures, one consisting of a single α helix and the other one consisting of an helix-turn-helix motif. They both have

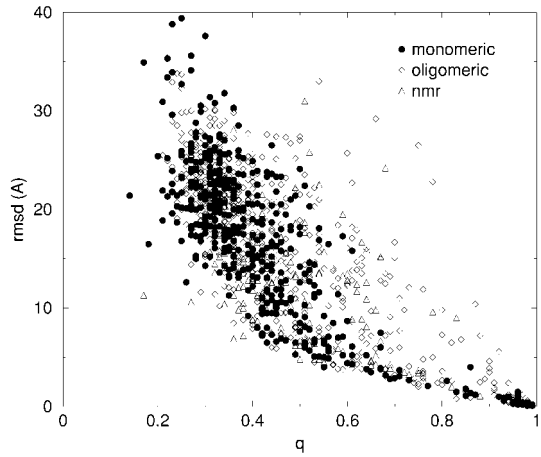

Fig. 3. RMSD plotted vs. the overlap for 1,234 protein pairs.

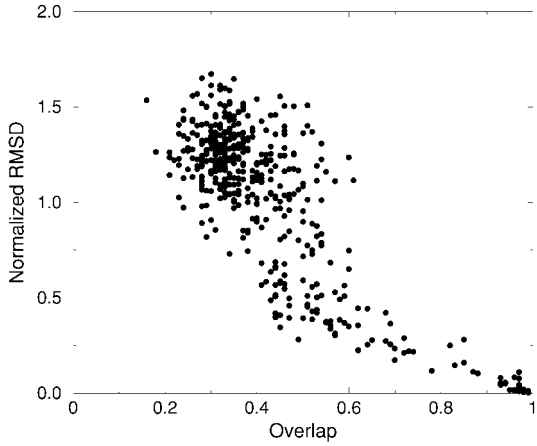

Fig. 4. Normalized RMSD vs. overlap for protein pairs involving monomeric proteins.

small values of c and ϕ. Despite the large overlap, their RMSD is large.

In this part of the study, we consider all protein chains in databases A and B, including chains with the same sequence, which take part in different oligomeric proteins and have slightly different structures. For each contact map $\mathbf{C}$, we selected the alternative contact map $\mathbf{C}'$ with the largest overlap $q(\mathbf{C}, \mathbf{C}')$. Then we computed the RMSD of the C_α atoms of the corresponding structures and the sequence similarity q_{seq} between the sequence of the first protein and the sequence of the protein from which $\mathbf{C}'$ was derived, aligning $\mathbf{C}$ and $\mathbf{C}'$ without allowing for gaps. The sequence similarity is defined as the fraction of positions in which the same amino acid appears in the two sequences. Pairs of chains with $q_{\mathrm{seq}} = 1$ have exactly the same sequence, but they can have different structures either because they appear in different protein complexes or because some experimental conditions change. The quantity q_{seq} gives a measure of the evolutionary relationship between the two sequences: pairs of chains with large q_{seq} are closely related evolutionarily. We also consider chains with large sequence homology in the database of protein structures in order to sample large values of q, which are rare for chains without sequence homology. In this context, gapless threading is a poor method to generate analogous chains, i.e., sequences with small q_{seq} and high structural overlap. Nearly all chains with high overlap generated by gapless threading are homologous, i.e., they also have high q_{seq}. This happens because during molecular evolution, a protein sequence undergoes a process of insertion or deletion.

Since we adopt the convention of considering only C_α atoms in the calculation of the RMSD, we define a contact when two C_α atoms are closer than $R_c = 10$ Å. We distinguish three classes of chain pairs: (1) those whose first chain is a monomeric protein; (2) those whose first chain is part of an oligomeric protein; (3) those whose first chain is determined by NMR. Figure 3 shows a scatter plot of the RMSD vs. the structural overlap. The points can be approximately fitted to an exponential curve:

$$\mathrm{RMSD} \simeq A \exp(a(1 - q)) \tag{15}$$

The fit yields a correlation coefficient $r = -0.89$. The correlation is stronger for monomeric proteins ($r = -0.94$) than for chains in oligomeric proteins ($r = -0.88$) and NMR structures ($r = -0.61$). For monomeric proteins, structures with $q > 0.7$ have RMSD < 4 Å, and the parameters of the fit are $A = 0.36 \pm 0.03$ Å and $a = 6.1 \pm 0.1$. Chains in oligomeric proteins and structures determined by NMR have on the average less contacts than crystal structures of monomeric proteins. This may explain why the correlation between the overlap and the RMSD is much weaker for them. The points with the largest RMSD values for a given q usually correspond to small chains in oligomeric proteins, which have few contacts per residue and few nonlocal contacts (they may even consist of a single α helix).

While the overlap is not very sensitive to chain length, the RMSD between protein chains depends strongly on their length. Maiorov and Crippen[36] proposed using a normalized RMSSD, which is much less sensitive to chain length. We studied the correlation of the overlap with this new measure, finding results very similar to those previously exposed. The best fit of the normalized RMSD as a function of the overlap is again exponential, with correlation coefficients $r = -0.92$ for monomeric proteins, $r = -0.87$ for oligomeric proteins, and $r = -0.69$ for NMR structures, close to the values obtained for normal RMSD. We show in Figure 4 the resulting scatter plot for monomeric proteins.

We also studied the RMSD as a function of sequence similarity q_{seq}. The dependence of the RMSD on q_{seq} is very similar to the dependence on the structural overlap. Also, in this case an exponential curve fits the dependence well, and the quality of the fit is better for monomeric than for oligomeric proteins and NMR structures. The values of the coefficients of correlation are similar to those given above. For crystal structures of monomeric proteins, we find the parameters $A = 0.30 \pm 0.02$ Å and $a = 4.37 \pm 0.08$. A scatter plot of the structural overlap q vs. sequence similarity q_{seq} is shown in Figure 5. The coefficient of correlation is $r = 0.80$ when all the chains are considered. Also, in this case the correlation is the best for monomeric

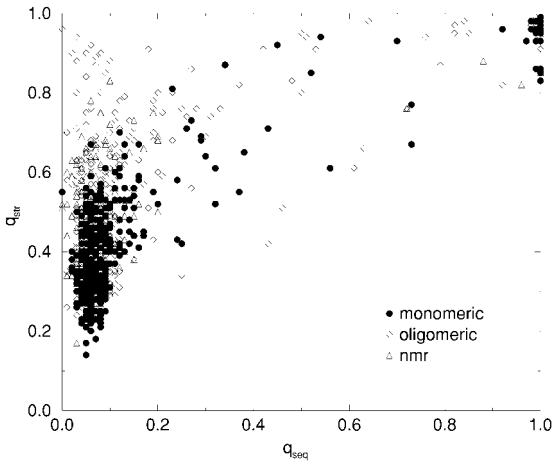

Fig. 5. Structural overlap q vs. sequence similarity q_{seq}.

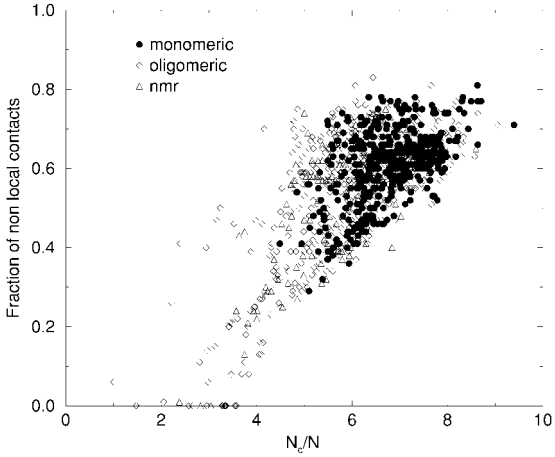

Fig. 6. Fraction of non-local contacts vs. contacts per residue.

proteins: $r = 0.90$, compared to $r = 0.77$ for oligomeric proteins and $r = 0.47$ for structures determined by NMR. The correlation increases if higher powers of q are considered: for instance, for monomeric proteins, the coefficient of correlation between q^2 and $q_{\rm seq}$ is $r = 0.95$.

The structural overlap q is, in general, larger than the sequence similarity $q_{\rm seq}$. This is in agreement with the observation that evolution conserves protein structures much more than protein sequences. For this reason, algorithms of structure comparison can detect large structural similarity even when the sequences have very low homology.[39]

Figure 6 shows that the fraction of non-local contacts is positively correlated to the number of contacts per residue. For all three sets, data are in agreement with the simple scaling law,

$$\frac{N_{\rm nl}}{N_c} \approx 1 - \left(1 + a\,\frac{N_c}{N}\right)^{-1}, \qquad (16)$$

where $N_{\rm nl}$ is the number of non-local contacts, N_c is the total number of contacts, and the best fit value is $1/a = 2.7 \pm 0.1$. Such scaling law can be expected from the fact

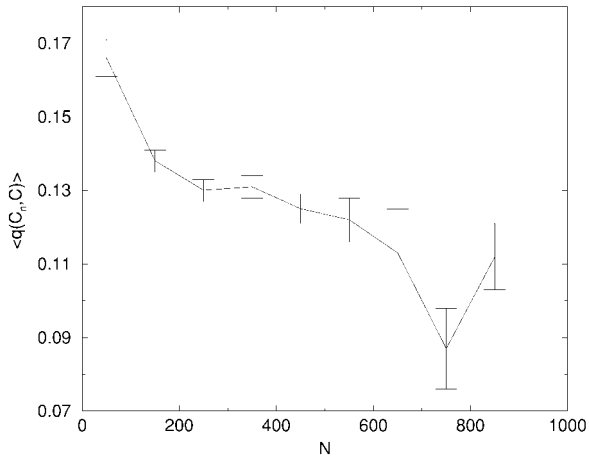

Fig. 7. Average overlap between structures with N residues.

that, as the number of contacts per residue increases, the number of local contacts reaches a limiting value. The number of non-local contacts is related to the contact order, a parameter that has been shown to be an important determinant of folding kinetics.[40]

The last question that we address in this section concerns whether the overlap is a suitable measure for the comparison of unrelated structures. Naively, it could be expected that it has to vanish in the limit of very long chains. Instead, it has been recently shown that even two random lattice structures have a finite fraction of common contacts.[41] This is due to local contacts between residues that are neighbors in the chain, which have a non-vanishing probability of being close in two random structures. In the case of protein structures, this effect is even larger, due to the presence of secondary structure, which determines regular patterns of contacts: the contacts $(i, i + 4)$ are present in α-helices, the contacts $(i, i + l), (i + 1, i + 1 + l), \cdots$ are present in parallel β-sheets, and the contacts $(i, i + l), (i + 1, i - 1 + l), \cdots$ are present in antiparallel β-sheets. These secondary structure contacts are a finite fraction of the total number of contacts even in the limit of very large chains.

For each chain, we measured the average overlap between the native structure and all alternative structures generated by threading, excluding those generated from homologous chains. We further averaged the overlap over chains whose length is in a bin of 100 residues width. We used the all atom definition of contact with $R_c = 4.5$ Å. In this case, the average number of contacts in the globule is 3.9. The result is shown in Figure 7. For long chains, the number of alternative structures is quite low and the statistical relevance is rather poor. Even for structures with more than 600 residues, the typical overlap does not decrease significantly below 0.1. These results are in qualitative agreement with a recent study by Micheletti et al.[42]

Energy Parameters

Here we report the derivation and test of contact energies for protein folding. We derived three different sets of

interaction parameters, $U^{(1)}$, $U^{(2)}$ and $U^{(3)}$, by increasing progressively the size of the training set.

In order to derive $U^{(1)}$, we chose as a training set 40 proteins from database A, whose chain lengths range from 36 to 128 residues. We generated alternative structures by threading the sequences along 100 different contact maps, including the ones derived from the training set. The optimization algorithm was run with the following parameters: $\delta = 0.2$, $\gamma = 0.02$, and $U^2 = 0.2$. Starting from a set of random energy parameters, after 14 iterations of the optimization procedure (which took a CPU time of 1.5 h), we found the set $U^{(1)}$ of interaction parameters. For this set, 38 chains had $q_0 = 1$, which means that ground state and native state coincide (for a definition of q_0, see Eq. 7), and two had $q_0 = 0.5$ and $q_0 = 0.1$, respectively. They were both from oligomeric proteins. The value of $\bar{Q}$ (eq. 9) for the 38 proteins with the correct ground state was $\bar{Q}(U^{(1)}) = 0.95$.

The predictive power of the parameters $U^{(1)}$ was blindly tested on the remaining proteins of database A with chain length $N \leq 390$, containing 402 chains. The number of alternative structures was always larger than 5,000. In 90% of the cases, we obtained $q_0 = 1$ and an average value of $\bar{Q} = 0.94$. Of the 39 protein chains whose native state was not recognized, 27 were from oligomeric proteins and only 12 were monomeric proteins, out of a total of 247 monomeric proteins. Thus, more than 95% of the monomeric proteins are recognized in this blind test.

Next, we increased the training set by adding some of the monomeric proteins that were not recognized. With a training set of 47 proteins we derived the parameter set $\mathbf{U}^{(2)}$. This performed better than $\mathbf{U}^{(1)}$: over the whole database A, for only 29 proteins was the ground state $C_0(\mathbf{S})$ different from the native structure $C_{\rm nat}(\mathbf{S})$, and all mistakes were due to oligomeric proteins. The proteins for which $\mathbf{U}^{(2)}$ selects the true native structure have $\bar{Q} = 0.93$; thus, they are, in general, very stable. We could not determine a significantly better parameter set: even using all of the 456 proteins as training set, the number of chains such that $C_0(\mathbf{S})$ differs from $C_{\rm nat}(\mathbf{S})$ decreased only from 29 to 27. We found that energy parameters optimized with respect to different training sets of proteins are rather similar, and most of the time they yield very similar stability for a given protein: For some proteins, the value of Q is close to unity for all different parameter sets, and for some others the value of Q is always low, even when they are inserted into the training set. For only a few proteins did Q change significantly when we put them in the training set. The robustness of the results with respect to the change of the training set is remarkable. This indicates that the overlap method gives a very high stability to native structures, which is not significantly affected when parameters are changed. We tested the energy parameters $\mathbf{U}^{(2)}$ also on database B. This was a blind test, because the proteins in the test set are not sequence homologous to the proteins in the training set. The test set was divided in three classes:

1. 196 monomeric proteins: in this class, only in one case did the ground state differ from the native structure and in another case the stability of the native structure was low ($Q = 0.5$). The fraction of proteins for which $Q > 0.6$ is thus 99.5%.
2. 383 oligomeric proteins: in this class, in almost 20% of the cases the ground state does not coincide with the native structure. The fraction of proteins for which $Q > 0.6$ is 74%.
3. 136 structures determined by NMR. About 50% of these proteins have a ground state that does not correspond to the native structure. The fraction of proteins for which $Q > 0.6$ is 36%.

These results indicate that predictions for NMR structures are more difficult than those for crystallographic structures. The necessity to distinguish between NMR and crystallographic structures while deriving energy functions was already pointed out by Godzik et al.[43] They observed that energy parameters obtained by a statistical analysis of X-ray structures perform rather poorly for the recognition of NMR structures, and vice versa.

Finally, we describe the derivation of the set $\mathbf{U}^{(3)}$ of energy parameters. To this end, we performed some runs of optimization using all the chains with $N \leq 200$ in database C as training set. Since this set contains sequence homologous chains, structures of high similarity with the native state are present in the database of structures, and the recognition becomes more difficult. We tested the resulting energy parameters $\mathbf{U}^{(3)}$ on a test set containing 1,079 proteins with $N \leq 455$. This was not a blind test, because the shorter chains, which have more alternative structures and thus are more difficult to recognize, appear in both the test set and the training set. The results are as follows:

1. For 401 of the 406 *monomeric* proteins we found $q_0 = 1$ and $Q(\mathbf{S}) > 0.7$. The exceptions are proteins with cofactors and small fragments.
2. For 83% of the 515 chains in *oligomeric* proteins we found $q_0 = 1$ and $Q(\mathbf{S}) > 0.7$.
3. For 68% of the 153 *NMR structures* we found the correct ground state but only 60% of them have $Q(\mathbf{S}) > 0.7$.

The energy parameters $\mathbf{U}^{(3)}$ are reported in Table I.

Stability Parameter α

The three classes of proteins discussed above are characterized by a different average stability. The observed ranking is confirmed by the analysis of the stability parameter α defined in Eq. (6). A large stability parameter implies a well-correlated energy landscape and a stable ground state. Monomeric proteins have, on the average, the largest values of α, followed by oligomeric proteins and then by NMR structures. The parameter $\alpha(\mathbf{S})$ is correlated to chain length N, and it increases approximately as $\sqrt{N}$. The average value and the standard deviation of $\alpha(\mathbf{S})/\sqrt{N}$ as a function of N for the three classes of proteins is shown in Figure 8. These data include only chains whose native

TABLE I. Best Contact Interaction Matrix Derived in This Work[†]

	Ala	Glu	Gln	Asp	Asn	Leu	Gly	Lys	Ser	Val	Arg	Thr	Pro	Ile	Met	Phe	Tyr	Cys	Trp	His
Ala	−0.0479	0.1346	−0.0457	0.1018	0.1049	−0.1711	0.1844	0.0691	0.0464	−0.1431	0.1049	0.0310	0.1462	−0.0737	−0.0847	−0.1119	−0.1408	−0.1085	−0.0880	0.0266
Glu		0.1259	−0.0413	0.1581	0.1146	0.0802	0.2311	−0.2403	0.0823	0.1010	−0.3511	0.0675	0.2241	0.1103	0.0637	0.0885	−0.0522	0.1550	−0.0967	−0.0827
Gln			−0.0550	−0.0728	−0.0050	−0.0172	0.1710	−0.0735	0.1169	0.1061	0.0059	−0.0243	0.1127	−0.0480	−0.1038	−0.0171	−0.1431	0.0715	−0.0540	−0.0125
Asp				0.0840	0.0192	0.2673	0.1115	−0.1154	0.0424	0.2728	−0.1859	0.1043	0.2386	0.1892	−0.0197	0.0827	−0.1165	0.1169	−0.0124	−0.0749
Asn					−0.0917	0.0890	0.1196	−0.0381	0.1452	0.1180	−0.0150	0.0155	0.1560	0.1485	0.0124	0.0018	−0.1149	−0.0844	−0.0250	0.0386
Leu						−0.5067	0.0782	0.0543	0.0959	−0.4593	−0.0651	−0.0316	0.0745	−0.5112	−0.1822	−0.5450	−0.2614	−0.1305	−0.2639	−0.0169
Gly							0.2219	0.1963	0.1075	0.1859	−0.0251	0.1763	0.2131	0.1174	−0.0573	0.0789	−0.0176	−0.0982	−0.1567	0.0979
Lys								0.1216	0.1690	0.0609	0.0839	0.0467	0.1099	0.0682	0.0866	−0.0416	−0.1120	−0.0330	−0.1152	0.0390
Ser									0.0941	0.1766	0.0442	0.0228	0.1626	0.0332	0.0185	0.0398	0.0214	−0.0132	−0.0802	−0.0005
Val										−0.5193	0.0475	0.0119	0.0868	−0.4223	−0.2127	−0.4001	−0.2792	−0.2349	−0.2898	−0.0039
Arg											0.0306	−0.0210	−0.0614	−0.0266	−0.0163	−0.0904	−0.1369	0.0544	−0.2070	−0.0184
Thr												0.0150	0.1908	−0.0700	0.0018	−0.1120	−0.1445	−0.0013	0.0052	0.0681
Pro													0.1077	0.0882	−0.0069	−0.0604	−0.1326	0.0545	−0.0910	0.0295
Ile														−0.5852	−0.2137	−0.3791	−0.3164	−0.2235	−0.1961	−0.0326
Met															−0.1059	−0.1785	−0.1621	−0.0557	−0.0775	−0.0345
Phe																−0.3088	−0.4212	−0.3262	−0.3405	−0.1250
Tyr																	−0.2793	−0.2444	−0.3209	−0.1976
Cys																		−1.0442	−0.1176	−0.0701
Trp																			−0.1066	−0.0200
His																				0.0005

[†]A contact is defined through a threshold of 4.5 Å for all atom pairs. Energies are given in units of k_BT.

Fig. 8. Average value of the stability parameter $\alpha(\mathbf{S})/\sqrt{\mathbf{N}}$ vs. N for three classes of structures. The error bars are estimates of the standard deviation.

Fig. 9. Z score vs. stability parameter $\alpha(\mathbf{S})$ for monomeric proteins.

state is recognized (i.e., with $q_0 = 1$), otherwise the stability parameter would be zero or negative, and with less than 400 amino acids. X-ray structures of monomeric proteins have a larger stability parameter than X-ray structures of oligomeric proteins and NMR structures. The corresponding distributions, however, have a considerable overlap. These distributions have also been reported in Bastolla et al.[17]

Finally, we note that the stability parameter $\alpha(\mathbf{S})$ is correlated to the Z score, Eq. (12), with coefficient of correlation $r = -0.64$ (see Fig. 9). There are, however, few cases of chains with small α and very negative Z scores.

Unstable Structures

We now take a closer look at the structures that we failed to stabilize. They are characterized by having $q_0 < 1$, i.e., the overlap method did not recognize their native structures. We discuss first structures derived from crystallographic data. NMR structures will be discussed in a following section.

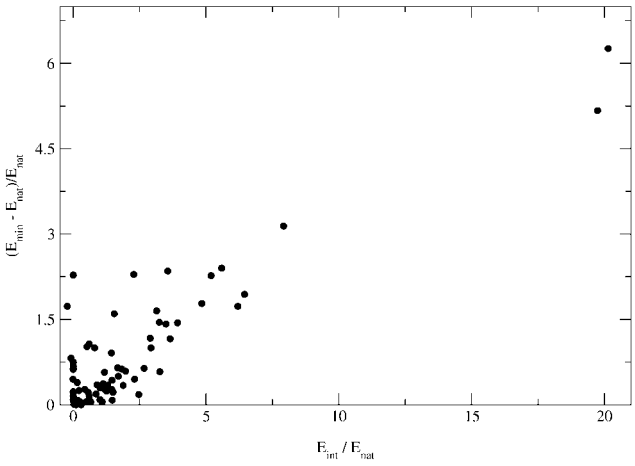

Fig. 10. Energy difference between the ground state and the native state vs. the energy of interaction with other chains for 72 chains from oligomeric proteins determined by X-ray analysis.

As already noted, interactions between different chains forming an oligomeric protein and interactions with cofactors are neglected in the present model. As a result, some ground states of chains in oligomeric proteins do not coincide with the experimentally observed native states, and those that do coincide are less stable than for monomeric proteins. We did not generate alternative configurations of a whole protein composed by several chains and did not test whether the native structure was the ground state of this system, but we did measure the interactions between chains in the native structure. We found that in nearly all cases, the total energy of the native structure was lower than the sum of the energies of the ground states of the isolated chains.

The plot of the difference between the ground state energy $E_{\rm min}$ and the native state energy of the isolated chain, $E_{\rm nat}$, vs. the interaction energy with other chains $E_{\rm int}$ is shown in Figure 10. These results refer to all chains in oligomeric proteins for which all coordinates are available in the PDB. Both energies are normalized by dividing them by $E_{\rm nat}$. Since all three quantities $E_{\rm nat}$, $E_{\rm nat} - E_{\rm min}$, and $E_{\rm int}$ are negative, the normalized energies $(E_{\rm min} - E_{\rm nat})/E_{\rm nat}$ (instability energy) and $E_{\rm int}/E_{\rm nat}$ represented in Figure 10 are positive. The only exceptions are three chains for which the interaction energy $E_{\rm int}$ turns out to be repulsive (1mabG, 2drpD, 1isuA). In these proteins, large cofactors are present; therefore it is possible that the cofactors are responsible for stabilizing the different chains together.

The two normalized energies $(E_{\rm min} - E_{\rm nat})/E_{\rm nat}$ and $E_{\rm int}/E_{\rm nat}$ are significantly correlated, with a correlation coefficient $r = 0.88$. This result suggests that the alternative structures that we generated are able to estimate the interaction energy neglected in the model. The instability energy is generally smaller than the interaction energy; thus, in most cases, the native energy of the system is lower than the sum of the energies of the isolated chains composing it. The highest value of the instability energy is found for two virus coat proteins, not represented in the plot, for which $(E_{\rm min} - E_{\rm nat})/E_{\rm nat}$ is, respectively, 15.0 and 19.0.

We also estimated the interaction energy with cofactors computing the number of contacts and assuming that one contact with a cofactor contributes with an energy equal to the average energy of a native contact. This is only an approximation and it might underestimate the interactions if the cofactors are charged or covalently bound. For 72 out of the 87 problematic crystallographic structures for which the coordinates of all chains and cofactors are available, we checked whether the sum of the native energy plus these interaction terms is lower than the ground state energy of the single chain. This is true in all but nine cases. The nine exceptions are listed in Table II and discussed below.

Two of these chains (3cyr and 1isuA) have large cofactors covalently bound, four heme groups for 3cyr and an iron-sulfur cluster for 1isuA, and usually ligands contribute significantly to the stability. In 1aws, four residues are modified into selenomethionine, and this may explain why the native energy estimated is too high by a small amount. Five of the remaining chains are small, and all of them (except 1ail) have three or four disulfide bonds. Two are inhibitors bound to enzymes and two are small fragments. It is possible that the interactions with other molecules of the crystal influence their X-ray structure.

We investigated in detail the case of crambin, which is a highly hydrophobic protein. There are 13 structures available for this protein in the PDB. Four of them were determined in X-ray experiments (1cbn, 1ab1, 1cnr, 1crn) and nine (eight conformers in 1ccm and one average structure 1ccn) were determined in one experiment of NMR spectroscopy. The four X-ray structures are very similar, with an average overlap of 0.98, and so are the nine NMR structures, with an average overlap of 0.91. However, the comparison between X-ray structures and NMR structures yields a significantly lower overlap, $q = 0.85$. Moreover, the average number of contacts is 7% larger for NMR structures, and the average contact energy is 15% lower, despite the fact that usually NMR structures have less contacts and are less stable than crystallographic structures. In the test of the present energy function, we used the structure 1cbn, determined by X-ray with resolution of 0.83 Å. The NMR structure of lowest energy has an energy 28% lower than the energy of 1cbn, and only slightly higher than the energy of the false ground state of the model.

A possible explanation of these differences is that the crystal environment significantly changes the structure of crambin with respect to the one in solution.[18] This is even more likely since the conditions of crystallization were peculiar, with temperatures between 130 and 150 K for different X-ray experiments.

The suggestion that the crystal modifies the structure of small proteins and protein fragments can explain all remaining cases in Table I where the native state is not the ground state of the model. The most problematic protein would then be interferon γ (1rfbA). The validation of this structure with WHAT-CHECK[37] identifies several serious

TABLE II. Nine Most Problematic Crystallographic Structures[†]

PDB id.	Protein	Remarks	N	q_0	E_0/E_{nat}	E_{int}/E_{nat}
1awsA	Isomerase	Selenomethionines	164	0.94	1.07	0.01
1rfb A	Interferon γ	Incorrect structure	119	0.27	1.25	0.22
3cyr	Cytochrome c3	Heme	107	0.16	1.53	0.35
1ail	RNA binding (fragm.)		70	0.36	1.19	0.00
1isu A	Electron transf.	Iron-sulfur cluster	62	0.20	2.73	−0.11
1tgs I	Inhibitor/tripsin.	SO_4, Ca	56	0.08	1.39	0.15
1fle I	Inhibitor/protease		47	0.16	2.07	0.60
1cbn	Crambin	Ethanol	46	0.14	1.34	0.02
1ajj	Receptor (fragm.)	SO_4, Ca	37	0.08	2.74	0.00

[†]The columns show PDB code, name, cofactors if any, number of residues N, overlap q_0 between the ground state of the model and the native state, ratio between their energies (note that both are negative), and ratio between the interaction energy that is neglected in the model and the ground state energy.

inconsistencies in the deposited structure: 1rfbA has highly anomalous bond lengths and bond angles; one third of the residues have bad torsion angles and the Ramachandran plot Z score is pathologically low; the distribution of residues in the inside and outside of the protein is highly unusual; there are 337 interatomic distances abnormally short; the packing is highly unusual with a pathologically low Z score; the backbone conformation is highly unusual and there are 11 water molecules without hydrogen bonds.

This observation indicates that the present energy function is able, using simple alternative structures determined by gapless threading, to identify incorrect structures in the PDB.

Disulfide Bridges

In this section, we discuss the problem of disulfide bridges, which are the strongest pairwise interaction in the present potential, as should be expected from the fact that these residues form covalent bonds. However, not all the cysteine-cysteine contacts form disulfide bridges: for instance, when three cysteine residues come together, only two of them form a disulfide bridge. A possible way to study this problem is to introduce a new energy parameter representing the cysteine-cysteine interactions that are real disulfide bridges. For every cluster of m cysteines in mutual contact, the number of disulfide bridges is evaluated as $[m/2]$, where the square brackets represent the integer part. Thus, we can distinguish between two different kinds of cysteine-cysteine contacts and apply the usual optimization procedure. Surprisingly, the two parameters for the cysteine-cysteine interaction are set to be nearly equal by the optimization scheme. The reason for this result is perhaps that it is very rare that cysteine residues come together without forming a disulfide bridge. In most cases, a cysteine not involved in a disulfide bridge forms a covalent bond either with another chain or with a cofactor, for instance with heme. Since we do not consider these interactions, the structure is much less stable, and the optimization of the energy function tends to stabilize it by increasing the value of the cysteine-cysteine interactions. We conclude that the database of structures that we used does not enable us to derive accurately the value of cysteine-cysteine interactions that do not form disulfide bridges.

NMR Structures

For structures determined by NMR spectroscopy, the fraction of proteins for which the native state and the ground state do not coincide is much higher than for X-ray structures. The situation does not improve by including in the training set only NMR structures, but improves using both X-ray structures and NMR structures to generate decoys. In this case, the fraction of correctly recognized NMR structures increases only from 68 to 72%.

There are at least two differences between crystallographic structures and NMR structures. The first difference is that NMR structures are usually represented in the PDB as an ensemble of several conformers (typically twenty), all compatible with the experimentally determined NOE restraints. Sometimes the average structure is also deposited. In the present study, we used the first conformer of the PDB file to compute the native contact map. This choice is arbitrary and in the following part of this section we discuss possible alternatives.

We observed that the contact maps of different conformers can be rather different, so that their contact energies can also be rather different. The overlap averaged over all groups of conformers and all NMR proteins with more than one conformer is shown in Figure 11 as a function of the threshold R_c used in the definition of the contact. The limiting value $q = 1$ should be reached when R_c exceeds the linear size of the protein. However, for all reasonable values of R_c the average overlap is considerably below 1, indicating that the differences in contact map (thus also in contact energy) for different conformers are sizeable. Although this result might be an artifact due to the contact energy function, it can also indicate a real difference in energy among conformers. As a consequence, we found that, among the 49 chains whose native state was not recognized, there were 11 conformers whose energy was lower than the one of the false ground state of the model. Therefore, they would have been recognized had we used them for recognition.

This observation suggests considering the whole ensemble of conformers as a "ground state." In order to do that in a self-consistent way, one can use for each conformer α a Boltzmann weight, $\exp(-\beta E_\alpha)$, determined using the contact energy function. The resulting optimiza-

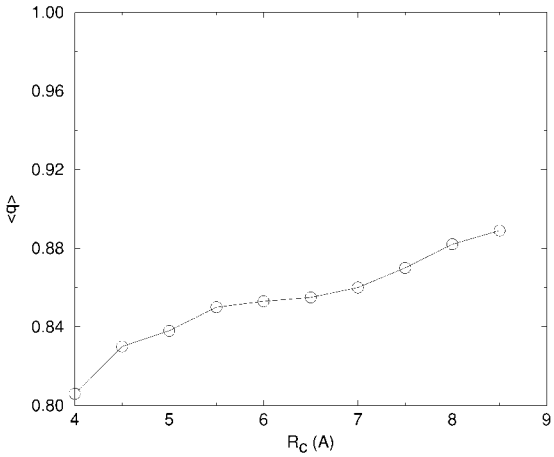

Fig. 11. Average overlap between conformers of NMR structures as a function of the threshold distance for contacts.

tion procedure is more complicated but still feasible. The average overlap to be maximized is

$$Q(\mathbf{S}, \mathbf{U}) = \frac{1}{Z_2} \sum_{\alpha \in \Omega', \beta \in \Omega} q(\mathbf{C}_\alpha, \mathbf{C}_\beta) e^{-(E(\mathbf{C}_\alpha)+E(\mathbf{C}_\beta))},$$

$$Z_2 = \sum_{\alpha \in \Omega', \beta \in \Omega} e^{-(E(\mathbf{C}_\alpha)+E(\mathbf{C}_\beta))}, \qquad (17)$$

where Ω is the threading set of possible configurations for sequence $\mathbf{S}$ including the native ones and $\Omega' \subset \Omega$ is the subset of native conformers.

The second difference that we discuss has a physical origin. NMR spectroscopy is used for the determination of a protein structure in solution. The solution structure is, in general, different from the crystal structure of the same protein. In particular, proteins in solution have larger structural fluctuations than in a crystal, as also suggested by the fact that loop conformations are typically different in NMR conformers. We observed a small but significant difference in the asymptotic number of contacts, determined by the surface scaling of Eq. 14. X-ray structures have asymptotically 3.9 ± 0.1 contacts per residue, while NMR structures have 3.3 ± 0.2 contacts per residue. Therefore, we derived energy parameters by using only NMR structures to generate alternative conformations. This makes the recognition problem much easier, since only a few thousands of alternative structures are produced for each chain. Instead of using Eq. 17, we derived the contact map from the first conformer in the list. We then tested the whole set of conformers. Quite often, when several conformers are listed, we found at least one of them with lower energy than the conformer used in the training set. For chains that were not recognized, however, this energy is still higher than that of the false ground state.

The recognition rate increases by increasing the threshold distance for contacts, R_c. This is probably due to the increase in the number of contacts per residue. The recognition rate improves also by lowering the effective temperature $1/U$ in the optimization procedure. As discussed above, a lower optimization temperature makes the model less stable (the parameters $\alpha(\mathbf{S})$ are reduced) but it can increase the number of chains recognized. The fact that for NMR structures a lower optimization temperature is needed is consistent with the fact that they are less stable.

Our best recognition has been obtained with $R_c = 5.5$ Å and $U^2 = 0.6$. For these parameters, and decoys generated only from NMR structures, the optimized energy function recognizes 93% of the NMR structures. Among the ten problematic chains, there are some that are peculiar: two oligomeric proteins (1wtuA, 1hrzA), one chain studied in reducing environment so that disulfide bonds are not formed (1tiv), one virus coat protein studied in a lipid environment (2cps), one coagulation factor studied with a large number of carboxy groups (1cfh), and one membrane protein (1aty). For all these proteins, it is not surprising that the recognition failed. In particular, since membrane proteins have a different environment, their effective energy function, obtained by averaging over the degrees of freedom of the membrane and of the solvent, should be different from that of water solvable proteins, and they should not be included in the training set. It is nevertheless possible, as in the present case, that a membrane protein mistakenly included in the training set is not sorted out when the PDB file is automatically read. Even in this case, the overlap method gives reasonable results, namely most water solvable proteins are stabilized, but the membrane protein is unstable, despite its inclusion in the training set. The remaining four chains are fragments, from bacteriorhodopsin (1bct), a glycoprotein (1tle), a hormone (1zwa), and one plasminogen activator (1tpn).

To summarize this part of the study, we found that NMR structures are more problematic than crystallographic ones, and that it is not convenient to include the two groups in the same training set. The reasons for this different treatment are listed below. First, native conformers constitute a heterogeneous ensemble. Second, solution structures and crystallographic ones have slightly different average properties, for example, different compactness. Third, the solution structure of a given protein might be different from its crystallographic one. We observed this for the case of crambin. Fourth, the kind of structures studied by NMR are often peculiar. They are smaller chains, often stabilized by disulfide bridges, or they are studied under special experimental conditions. One possible reason for these differences is that structures in solution are generally more flexible and less stable than those in a crystal. This is suggested in the present analysis by the fact that for NMR structures, a larger threshold contact distance and a lower effective temperature than for crystallographic structures were needed.

Other Definitions of Contact

In this section, we discuss other possible definitions of contact. We first studied the effect of varying the threshold distance R_c. With the definition of contact based on all

heavy atoms, we analyzed $R_c = 4.0$ Å and $R_c = 5.0$ Å. The recognition for $R_c = 4.0$ Å was worse than for $R_c = 4.5$ Å: we obtained $\bar{Q} = 0.76$ and $\bar{Q} = 0.85$, respectively, for a test set including all crystallographic structures. This is consistent with the fact that the performances of the energy function for NMR structures are worse when a threshold $R_c = 4.5$ Å instead of $R_c = 5.0$ Å is used. Interestingly, the asymptotic number of contacts for NMR structures using $R_c = 4.5$ Å and $R_c = 5.0$ Å is the same as that for X-ray structures using $R_c = 4.0$ Å and $R_c = 4.5$ Å, respectively. With the threshold distance $R_c = 5.0$ Å, we found the average overlap $\bar{Q} = 0.87$, slightly larger than for $R_c = 4.5$ Å. However, for a larger number of chains the ground state and the native state do not coincide. Moreover, the chains that were problematic with $R_c = 4.5$ Å remain problematic in this case. We, thus, conclude that the optimal threshold distance for the all atoms definition of contacts, at least within the present optimization procedure, lies between 4.5 and 5.0 Å (see also ref. 9).

Next, we tried a definition of contact based on the distance between C_α atoms. The results were generally worse than with the contact definition based on all atoms. We found the best recognition for $R_c = 11$ Å (see also ref. 9). In this case, the optimized energy function recognized 88% of monomeric native structures, but only 42% of the oligomeric proteins.

Comparison With Other Energy Functions

Several sets of contact energy parameters have been published.[1] We chose for comparison three well-known parameter sets: Miyazawa-Jernigan (MJ),[2] Skolnick et al. (SJ),[4] and Thomas and Dill (TD).[5] We compared these parameter sets with $U^{(3)}$ by computing their coefficient of correlation r, following Betancourt and Thirumalai.[44] MJ and SJ adopted a definition of contact very similar to ours while TD used a definition of contact based on the distance between C_β atoms with an amino acid–dependent R_c. We excluded cysteine-cysteine interactions from the computation of r, since it has a different physical origin and it is much stronger than the others. We found the following correlation coefficients: $r = 0.74$ with MJ, $r = 0.83$ with SJ, and $r = 0.70$ with TD. For comparison, we recall that Betancourt and Thirumalai measured a coefficient of correlation $r = 0.82$ between MJ and SJ.[44] This high correlation is obtained despite that the present method to derive parameters is completely different from the ones used by these authors.

We then applied the MJ, SJ, and TD parameters to the training set C. The best results for recognition were obtained for TD interaction set, with only 139 errors in recognition of 921 X-ray structures, followed by the SJ interaction set with 163 errors and by the MJ set with 301 errors. It is noteworthy that the TD energy function performed better than the other ones although it was based on a different definition of contact with respect to the one that we used for computing the contact maps. For comparison, the present energy function makes 92 errors in recognition of X-ray structures. This comparison should not be taken too seriously, because most failures in recognition are due to the neglect of interchain and cofactor interactions, and because the test is not blind, being performed on chains that have been used in the training set. However, these comparisons indicate two interesting facts: first, all chains that are problematic with the present energy function are also problematic with at least two, and mostly all three, other energy functions. Second, all the energy functions examined reproduce the same ranking in stability: monomeric proteins are most stable, oligomeric proteins are of intermediate stability, and NMR structures are the least stable.

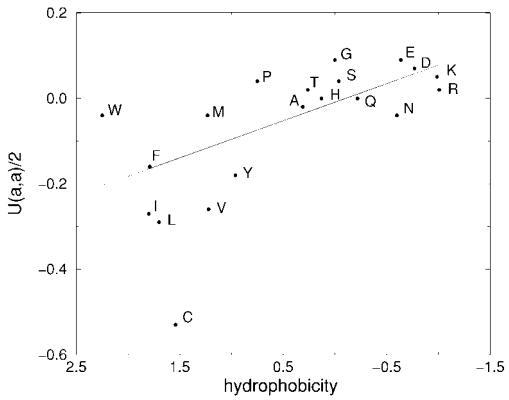

Fig. 12. Correlation between hydrophobicities and interaction energies.

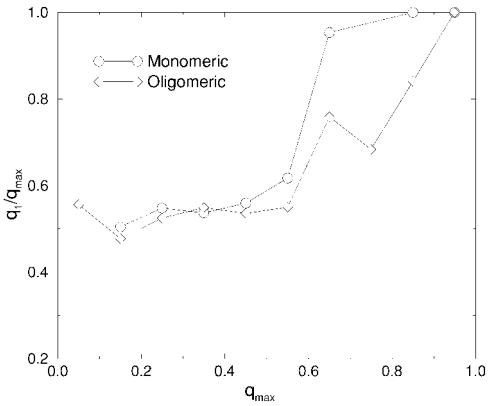

Fig. 13. Dependency of the quality of the prediction, q_1/q_{max}, on q_{max}.

The present energy parameters are also correlated to hydrophobicity, if cysteine is excluded. The existence of a correlation between the self-interaction elements $U(a, a)$ of knowledge-based potentials and hydrophobicity was first noted by Godzik et al.[43] and, independently, by Li et al.[46]; successively, it was studied by Betancourt and Thirumalai[44] and recently reviewed by Chan.[45] The set of hydrophobicities $h(a)$ measured by Fauchere and Pliska[47] has a coefficient of correlation $r = 0.72$ with the self-interactions $U(a, a)$ (Fig. 12) and $r = 0.77$ with the average interactions $1/20\ \Sigma_b\ U(a, b)$. Interestingly, Betancourt and Thirumalai found a significantly higher coefficient of correlation, $r = 0.92$, between the diagonal elements $B(a, a)$ of a rescaled MJ interaction matrix and the same measure of hydrophobicity.

Structure Prediction

For every chain for which the native state was recognized, we considered a "fictitious" kind of structure prediction by looking for the structure of the second lowest energy after the native state. To measure the quality of such prediction, we introduce the overlap q_1 between this structure and the native state and the maximum value of the native overlap for alll alternative structures, $q_{\max}$. A prediction is successful if the predicted structure is the one with the highest similarity, i.e., if $q_1 = q_{\max}$.

We found that the value of $q_1/q_{\max}$, which is always less than or equal to unity, depends critically on $q_{\max}$ itself (see Fig. 13). For $q_{\max} > 0.5$, with very high probability the structure of highest similarity has also the lowest energy and the recognition is successful. Instead, for $q_{\max} < 0.5$ the value of $q_1/q_{\max}$ is 0.5, close to the threshold of randomness. This result is reminiscent of the findings of Mirny and Shakhnovich.[48]

In order to explain such a result, we consider the subset $\Omega(q)$ of structures at overlap q with the native structure. The number of such structures, their average energy, and the variance of the energy are all decreasing functions of q. The minimum energy for the structures in $\Omega(q)$ depends on all these quantities, which have competing effects. The decrease of the average energy as q increases tends to decrease the minimum energy. In contrast, the decrease of the variance and the number of structures tends to increase it. In the present case, we compare a few structures with $q \approx q_{\max}$ with a bulk of structures with typical overlap $\bar{q} < q_{\max}$. In this way, $q_{\max}$ has to satisfy an inequality so that the energy of the structure with $q = q_{\max}$ is lower than the minimal energy of the set of structures with typical overlap $\bar{q}$. The better correlated the energy landscape is, the easier it is to fulfill this condition.

CONCLUSIONS

In this work, we analyzed in detail the optimization method that we recently proposed to derive energy parameters for protein fold recognition.[17]

The method is based on the maximization of the thermodynamic average of the overlap between the native structure and an ensemble of alternative conformations simultaneously for all proteins in a suitable training set. By using this procedure, energy parameters are optimized to give unfavorable energy to structures very different from the native one. More precisely, for each protein in the training set, low energy values are assigned to alternative structures whose overlap with the native state is large, whereas high energy values are assigned to structures with low overlap. Previous studies suggested that a good optimization procedure for deriving energy parameters should include (1) a way to assign to the native state a very low energy with respect to the average energy of alternative conformations, measured in terms of the width of the energy distribution,[10–14] (2) a way to discriminate the native structure against individual alternative conformations,[6–8,16] and (3) a way to identify "troublemakers," namely proteins that are impossible to stabilize and that tend to be present in large, automatically read databases. Such features are built in the method discussed here. This method produces protein models with correlated energy landscapes, a condition favoring thermodynamic stability and fast folding and crucial for structure prediction.

To generate alternative conformations, we used gapless threading, which has the advantage of being extremely efficient and the disadvantage of being a poor strategy for the generation of low-energy decoys. In particular, since insertions and deletions are common during evolution, all native structures that are very similar to a target structure also have a high degree of sequence homology, and methods of structure prediction based on threading without gaps give in most cases the same result as a simpler method based on sequence homology.

In order to generate better alternative conformations, one possibility is to mimic evolution, by building candidate structures using the database of existing structures as in threading, but allowing gaps in the structure. This method has been used very successfully for structural comparison.[39,48]

Another possibility is to use off-lattice Monte Carlo simulations. In this case, it has been shown that by using the pairwise contact approximation of the energy, even the correct recognition of the native state for a single protein is impossible.[8] This result is not at odds with the present finding that a pairwise contact energy function is able to stabilize the native state of most monomeric proteins with respect to decoys generated by threading. In fact, in case of threading, the energy function does not have to take into account steric repulsion, hydrogen bonds, secondary structure, dihedral angles preferences, and so on, since all these features are automatically fulfilled in native protein structures. Nevertheless crambin, the protein studied in Vendruscolo and Domany,[8] is a rather peculiar protein, since it is strongly hydrophobic. It is possible that the negative result found in Vendruscolo and Domany,[8] is in part due to this peculiarity of crambin. However, Vendruscolo and Domany also showed that a family of 6 immunoglobulines cannot be stabilized in this way.[33] Thus, it would be very interesting to find an effective energy function that can stabilize at least one target protein with respect to decoys generated by Monte Carlo, using a tentative energy function more complex than the simple sum of pairwise contact terms and a more detailed protein model (see, e.g., refs. 13, 49).

The energy function derived in this work stabilized the native states determined by X-ray crystallography for most representative proteins in the Protein Data Bank, by assigning to them a lower energy than for alternative structures generated by gapless threading. Exceptions to this behavior were nearly always structures where interactions with neighboring chains or with cofactors were neglected, as also discussed by Maiorov and Crippen.[7] Despite the limitations of gapless threading, we found that the energy difference needed to stabilize the native state in the set of alternative conformations is significantly correlated with this neglected interaction energy. This suggests that alternative structures generated by gapless threading might be sufficient for the study of thermodynamic stabil-

ity, and that the overlap method is effective in identifying chains for which the neglected interactions play a dominant role. Crystal structures of small fragments are also problematic to recognize. We suggest that this is due to neglect of the stabilizing effect of interactions with neighboring crystal cells. The other failures in recognition involve three structures with covalently bound cofactors and one structure that appears to have been incorrectly determined: thus it seems that the energy function is able to identify some errors in PDB structures.

The present energy function selects the right pairing between sequence and structure when the following conditions hold: (1) alternative structures are generated by gapless threading, (2) native structures are obtained by X-ray crystallography, and (3) all relevant pairwise contact interactions (including those with cofactors and between different chains) are taken into account. In Vendruscolo et al.,[9] the last two conditions were not considered, hence the difference from the present results. In future work, it will be important to investigate whether our results also hold for decoys obtained by gapped threading.

The recognition of NMR structures, on the other hand, gives worse results. We suggested four complementary explanations: (1) Solution structures, as determined from NMR data, are slightly different from structures in a crystal. For example, the average number of contacts is slightly lower in the former case; (2) Since effective interactions are obtained by averaging out the degrees of freedom of the environment, different parameters might be suitable for describing solution and crystal structures; (3) NMR structures are usually provided as a list of conformers. It is not clear which is the best way to use such conformers when deriving energy parameters; for example, all such conformers can be used self-consistently as a kind of "native ensemble"; (4) NMR spectroscopy is often used to study small and peculiar proteins, which are more likely to be stabilized by effective energies different from those of typical globular proteins.

We finally suggest that the overlap method can also be applied to sequence design: i.e., for a given energy function, one can determine a sequence $\mathbf{S}$ such that the target structure $\mathbf{C}^*$ is its thermodynamically stable ground state by maximizing the average overlap between $\mathbf{C}^*$ and the Boltzmann ensemble of alternative conformations.

ACKNOWLEDGMENTS

M.V. is supported by an EMBO long-term fellowship. U.B. is grateful to Bill Eaton for interesting discussion.

REFERENCES

1. Hao MH, Scheraga HA. Designing potential energy functions for protein folding. Curr Opin Struct Biol 1999;9:184–188.
2. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
3. Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casari G, Sippl M. 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 1990;216:167–180.
4. Skolnick J, Jaroszweski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. When is the quasi-chemical approximation correct? Protein Sci 1997;6:676–688.
5. Thomas PD, Dill KA. An iterative method for extracting energy-like quantities from protein structures. Proc Natl Acad Sci USA 1996;93:11628–11633.
6. Maiorov V, Crippen G. Contact potential that recognizes the correct folding of globular proteins. J Mol Biol 1992;227:876–888.
7. Maiorov VN, Crippen GM. Learning about protein folding via potential functions. Proteins 1994;20:167–173.
8. Vendruscolo M, Domany E. Pairwise contact potentials are unsuitable for protein folding. J Chem Phys 1998;109:11101–11108.
9. Vendruscolo M, Najmanovich R, Domany E. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? Proteins 2000;38:134–148.
10. Goldstein R, Luthey-Schulten ZA, Wolynes PG. Optimal protein-folding codes from spin-glass theory. Proc Natl Acad Sci USA 1992;89:4918–4922.
11. Goldstein R, Luthey-Schulten ZA, Wolynes PG. Protein tertiary structure recognition using optimized Hamiltonians with local interactions. Proc Natl Acad Sci USA 1992;89:9029–9033.
12. Koretke KK, Luthey-Shulten ZA, Wolynes PG. Self-consistently optimized energy functions for protein structure prediction by molecular dynamics. Proc Natl Acad Sci USA 1998;89:2932–2937.
13. Hao MH, Scheraga HA. How optimization of potential function affects protein folding. Proc Natl Acad Sci USA 1996;93:4984–4989.
14. Mirny LA, Shakhnovich EI. How to derive a protein folding potential? A new approach to an old problem. J Mol Biol 1996;264:1164–1179.
15. van Mourik J, Clementi C, Maritan A, Seno F, Banavar JR. Determination of interaction potentials of amino acids from native protein structures: Tests on simple lattice models. J Chem Phys 1999;110:10123–10133.
16. Dima R, Settanni G, Micheletti C, Banavar JR, Maritan A. Extraction of interaction potentials between amino acids from native protein structures. J Chem Phys 2000;112:9151–9166.
17. Bastolla U, Vendruscolo M, Knapp EW. A statistical mechanical method to optimize energy functions for protein folding. Proc Natl Acad Sci USA 2000;97:3977–3981.
18. Sheinerman FB, Norel R, Honig B. Electrostatic aspects of protein-protein interactions. Curr Opin Struct Biol 2000;10:153–159.
19. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–230.
20. Bowie D, Luthy JU, Eisenberg D. A method to identify protein sequences that fold into a known 3-dimensional structure. Science 1991;253:164–170.
21. Lazaridis T, Karplus M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. J Mol Biol 1998;288:477–487.
22. Bringelson JD, Wolynes PG. Spin-glasses and the statistical-mechanics of protein folding. Proc Natl Acad Sci USA 1987;84:7524–7528.
23. Ellis RJ. Discovery of molecular chaperones. Cell Stress Chaperones 1996;1:155–160.
24. Dinner AR, Abkevich V, Shakhnovich EI, Karplus M. Factors that affect the folding ability of proteins. Proteins 1999;35:34–40.
25. Klimov DK, Thirumalai D. Factors governing the foldability of proteins. Proteins 1996;26:411–441.
26. Gutin AM, Abkevich VI, Shakhnovich EI. Proc Natl Acad Sci USA 1995;92:1282–1286.
27. Abkevich VI, Gutin AM, Shakhnovich EI. Free energy landscapes for protein folding kinetics: intermediates, traps and multiple pathways in theory and lattice model simulations. J Chem Phys 1994;101:6052–6062.
28. Tiana G, Broglia RA, Roman HE, Vigezzi E, Shakhnovich EI. Folding and misfolding of designed proteinlike chains with mutations. J Chem Phys 1998;108:757–761.
29. Bastolla U, Roman HE, Vendruscolo M. Neutral evolution of model proteins: Diffusion in sequence space and overdispersion. J Theor Biol 1999;200:49–64.
30. Grassberger P. Prune-enriched Rosunbluth method: simulations of theta polymers of chain length up to 1,000,000. Phys Rev E 1996;56:3682–3693.
31. Frauenkron H, Bastolla U, Gerstner E, Grassberger P, Nadler W. New Monte Carlo algorithm for protein folding. Phys Rev Lett 1998;80:3149–3152.

32. Bastolla U, Frauenkron H, Gerstner E, Grassberger P, Nadler W. Testing a new Monte Carlo algorithm for protein folding. Proteins 1998;32:52–66.
33. Vendruscolo M, Domany E. Protein folding using contact maps. Vitam Horm 2000;58:171–212.
34. Vendruscolo M, Mirny LA, Shakhnovich EI, Domany L. Comparison of two optimization methods to derive energy parameters for protein folding: perceptron and *Z* score. Proteins 2000;41:192–201.
35. Xia Y, Levitt M. Extracting knowledge-based energy functions from protein structures by error rate minimization: comparison of methods using lattice models. J Chem Phys 2000;113:9318–9330.
36. Maiorov VN, Crippen GM. Size-independent comparison of protein three-dimensional structures. Proteins 1995;22:273–283.
37. Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. Nature 1996;381:272. The program WHAT-CHECK is available on line at http://www.cmbi.kun.nl/gv/pdbreport/
38. Hobohm U, Sander C. Enlarged representative set of protein structure. Protein Sci 1994;3:522–524.
39. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
40. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–994.
41. Bastolla U, Grassberger P. Exactness of the annealed and the replica symmetric approximation for random heteropolymers. Phys Rev E 2001;63:031901.
42. Micheletti C, Banavar JR, Maritan A, Seno F. Protein structures and optimal folding from a geometrical variational principle. Phys Rev Lett 1999;82:3372–3375.
43. Godzik A, Kolinski A, Skolnick J. Are proteins ideal mixtures of amino-acids: analysis of energy parameter sets. Protein Sci 1995; 4:2107–2117.
44. Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci 1998;8:361–369.
45. Chan HS. Folding alphabets. Nature Struct Biol 1999;6:994–996.
46. Li H, Tang C, Wingreem NS. Nature of driving force for protein folding: a result from analyzing the statistical potential. Phys Rev Lett 1997;79:765–768.
47. Fauchere JL, Pliska V. Hydrophobic parameters-pi of amino acid side chain from the partitioning N-acetyl amino acid amides. Eur J Med Chem 1983;18:369–375.
48. Mirny L, Shakhnovich EI. Protein structure prediction by threading. Why it works and why it does not. J Mol Biol 1998;283:507–526.
49. Micheletti C, Seno F, Banavar JR, Maritan A. Learning effective amino acid interactions through iteractive stochastic techniques. Proteins 2001;42:422–431.