PROTEINS: Structure, Function, and Genetics 26:391–410 (1996)

Protein Fold Recognition and Dynamics in the Space of Contact Maps

Leonid Mirny[1] **and Eytan Domany**[2*]

[1]Departments of Structural Biology and [2]Physics of Complex Systems, Weizmann Institute of Science, Rehovot, Israel

ABSTRACT **We introduce an energy function for contact maps of proteins. In addition to the standard term, that takes into account pairwise interactions between amino acids, our potential contains a new hydrophobic energy term. Parameters of the energy function were obtained from a statistical analysis of the contact maps of known structures. The quality of our energy function was tested extensively in a variety of ways. In particular, fold recognition experiments revealed that for a fixed sequence the native map is identified correctly in an overwhelming majority of the cases tested. We succeeded in identifying the structure of some proteins that are known to pose difficulties for such tests (BPTI, spectrin, and cro-protein). In addition, many known pairs of homologous structures were correctly identified, even when the two sequences had relatively low sequence homology. We also introduced a dynamic Monte Carlo procedure in the space of contact maps, taking topological and polymeric constraints into account by restrictive dynamic rules. Various aspects of protein dynamics, including high-temperature melting and refolding, were simulated. Perspectives of application of the energy function and the method for structure checking and fold prediction are discussed. Proteins 26:391–410**

Key words: **contact maps; Monte Carlo dynamics; protein structure prediction; contact map energy function; sequence specificity**

INTRODUCTION

Under physiological conditions native enzymatic proteins have unique compact and functionally active conformations. The 3D structure of the protein is thought to be determined in a unique fashion by its amino acid sequence. A central problem of protein research[1–5] is to identify the rules that determine how the amino acid sequence directs its own folding process. This "folding code" is nonlocal, that is, the conformation of a particular amino acid depends on the identity and conformation of amino acids located distantly along the protein chain. Progress in x-ray crystallography and NMR spectroscopy yielded well-solved structures for about 300 different proteins[6] since the construction of the first three-dimensional model of myoglobin.[7,8] Serving as a source of data for statistical analysis, these structures can be used to obtain information about folding motifs, energetics, and other aspects of the folding code.

There are two main approaches to determination of structure from sequence. The first[9–11] assumes that a protein's native state corresponds to the global minimum of its (free) energy and attempts to find it. An alternative hypothesis[12–14] is that the unique native structure of a protein may correspond to a local minimum of the free energy, which, however, is easily reachable from the unfolded state. This approach requires detailed simulation of protein dynamics. The two scenarios are not contradictory: evolution may have selected those sequences that fold rapidly into their global free-energy minimum.[10,15] Three major choices must be made for both approaches: that of the representation of the protein's structure, the energy function to be used and of the method one uses to sample protein conformations.

Several representations of protein structure have been suggested, which vary in complexity and accuracy, all the way from specifying all atomic coordinates[16] to very simple off-[17,18] and on-lattice[15,19–21] models of the protein chain in 3 or 2 dimensions.[22] Working with highly detailed representations requires allocation of substantial computational resources to sample even very limited regions of conformational space.[16] On the other hand, using a simple representation one can extensively sample

This article is based on a thesis submitted by L.M. to the Weizmann Institute of Science in partial fulfillment of the requirements for the M.Sc. degree.

Dr. Mirny's current address is Department of Chemistry, Harvard University, Cambridge, MA.

*Correspondence to: Dr. Eytan Domany, Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel.

Received 3 June 1996; accepted 18 June 1996.

large regions of conformational space[23,24] and even perform complete enumeration of all conformations[15], but at the cost of providing the structure with low accuracy.[25]

Energy functions that were introduced for protein folding take into account different interactions involved in stabilizing a protein structure (reviewed in Wodak and Roman[26]), such as local conformational preferences of amino acids[27], pairwise amino acid interactions[15,18,28,29], hydrophobic interactions of the amino acids with the solvent[30], interactions of an amino acid with its polar and charged environment.[31] The form of the energy function might vary from a simple contact hydrophobic-hydrophilic potential[22] to very complicated atom-atom distance dependent potentials.[16,32]

The sampling techniques that are being used also cover a wide range, varying from slow molecular dynamic simulations[16] to fast Monte Carlo and simulated annealing methods, aimed at locating a global energy minimum in conformational space.[15,17]

In this work we chose a representation of intermediate complexity to provide a satisfactory description of a protein's structure, for which we can, however, define a simple energy function and, at the same time, use a fast sampling technique.

We chose to represent the structure of a protein of N amino acids by a two dimensional contact map.[33] Each (i,j) cell of this $N \times N$ matrix contains binary values: $S_{ij} = 1$ if amino acids i and j are in contact and $S_{ij} = 0$ otherwise. Amino acids i and j are in contact when the distance between them is less than some threshold. The value of the threshold is taken to be equal to some effective radius of amino acid interaction. Here i and j denote the position of amino acids along the polypeptide chain of length N. The principal difference of this type of structure representation from all the others described above is that it presents the spatial interactions between sequentially distant amino acids, rather than conformation of the polypeptide chain. In contrast to Cartesian coordinates, the map representation of protein structure is independent of the coordinate frame. This property made map representation attractive for protein structure comparisons and for search for similar structures in a database.[29,34–36]

For this reduced representation of protein structure we introduce an energy function, that states the value of the "energy" for any contact map and any amino acid sequence of a given length. This function does not yield the proper energy, obtained by taking all molecular interactions into account. Rather, we use the logarithm of the probability of occurrence of a given contact map as an estimate of (minus) the "energy" associated with it. Therefore, even though we refer to this value as "energy", it really is an approximation of the free energy of a contact map. We explain below what is meant by this concept. The "energy" associated with a contact map consists of three components; a previously used term, which represents residue-residue interactions[20], a *new hydrophobic* term, representing interactions of the residue with the solvent, and a constraint term, which excludes contact maps that correspond to nonphysical structures. The parameters of the first two terms in our energy function were determined by using our definition of the energy, which is tantamount to stating that the probability of observing a particular contact map for a given sequence satisfies, with this energy, Boltzmann's law for thermodynamic systems in equilibrium.[32] In the framework of this statistical approach we computed the values of various parameters of our energy function. For example, in order to calculate the contact energy between two amino acids types, we found the frequency of these contacts in all native proteins of our data set, normalized these frequencies, and translated them into energy values.

An important part of our work is testing and investigation of the energy function thus obtained. To prove applicability of the introduced energy function for the task of protein folding, we demonstrated its sequence specificity by various tests (we thank M. Levitt for suggesting some of these). For a given native protein structure the contact map was built, and then energy values were calculated for this native map substituting in it sequences obtained by random scrambling of the native sequence. We found that the true native sequence, corresponding to our map, has an energy which lies five standard deviations below the average energy of the scrambled sequences.

Another test and possible application of our model is fold recognition. For a given fixed sequence we generated a large set of possible contact maps, (obtained from the bank of known structures) and calculated the energy that corresponds to adopting each structure in the dataset by the sequence. In such a screening test we expect that the native map of the sequence gets the lowest energy.

We performed this test for 249 sequences of proteins with known structure. For every sequence a set of about 10,000 candidate contact maps was generated. In 238 cases indeed the native map was assigned the lowest energy. We also performed various studies of the distribution of the energy values obtained this way. In particular, the roles played in the fold recognition task by the two terms of our energy function were investigated in detail. We found that the pairwise contact energy is better at identifying the native map as the one of lowest energy. The hydrophobic part of the energy is, on the other hand, much more strongly correlated with the extent to which a map resembles the native one. Such correlations are of central importance in fold prediction, in that they are essential to select known maps that are close in some sense to that of a sequence whose structure one is trying to predict.

To test further the ability of our energy function in this respect we considered in detail the 20 maps that were assigned the lowest (next to the native) values of our energy function. The parent structures from which we extracted these maps were identified, pulled out of the database and compared with the native structure. These comparisons revealed in a large number of cases that our energy function indeed identified pairs of proteins that are known to have high sequence similarity (and low sequence homology).

In addition to the tests described above we also performed detailed comparisons of our energy function's performance with that of other related available potentials. In particular, we considered in detail a few cases that are known to constitute problems for fold recognition by various potentials, such as BPTI,[37] cro-protein,[38] and spectrin.

These tests are appropriate for the energy associated with residue-residue and hydrophobic interactions, which are is easy to evaluate for a given map and amino acid sequence. The constraint term, on the other hand, cannot be evaluated directly. It was taken into account in an indirect way, through introduction of restrictive dynamic rules. When a Monte Carlo-type dynamic procedure is started from a physically realizable contact map, and only changes that conform with these rules are allowed, we believe that the new map is also physically realizable. In order to identify dynamic rules that ensure this, we performed Monte-Carlo simulations for a fixed amino acid sequence, making dynamic steps in the space of contact maps. Starting from the native state we minimized the energy, allowing only changes that do not increase it. We formulated dynamic rules that excluded appearance of nonphysical structures in the course of this procedure.

These dynamic rules were then used to simulate protein melting. Starting from the native contact map, a limited number of Monte Carlo steps were performed at a high temperature. We observed that β sheets unfold before α helices. Next, we studied relaxation from a partially melted state, and found that the protein returned to conformations very similar to the native ones.

Finally, we combined screening as described above with dynamics, using a map obtained from screening a large set of candidate maps as the starting point of our dynamic procedure. In a few cases we were able to show improvement of our results as a consequence of the dynamics. Such was the case, for example, of the plastocyanin sequence, for which the combination of our screening method with Monte-Carlo dynamics demonstrated that the structure of pseudoazurin (known[39] to be close to plastocyanin) has the nearest energy to the native one.

We conclude the paper by summarizing what we have done, and discussing possible extensions and perspectives for future applications of our work.

CONTACT MAP REPRESENTATION OF PROTEIN STRUCTURE AND THE ASSOCIATED ENERGY FUNCTION

The Contact Map

The contact map is a 2D representation of protein structure. For a protein of N_p amino acids the contact map is an $N_p \times N_p$ matrix S whose elements are given, for all $i,j = 1, 2, \ldots, N_p$, by

$$S_{ij} = \begin{cases} 1 & \text{if there is contact between amino acids } i \text{ and } j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Two amino acids, u and v, are in *contact* if their separation D_{uv} satisfies

$$D_{uv} = \min_{l,m} \rho(R_l^u, R_m^v) < 4.5 \text{ Å} \tag{2}$$

where is the minimal distance between heavy atoms belonging to the side chain or the backbone of amino acids u and v, with corresponding coordinates R_l^u and R_m^v. The threshold value of 4.5 Å, taken from the literature,[20] guarantees that there is neither a water molecule nor a third residue between two amino acids in contact. Hereby we made an assumption that there is no difference between side chain and backbone contacts.

The contact map serves not only as a representation of structure, but also as an energy fingerprint of the protein conformation. The contact map summarizes the amino acid interactions in the structure. As far as pair interactions give the dominant contribution to protein stability and sequence specificity, the energy derived from the contact map can be a good approximation of the real protein energy. Using this reduced description of protein structure, we can calculate the approximate energy of any conformation, skipping or averaging over less important contributions to the protein energy (such as main chain bending, side chain orientational energy, etc.). Contact maps are very demonstrative representations of protein structure and, therefore, very attractive for visual analysis. For example, one can see all secondary structure elements on the map. A helix contains contacts between pairs with indices $i,i \pm 4$ and $i,i \pm 3$, and is represented as lines of contacts running parallel to the main diagonal and just near it (Fig. 1). Antiparallel (and parallel) β strands give rise to sets of $i + k, j - k; k = 0,1, \ldots$ contacts (or $i + k, j + k; k = 0,1, \ldots$ contacts), respectively. Consequently β sheets appear as straight lines of contacts that are perpendicular (or parallel) to the main diagonal of the contact map (Fig 1).

It is possible to reproduce, with considerable accuracy, the three-dimensional structure of the protein's backbone from its contact map. This fact has been demonstrated in the early work of Havel and colleagues,[40] who used methods of distance geometry developed by Crippen.[41] For more recent work that uses geometrical constraints and distance ine-

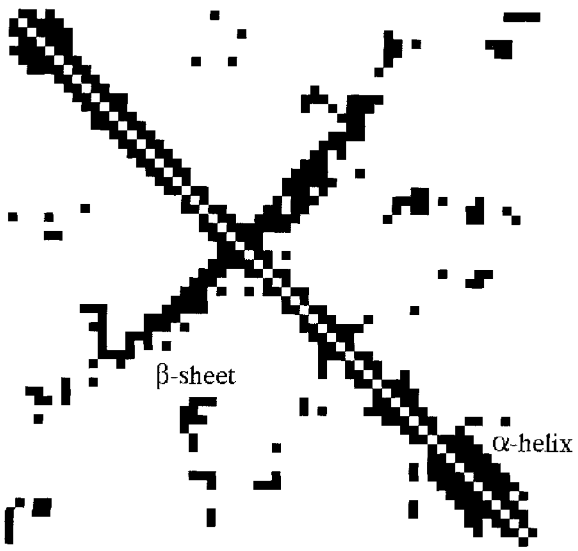

Fig. 1. Contact maps of BPTI. The half below the diagonal is the map of the native structure. The part above the diagonal was obtained by a constrained dynamic procedure. Note the secondary structure elements (α helices and β sheets), represented by typical patterns on the contact map.

qualities to reconstruct the backbone, see Saitoh and colleagues[42] and Bohr and colleagues.[43] An efficient and direct Monte-Carlo based evaluation of backbone conformation from contact maps will be published in the future (E. Kussel and E. Domany, unpublished observations).

Energy Function

We turn now to define a function H[S,**a**] that yields the value of the "energy" for a given amino acid sequence **a** = $(a_1, a_2, \ldots a_{N_p}$ that forms a structure whose contact map is S. Our energy function consists of three terms:

$$H = H^{pair}[\mathbf{S},\mathbf{a}] + H^{hydrophobic}[\mathbf{S},\mathbf{a}] + H^{constraints}[\mathbf{S}]. \quad (3)$$

The first term of our energy function contains pairwise residue-residue interactions.[20,21] Since we represent interactions in a discrete manner; that is, two amino acids are either in contact or with no contact between them, the total pair energy equals to the sum over energies of the existing contacts:

$$H^{pair}[\mathbf{S},\mathbf{a}] = \sum_{\substack{i,j=1\\ i<j}}^{N_p} s_{ij}E_{a_i,a_j}, \quad (4)$$

where $E_{u,v}$ is the energy of interaction between an amino acid of type u and one of type v. The parameters $E_{u,v}$ constitute a 20 × 20 matrix, which is computed (see below) from the frequency of different amino acid contact appearances in the contact maps of native proteins, obtained from the Protein Data Bank.[6]

The next term in our energy function serves two purposes: to take into account excluded volume and interactions of the various amino acids with water. Since amino acids are rather bulky groups, the possible conformations of a protein are restricted by the excluded volume interactions between them. Particularly, the number of contacts that amino acid u can possibly have in a structure cannot be too large, since it is not possible to pack too many amino acids into the volume which lies at a sufficiently small distance from u. Consequently, the number of contacts for each amino acid in a structure is constrained by excluded volume interactions, irrespective of the energy of these contacts. In addition, a hydrophobic amino acid will prefer to have a large number of contacts, even if they are not most favorable energetically, in order to avoid water molecules. In order to take into account these physical effects we introduce in the energy function the second term:

$$H^{hydrophobic}[\mathbf{S},\mathbf{a}] = \sum_{i=1}^{N_p} \left[\beta_{a_i} \left(n_{a_i} - \sum_{k=1}^{N_p} S_{ik} \right)^2 \right] \quad (5)$$

The parameters β_u and n_u depend on the identity of amino acid u. Clearly, $\Sigma^{N_p}{}_{k=1}S_{ik}$ is the number of contacts that the amino acid of index i has on a given contact map. Hydrophobic amino acids will have large values of n_u. The contribution of any amino acid u to $H^{hydrophobic}$ is minimal when the number of its contacts equals the optimal number of contacts, n_u (which depends on the amino acid type). The values of the parameters β_u and n_u were derived from a statistical analysis of native protein structures.

The third term of the energy function ensures that a contact map S satisfies certain constraints that originate from the polymeric nature of a protein, that is, topological and geometrical restrictions on possible protein conformations and, consequently, on possible contact maps. For example, one constraint follows from the sequential connection of amino acids in the polypeptide chain, which limits the chain's possible conformations. This and other constraints are discussed in detail below. These constraints could be expressed formally in the energy function by assigning infinite energy to all contact maps that violate them:

$$H^{constraints}[\mathbf{S}] = \begin{cases} 0 & \text{if map S corresponds to} \\ & \quad \text{physically possible structure} \\ \infty & \text{otherwise} \end{cases} \quad (6)$$

Since it is rather difficult to write down explicitly such a function, we approach the problem of constraints by introducing restrictive rules in our dynamic procedure. Adhering to these rules ensures that our dynamic steps do not leave the space of "realistic" contact maps. The dynamic rules are formulated and discussed below.

PARAMETER ESTIMATION FROM STATISTICAL ANALYSIS

We introduce now the statistical approach used to derive the energy function for a given amino acid sequence and any associated contact map. It is assumed, following Sippl,[32] that the probabilities of protein conformations $\mathfrak{R}_n$ are dictated by the Boltzmann distribution,

$$P(\mathfrak{R}_n(\mathbf{a})) = \frac{1}{Z} exp\left[\frac{-\varepsilon(\mathfrak{R}_n(\mathbf{a}))}{\kappa T}\right],\ Z = \sum_k exp\left[\frac{-\varepsilon(\mathfrak{R}_k(\mathbf{a}))}{\kappa T}\right] \quad (7)$$

Here $\mathfrak{R}_n(\mathbf{a})$ represents a fully specified conformation of a protein with amino acid sequence $\mathbf{a}$ (i.e., the coordinates of all the atoms that constitute the protein). $\epsilon[\mathfrak{R}_n(\mathbf{a})]$ is the (microscopic) energy associated with this configuration. Given the configuration $\mathfrak{R}_n(\mathbf{a})$, it is easy to determine the corresponding contact map $\mathbf{S}[\mathfrak{R}_n(\mathbf{a})]$. Define now the "projection operator" for any configuration $\mathfrak{R}(\mathbf{a})$ and any contact map $\mathbf{S}$:

$$\Delta(\mathfrak{R}(\mathbf{a}), \mathbf{S}) = \begin{cases} 1 & \text{if } \mathbf{S} = \mathbf{S}(\mathfrak{R}(\mathbf{a})) \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

Since any configuration has one single associated contact map we clearly have

$$\sum_{\mathbf{S}} \Delta(\mathfrak{R}(\mathbf{a}), \mathbf{S}) = 1 \quad (9)$$

We can now calculate (formally) the probability of appearance of a contact map $\mathbf{S}$ for our protein, as the sum of the probabilities of all microscopic configurations, that give rise to contact map $\mathbf{S}$.

$$P(\mathbf{S},\mathbf{a}) = \sum_{\mathfrak{R}} P(\mathfrak{R}(\mathbf{a}))\Delta(\mathfrak{R}(\mathbf{a}), \mathbf{S}) \quad (10)$$

This complicated function of $\mathbf{S}$ can be written as

$$P(\mathbf{S},\mathbf{a}) = e^{-E(\mathbf{S},\mathbf{a})} \quad (11)$$

It is this effective (free) energy $E(\mathbf{S},\mathbf{a})$, which we approximate by the function $H(\mathbf{S},\mathbf{a})$ of Equation (3).

The basic premise used to estimate the interaction parameters E_{uv} is based on the assumption that the probability of forming a contact between amino acids u and v can be obtained from Equation (11), and a statistical analysis of the known protein contact maps. A formal derivation of our estimates for E_{uv} is given in the Appendix. Here we present only intuitive arguments.

Set the zero of energy as corresponding to an average or "neutral" contact. Compared to this, contacts formed by some amino acids are favorable, and we expect negative E_{uv} values for these. On the other hand,some contacts could be unfavorable, that is, the corresponding amino acids must be forced to form this contact. For these we expect to have positive values of E_{uv}.

In order to calculate the energy of a $u - v$ contact, one has to estimate correctly the probability of finding this contact in the set of proteins whose structure is known. Our set consists of 54 protein structures from the Brookhaven Protein Data Bank.[6]

For each protein p in the set we calculated the probability of having a $u - v$ contact, w^p_{uv}. Let us denote the number of $u - v$ contacts in a given protein structure by N_{uv}, and the number of amino acids of type u in its sequence by N_u. We do not count contacts between amino acids i and j if they are neighbors along the polypeptide chain. The maximal number of $u - v$ contacts that can be formed by the sequence is M_{uv}. It is equal to $N_uN_v/2$ less the number of $u - v$ neighbor pairs along the polypeptide chain. We calculate the frequency of $u - v$ contacts as the ratio of their actual number and their maximal possible number:

$$F_{uv} = \frac{N_{uv}}{M_{uv}} \quad (12)$$

This frequency, obtained for each pair, has to be compared with the frequency of any spatial contact on the map, determined as the ratio of actually formed contacts N_o, to the number of all possible contacts of a protein of length N_p:

$$F^{total} = \frac{N_o}{N_p(N_p - 3)/2}, \quad (13)$$

Finally, the relative probability of $u - v$ contact formation equals the ratio of these frequencies:

$$w^p_{uv} = \frac{F_{uv}}{F^{total}} = \frac{N_{uv}}{N_o} \cdot \frac{N_p(N_p - 3)/2}{M_{uv}} \quad (14)$$

and we have shown that to a good approximation

$$E^p_{uv} = -log\ w^p_{uv} \quad (15)$$

To calculate this quantity for a set of proteins we first averaged the values of w^p_{uv} over the set.

$$E_{uv} = -log\langle w_{uv}\rangle \quad (16)$$

where $\langle\bullet\rangle$ denotes a weighted average over all proteins in the data set. Since the approximation used in the Appendix for the derivation of Equation (15) works better for longer proteins, we assign to each protein a weight proportional to N_p:

$$\langle w_{uv}\rangle = \frac{\sum_p N_p w^p_{uv}}{\sum_p N_p} \quad (17)$$

Substitution into the expression for the energy gives finally

$$E_{uv} = -log\left[\frac{\sum_p N_p w^p_{uv}}{\sum_p N_p}\right] \quad (18)$$

The values of obtained by the method described above are given* in Table I.

Turning now to evaluating the parameters and that enter the expression (5) for $H^{hydrophobic}$, we follow the same strategy as used above. For each amino acid type we calculated the frequency of the number of contacts in a set of native structures, and plotted them as histograms. Table II presents mean values and standard deviations for the number of contacts of different amino acids. The frequency distributions of the number of contacts can be approximated by gaussians, with different parameters for different amino acid types. If the distribution is approximated by

$$f_n^u \propto \exp\left[-\frac{(n-n_u)^2}{2\sigma_u^2}\right] \qquad (19)$$

then following the assumption of Boltzmann distribution, the energy of volume and hydrophobic interactions for each amino acid type is given by

$$E_u^{vol}(n) = -log\, f_n^u = \frac{(n_u - n)^2}{2\sigma_u^2} \qquad (20)$$

where n is the actual number of contacts this amino acid has in the given structure,

$$n = \sum_{k=1}^{N_p} S_{ik}. \qquad (21)$$

and the parameter appearing in Equation (5) is identified as

$$\beta_u = \frac{1}{2\sigma_u^2}$$

The variation of for different amino acids was found to be very small, $\sigma_u = 2.5 \pm 0.2$. The most probable number of contacts does vary significantly (see Table II), in the range from 6 to 11. The large variation of n_u values for different amino acids is due to the fact that hydrophobic amino acids are usually buried deep into the protein structure, forming hydrophobic cores of protein globules and, consequently, have many contacts, whereas hydrophilic amino acids prefer to be exposed to the solvent and have a small number of contacts with the other amino acids. We found that the hydrophobic index used by other authors[30] and the average number of contacts determined by us are correlated; the Pearlson correlation coefficient for n_u and the Casari-Sippl[30] hydrophobic index is $\rho = 0.77$.

TESTING H^{PAIR} AND H^{HYDRO}

We present now rather extensive tests that were performed to evaluate the quality of the energy function (Eqs. 3–5). Here we concentrate on the two terms H^{pair} and H^{hydro}, for which explicit forms were presented and the values of the parameters were determined as described above. We postpone discussion of the third term, $H^{constraint}$, which is much more difficult to express explicitly as a function of the sequence **a** and the map **S**, to a later section in this article. This limits the tests to be performed to those for which we are certain that the proposed map corresponds to a physically realizable structure, so that $H^{constraint} = 0$ for the map tested.

Sequence Specifity of the Energy Function

The simplest way to test the energy function H[**S**, **a**] is to pick a fixed legitimate contact map **S** and thread through it varying different sequences **a**. If we take a protein p of known native contact map $\mathbf{S}^p$ and substitute in this template map a large number of randomly assembled sequences, we expect these to produce higher values of the energy than that obtained when the true sequence $\mathbf{a}^p$ is used. In effect we took the native sequence and *randomly scrambled* its amino acids, generating this way a large number of random sequences, which we called the shuffled ensemble. Each of these random sequences $\mathbf{a}_k$ was "threaded" through the native contact map and its energy $H_k = H[\mathbf{S}^p, \mathbf{a}_k]$ was calculated using Equations (3–5). These energies were compared with that of the original unscrambled native sequence $\mathbf{a}^p$. For the BPTI protein we generated a shuffled ensemble of 1000 sequences. The average and standard deviation of the energies obtained when these sequences were threaded into the template native contact map of BPTI are

$$<\mathbf{E}> = 55.9 \quad [<(\mathbf{E} - <\mathbf{E}>)^2]^{1/2} = 10.3 \qquad (22)$$

The energy of the native sequence,

$$E^{native} = 4.2 \qquad (23)$$

lies about five standard deviations below the average. This large difference clearly indicates that our energy function is sequence specific. That is, the native sequence "fits" the native map much better than a randomly scrambled sequence of the same amino acids. This fact, however, does not imply that the native sequence has the *lowest* energy among all shuffled sequences. In fact, shuffled sequences of lower than native energy were found by the following Monte Carlo procedure. Starting with the native sequence, a pair of amino acids was picked randomly and permuted. The resulting new sequence was accepted if energy decreased as a result of the permutation. The process was stopped when no permutaion was accepted after 100 attempts. Using different randomly selected pairs we generated an ensemble

*Slightly modified values of these parameters were used for some of the calculations reported in this paper. These parameters, as well as additional information can be obtained by contacting mirny@husc.harvard.edu.

TABLE I. Contact Energies E_{uv} of amino acids u and v, as obtained [Using Equation (18)] From 54 Known Protein Structures

Amino																				
Acid	ALA	GLU	GLN	ASP	ASN	LEU	GLY	LYS	SER	VAL	ARG	THR	PRO	ILE	MET	PHE	TYR	CYS	TRP	HIS
ALA	0.175	0.206	0.150	0.466	0.242	-0.168	0.355	0.216	0.303	-0.151	0.154	0.272	0.479	-0.279	-0.162	-0.200	-0.156	0.350	-0.345	0.015
GLU	0.206	-0.173	0.153	0.336	-0.074	0.099	0.522	-0.175	0.286	0.188	-0.372	0.148	0.298	-0.097	-0.053	-0.184	0.085	0.445	-0.253	-0.320
GLN	0.150	0.153	-0.036	0.231	-0.128	0.034	0.291	0.124	0.234	-0.017	-0.349	0.087	-0.072	0.098	0.200	-0.197	-0.141	-0.389	-0.251	-0.130
ASP	0.366	0.336	0.231	0.285	0.120	0.217	0.358	-0.143	0.226	0.332	-0.442	0.184	0.539	0.157	0.080	0.036	-0.230	0.510	-0.149	-0.103
ASN	0.242	-0.074	-0.128	0.120	0.264	0.045	0.267	0.081	0.339	0.122	-0.164	0.216	0.466	0.111	-0.382	0.166	-0.333	0.039	0.039	-0.057
LEU	-0.168	0.099	0.034	0.217	0.045	-0.685	0.102	0.067	0.212	-0.591	-0.353	-0.020	0.093	-0.781	-0.393	-0.742	-0.529	-0.380	-0.654	-0.169
GLY	0.355	0.522	0.291	0.358	0.267	0.102	0.327	0.223	0.550	0.211	0.184	0.327	0.414	0.296	0.296	0.312	0.043	0.019	-0.173	0.275
LYS	0.216	-0.175	0.124	-0.143	0.081	0.067	0.223	0.397	0.253	-0.004	0.238	0.092	0.218	-0.040	-0.021	-0.052	-0.279	0.212	-0.185	0.187
SER	0.303	0.286	0.234	0.226	0.339	0.212	0.550	0.253	0.250	0.207	0.047	0.215	0.232	0.132	0.140	-0.021	0.034	0.119	-0.120	0.250
VAL	-0.151	0.188	-0.017	0.332	0.122	-0.591	0.211	-0.004	0.207	-0.451	0.019	0.033	0.280	-0.588	-0.326	-0.600	-0.261	-0.205	-0.598	0.054
ARG	0.154	-0.372	-0.349	-0.442	-0.164	-0.353	0.184	0.238	0.047	0.019	-0.017	-0.130	0.052	-0.433	-0.397	-0.466	-0.544	-0.223	-0.335	0.043
THR	0.272	0.148	0.087	0.184	0.216	-0.020	0.327	0.092	0.215	0.033	-0.130	0.044	0.254	-0.178	-0.044	-0.189	-0.131	0.110	-0.286	0.000
PRO	0.479	0.298	-0.072	0.539	0.466	0.093	0.414	0.218	0.232	0.280	0.052	0.254	0.341	0.403	-0.242	0.092	-0.289	0.216	-0.345	0.309
ILE	-0.279	-0.097	0.098	0.157	0.111	-0.781	0.296	-0.040	0.132	-0.588	-0.433	-0.178	0.403	-0.716	-0.727	-0.748	-0.547	0.042	-0.849	0.016
MET	-0.162	-0.053	0.200	0.080	-0.382	-0.393	0.296	-0.021	0.140	-0.326	-0.397	-0.044	-0.242	-0.727	-0.869	-0.664	-0.705	-0.632	-1.281	-0.563
PHE	-0.200	-0.184	-0.197	0.036	0.166	-0.742	0.312	-0.052	-0.021	-0.600	-0.466	-0.189	0.082	-0.748	-0.664	-0.953	-0.501	-0.650	-1.114	-0.367
TYR	-0.156	0.085	-0.141	-0.230	-0.333	-0.529	0.043	-0.279	0.034	-0.261	-0.544	-0.131	-0.289	-0.547	-0.705	-0.501	-0.258	-0.263	-0.619	-0.609
CYS	0.350	0.445	-0.389	0.510	0.039	-0.380	0.019	0.212	0.119	-0.205	-0.223	0.110	0.216	0.042	-0.632	-0.650	-0.263	-1.843	-0.147	-0.814
TRP	-0.345	-0.253	-0.251	-0.149	0.039	-0.654	-0.173	-0.185	-0.120	-0.598	-0.335	-0.286	-0.345	-0.849	-1.281	-1.114	-0.619	-0.147	-0.826	-0.682
HIS	0.015	-0.320	-0.130	-0.103	-0.057	-0.169	0.275	0.187	0.250	0.054	0.043	0.000	0.309	0.016	-0.563	-0.367	-0.609	-0.814	-0.682	-0.568

TABLE II. Mean (n_u) and Standard Deviation (σ_u) of the Number of Contacts of Amino Acid U, as Obtained From Analysis of 54 Known Protein Structures

Amino Acid	ALA	GLU	GLN	ASP	ASN	LEU	GLY	LYS	SER	VAL	ARG	THR	PRO	ILE	MET	PHE	TYR	CYS	TRP	HIS
n_u	6.7	6.4	7.0	6.1	6.4	9.3	5.6	6.5	5.8	8.7	8.0	6.7	5.8	9.5	9.4	10.4	9.7	8.1	11.4	8.0
σ_u	2.4	2.4	2.7	2.6	2.6	2.6	2.3	2.4	2.6	2.4	3.2	2.6	2.3	2.4	2.8	2.3	2.8	2.2	2.8	2.7

of "superstable" sequences.[23] Their existence may be due to the fact that our energy function is imperfect; even if these are truly of lower energy than the native sequence, this does not contradict at all our previous finding of sequence specifity. There is no reason to exclude the possibility that some permutations of the amino acids of a given native sequence can lower the energy of the corresponding native map.

Fold Recognition

A complementary test of our energy function is to substitute in it a fixed sequence a^p (of a protein of known structure), and vary the contact maps. In order to do this we need a large set of contact maps S^{μ}, all corresponding to possible three-dimensional structures! Once such an ensemble of maps is generated, the fixed sequence $\mathbf{a}^p$ is threaded through all of them and the energies

$$E^{\mu,p} = H[\mathbf{S}^{\mu},\mathbf{a}^p] \qquad (24)$$

are evaluated and compared with the native energy

$$E^p = H[\mathbf{S}^p,\mathbf{a}^p]. \qquad (25)$$

We performed extensive studies of this nature for a large number of sequences. Detailed results on the low-energy maps that were "identified" by various sequences and the effect of dynamics on these maps will be discussed later. Here we present only some general results on large samples of sequences, from which we can learn about the merits and shortcomings of our energy function, especially in comparison with other knowledge-based potentials.

As a first step we generated an extensive library of allowed contact maps. We obtained allowed maps for a protein of length N from the native contact maps S^p of all proteins of known structure and length $N' \geq N$. From such a map we can extract $N' - N + 1$ submaps $\mathbf{S}^{\mu}$, $\mu = 1, 2, ..., N' - N + 1$, using all submatrices of size $N \times N$ that lie along the diagonal of the large map:

$$\mathbf{S}^{\mu}_{ij} = \mathbf{S}^{p}_{i+\mu,j+\mu} \qquad (26)$$

Clearly each such submap corresponds to the structure of an actually realized polypeptide chain of length N, and hence for each of these submaps $H^{constraint} = 0$. For example, from a set of 355 protein structures and a template sequence of length $N = 100$ one obtains this way about 50,000 allowed submaps. Clearly, the native map of every protein tested was included in the set of its candidate allowed maps. After evaluation of all the energies $E^{\mu,p}$ obtained from the sequence of protein p and the submap μ, the following questions were posed:

1. Can a sequence identify its native fold (if present in the set of tested maps)? That is, does one get $E^p \leq E^{\mu,p}$?
2. How large is the gap between the native energy and the mean energy of the entire set of candidate maps?
3. To what extent are the energies $E^{\mu,p}$ obtained for various maps correlated with the deviation of the map S^{μ} from the native map S^p?
4. What are the roles of the different terms in our energy function in determining the answers to the questions posed above?
5. Can the energy function be used to recognize proteins with homologous structures?

The manner in which we investigated these questions and the specific answers are as follows.

We tested 249 proteins of lengths $52 \leq N \leq 456$, using about 10,000 maps (on the average) for each sequence. The native map had the lowest energy for 238 out of the 249 proteins tested; that is, for a fraction

$$f_{correct} = 0.956 \qquad (27)$$

of the sequences the native map was correctly identified. Hence the answer to the first question posed above is positive for an overwhelming majority of the cases.

A quantitative answer to the second question is provided in terms of the parameters Z^p , which measure the deviation of the minimal energy $E^p_{min} = \min_{\mu} E^{\mu,p}$ from $<E>^p$, the mean energy of the set of maps tested for protein p, measured in units of σ^p, the standard deviation of the energies in the set

$$Z^p = \frac{E^p_{min} - <E>^p}{\sigma^p} \qquad (29)$$

As mentioned above, for all but 11 proteins we found $E^p{}_{min} = E^p$. We calculated the average of Z^p for all 249 tested proteins, and found the following Z score:

$$Z = -5.9. \qquad (29)$$

That is, the lowest energy obtained for a given sequence lies, on the average, about 6 standard deviations below the mean energy of the set of maps tested. The answers to questions 3 and 4 are related and hence will be discussed together. First, to our surprise we found that when only the pairwise term was used, the fraction of correctly identified native maps *increased* to $f^{pair}_{correct} = 0.98$; that is, only for 5 sequences (out of the 249 tested) were the lowest energies obtained with nonnative maps. Furthermore, when the pair-potential alone was used to calculate the Z score, we found $Z^{pair} = -6.6$. That is, the pairwise part of the potential apparently captures more of the energetics of contact maps than the hydrophobic part, with respect to both aspects addressed above, that is, identification of the native map and generating a larger gap. Nevertheless, the hydrophobic term is of considerable importance in our analysis, since it provides a strong *correlation between the energy of a map and the extent to which it differs from the native map.* These correlations are very important for using our method to screen and identify maps of proteins of unknown structure.

In order to present a quantitative measure of these correlations, we used a definition of the distance between two contact maps $\mathbf{S}^A$ and $\mathbf{S}^B$ that was introduced by Yee and Dill[34]:

$$d_{A,B} = \frac{\sum_{i=1}^{N}\sum_{j=i+1}^{N}|\mathbf{S}^A_{ij} - \mathbf{S}^B_{ij}|}{N^A_c N^B_c} \qquad (30)$$

where N^A_c and N^B_c are the numbers of contacts on maps A and B. The numerator of this expression counts the number of mismatches (e.g., the Hemming distance) between the two maps. For about 10 proteins selected at random we calculated the correlation coefficient between distance form the native map and the two energy terms, pair and hydrophobic. The results were

$$\rho_{pair} = 0.0 \pm 0.1 \quad \rho_{hydro} = 0.75 \pm 0.1. \qquad (31)$$

The significance of the hydrophobic interactions in identifying maps that lie close to the native one is therefore evident. We view the very high value of the correlation coefficient ρ_{hydro} as one of the most promising aspects of our proposed method.

Recognition of Homologous Structures

We carried out a detailed investigation of the ability of our potential to recognize folds with different sequence but relatively high structural similarity. As mentioned above, the capability for such recognition (obtained by means of assigning low energy values to the homologous partial contact maps) is of central importance for predicting the structure of proteins. When the native structure is not present in the dataset, a fold recognition procedure has to find a structure which is similar to the native one.

Several examples of proteins with similar structures and nonhomologous sequences are known. These proteins can be used to test our energy function's ability to recognize homologous proteins. To carry out this test we selected for each sequence in our dataset the twenty nonnative structures (maps) to which the lowest energy values were assigned. These 20 structures were compared with the native structure using unbiased protein structure comparison techniques.[36]

Since the dataset of maps whose energy was evaluated contained between 6,000 and 20,000 candidate maps for each sequence (the precise number depends of course on the length of the sequence), the 20 selected maps constitute a minute fraction (between 0.1% and 0.3%) of the set.

For many proteins our potential successfully recognized similar structures. A list of 80 homologous pairs identified this way will be provided upon request. To characterize similarity between structures of recognized proteins we refer to the FSSP database.[44] This list of 80 homologous pairs was sorted on the basis of their FSSP Z score; only pairs with Z score >3.5 were included.

The list of homologous proteins successfully recognized by our potential contains also several pairs of proteins that are functionally different, sharing, however, a great deal of structural similarity (e.g., TIM barrels, globin folds; see the recent review by Orengo and colleagues.[45] for an extensive list of such protein pairs). In Table III we highlight only a short representative list of successfully recognized homologous structures, containing known[45] examples of proteins with high structural similarity and low sequence homology.

Our potential was also able to recognize similarities between proteins of the same structural family even when sequence homology between proteins is lower than 20%. One of the most pronounced example of family recognition is the family of L-arabinose binding proteins. The sequence of D-ribose-binding protein (2dri) recognizes the structure of Leu-,isoleu-,val-(LIV)-binding protein (2liv) as one of very low energy. The structure of this protein (2liv) is recognized also by the sequence of the L-arabinose-binding protein (8abp), which belong to the same family and indeed has high structural similarity with the LIV-binding protein.

Proteins of the L-arabinose binding family also recognize the B chain of tryptophane synthase (1wsy-B), ae well as thioredoxin reductase (1trb) and two proteins of the thioredoxin family: glutathione peroxidase (1gpl) and thioredoxin (3trx). These proteins have mixed a b folds and share significant amount of structural similarity according to Dali structure alignments.[36]

There also were several sequences that have known homologous structures which our fold recognition procedure failed to identify. For example, the structure of carboxipeptidase A did not appear among the best fitting for the sequence of Ras p21

TABLE III. Short List of Representative Homologous Structures Recognized by Our Energy Function

Globins	Myoglobin	–	Colicin A
	Myoglobin	–	Phycocyanin C
	Cytochromes c and B256	–	Phycocyanin C
	Hemerythrin	–	Apolipoprotein E2
TIM barrel	Adenosine deaminase	–	D-xylose isomerase
	Glycolate oxidase	–	D-xylose isomerase
	Aldose reductase	–	Adenosine deaminase
	Triosephosphate isomerase	–	D-xylose isomerase
Dehydrogenases	Lactate dehydrogenase	–	Malate dehydrogenase
Hematopoetic cytokins	Interleukin 4	–	Granulocyte colony stimulating factor
Others	Ferritin H	–	Ribonucleotide reductase R2
	Chloramphenicol acetyltransferase	–	Dihydrolipoamide acetyltransferase
	Class I MHC	–	Immunoglobin
	Tumor necrosis factor	–	Viral coat and capsid protein (SBMW)

proteins, although the structures of these two.* We identified, as the main reason for these failures, absence of alignment of the sequence to a structure. That is, when the sequence is threaded through a structure we do not allow for gaps. This renders impossible the recognition of a structure which is basically similar to the native one, but differs in the length of a few loops. In such a case the sequence assigns low energies only to small regions of the structure, which might be insufficient for successful recognition. In some of the successful cases presented above even this partial recognition was sufficient to provide low energies for the homologous structures.

We conclude this section by presenting in detail the lower structures obtained using the sequence of myoglobin (1mba) as the template. Note that neither myoglobin nor hemoglobin were used for the derivation of the potential energy; these proteins were not "seen" when the parameters of our energy function (or dynamic rules) were obtained. The histogram of energies is shown in Figure 2 (only the energies of a few hundred maps are shown). The structures are sorted in order of increasing energy. Structures with the first 10 lowest energies are shown in Table IV. Clearly the native fold of myoglobin has the lowest energy by far. Then come the hemoglobins, which are the other members of the globin family, followed by hemerythrin which is also a heme-containing oxygen transport protein. Hemerythrin consists of four domains, each of four α helices, and has a similar fold to myoglobin which is a one-domain protein of eight helices. Having sequence identity below 15% with myoglobin, hemerythrin was shown to have a similar globin fold.[46] Next come two proteins which have mixed α/β structure. The structures of these proteins contain regions of several α helices that were recognized to have a globinlike fold. The second of them—phosphogluconate dehydrogenase (1pgd)—has a large domain yhat consists of 8 α helices and has a fold similar to the globin fold. Finally there is one more protein, colicin A, that adopts globin fold. Unlike the majority of globins that serve for oxygen transport or storage, colicin A is an antibiotic protein.[36] We calculated the root mean square deviation (RMSD) between structures using the program[47] StructAl. The RMSd of colicin A from myoglobin is 4.0 Å for 115 Cα atoms. Hence their 3D structures are rather similar, and colicin A was correctly recognized by our method.

Comparison With Other Potentials

An important issue that has to be addressed is how does the quality of our results compare with similar results obtained using other knowledge-based potentials? We tried to answer this question in as fair and impartial manner as possible. One should bear in mind that not all potentials are published and available for detailed comparison. Furthermore, reported values of various indicators (such as $f_{correct}$ and the Z score) are sensitive to the selection of the protein sequences and the library of maps that are tested. In what follows we present a variety of comparisons with different potentials.

The first figure of merit, the fraction of correctly identified sequences (out of our pool of 249 proteins) had the value $f_{correct}$ = 0.956 for our total energy and $f^{pair}_{correct}$ for our pair potential part. The same screening experiment on the same set of sequences and maps, using the (pair) potential of Kolinsky and Skolnick[19] had a slightly lower success rate, $f^{KS}_{correct}$; the number of sequences whose native map was not identified was 16. On the other hand, the Z score obtained for this potential is better than ours: $Z^{KS} = -8.1$. Similarly, the Z score (as obtained and reported by other authors on different sets of sequences and templates) of the Sippl and Levitt potentials is better than ours. Sippl[48] reported $Z^{pair} = -6.8$ and a score $Z = -9.6$ for a potential combined of pairwise interactions and interactions with

*Proteins have similar topology. Our method did not find homology between staphylococcal nuclease and a group of toxins (verotoxin, enterotoxin).

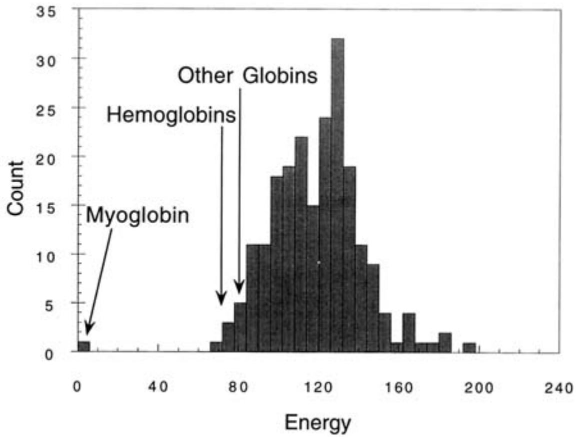

Fig. 2. Fold recognition test. Histogram of energies obtained by threading the myoglobin sequence through the different conformations of a protein set. The native structure has the lowest energy and other members of the globin family come next.

the solvent. For a hydrophobic potential Huang and colleagues.[49] reported $Z = -8.2$.

These values were obtained from different sets of proteins and different banks of maps; hence using them, either to compare the Sippl[48] and Levitt[49] potentials to each other or with our potential is not too meaningful. As a different basis for comparison we give now results for three specific cases that are known to be difficult for such screening tests. The first sequence is that of BPTI, for which we compare the energies of the native map with two misfolded structures that were obtained in a recent paper of Elber and Kaesar.[37] They used potentials of Sippl[50] and Crippen[51] and found that the two misfolded structures have lower energies than the native state. As can be seen from Table V, our energy function does identify the native state as that of lowest energy. The same is true for the Hinds/Levitt potential,[20] whereas the potentials of Kolinsky and Skolnick, as well as that of Sippl (which reportedly selected the misfolded structures) assign lower energies to nonnative folds. Interestingly, the success of our potential can be attributed to the relatively low value that was assigned to the native fold by the *hydrophobic* part of our energy function.

The second special case is that of Cro-protein (pdb code 3cro), known to be problematic for fold-recognition procedures.[38] For this protein the native fold was ranked as that of lowest energy by our potential[19] and also by that of Hinds and Levitt. The Kolinsky-Skolnick potential found 9 folds with lower energy than the native one: the energy function of Miyazawa and Jernigan[52] ranked the native map as number 60 in its energy.

The third special case is that of spectrin (pdb code 2cps, chain A). The potential of Kolinsky and Skolnick failed to recognize this protein; the lowest energy was assigned to the spectrin sequence folded into a fragment of Theiler's murine encephalomyelitis virus (pdb code 1tme, chain 1). The potential of Hinds and Levitt[20] also failed to recognize spectrin and identified a fragment of β-actin (2btf, chain A) as the best fitting map for the spectrin sequence. Our potential managed to recognize spectrin correctly. Interestingly, again our pairwise potential was not able to do it alone and only the combination of both pairwise and hydrophobic terms provided the correct result.

In summary, we believe to have demonstrated that our energy function is of comparable quality to existing knowledge-based potentials, and in some aspects works better than any particular potential. We also identified the special role of the two different terms in our energy function, and demonstrated that the novel hydrophobic term is of importance, especially in that it provides high correlation between the energy of a screened map and its distance from the native fold.

DYNAMICS IN THE SPACE OF CONTACT MAPS

Having performed extensive tests of our energy function, we tried to use it in a dynamic Monte Carlo process. The idea is to use a fixed sequence and change its contact map by eliminating existing contacts (turning $S_{ij} = 1$ to $S_{ij} = 0$) or addding new ones ($S_{ij} = 0$ becomes $S_{ij} = 1$). To our knowledge, this is the first attempt at using the contact map reduced representation in such a dynamic process. As a first step towards establishing a reasonable dynamic procedure, we investigated the stability of native maps against their modification.

Stability of the Native Contact Map

The native structure, obtained by x-ray analysis of protein crystals, constitutes an average over many conformations of the protein. Molecular dynamics simulations done in the vicinity of the native state confirm that there is a set of protein conformations very similar to the structure solved by x-ray crystallography, each corresponding to local minima of an energy function.[24] Consequently, starting minimization from the native state one expects, for realistic energy functions of a protein, to find a set of energy minima in the vicinity of the native conformation. One should note an important point: the answer to the question whether a given conformation (or contact map) lies at a local minimum of the energy function is not determined by the energy function alone; the answer depends on the dynamic procedure used, since this determines the possible steps that can be taken from a given state. We try to choose the allowed changes of a contact map in a way that corresponds to physically plausible moves of a protein. Furthermore, we start from a map for which $H^{constraints} = 0$, and wish to formulate dynamic rules that ensure that the map obtained after

TABLE IV. Proteins for Which the Lowest Energies* Were Obtained When Substituting the Myoglobin Sequence Into Their Contact Maps

Code	Length	E	E^{pair}	E^{hyd}	Distance from 1mba	Compound
1mba	146	0.7	−38.6	39.4	0.000000	Myoglobin
1dxu	574	70.9	1.2	69.6	0.001478	Hemoglobin
1hbb	574	73.0	2.5	70.5	0.001486	Hemoglobin A
3sdh	290	74.4	−1.6	76.1	0.001319	Hemoglobin I
4hhb	574	74.5	3.7	70.8	0.001481	Hemoglobin
2hmq	452	79.6	−5.7	85.3	0.001465	Hemerythrin
1cmy	574	79.6	−3.9	83.6	0.001586	Hemoglobin Ypsilanti
1gly	470	80.6	0.3	80.2	0.001852	Glucoamylase
1pgd	469	81.3	0.2	81.0	0.001806	Phosphogluconate dehydrogenase
1hmd	452	81.4	−4.9	86.3	0.001465	Hemerythrin
1col	394	85.5	3.6	81.9	0.001369	Colicin A

*The total energy E and the parts E^{pair}, coming from the contacts, and E^{hyd}, from the hydrophobic term, are presented together with the distances of the corresponding maps from that of myoglobin.

TABLE V. Comparison of the Performance of Our Potential With Others, for a Few Difficult Fold-Recognition Tasks

	This work			Hinds/	Kolinsky/	
Structure	E	E^{pair}	E^{hydro}	Levitt E^{pair}	Skolnick E^{pair}	Miyazawa Jernigan
BPTI						
Native	17	−19	36	−53	−0.8	
Misfolded 1	33	−23	56	−46	−6.7	
Misfolded 2	27	−18	45	−32	−3.1	
Summary	Correct			Correct	Incorrect	
Cro-protein (3cro)	Correct			Correct	Incorrect	Incorrect
Spectrin (2 cps, chain A)	Correct			Incorrect	Incorrect	

the change is also physically realizable, and hence also has $H^{constraints} = 0$. To demonstrate that this is indeed necessary, we started our investigation by using a "naive" Monte-Carlo procedure for simulation of contact map dynamics, which ignores $H^{constraints}$. The simulation started at a native contact map.

At each step of the dynamics a randomly picked pair of amino acids i and j was considered. An attempt was made to change the contact variable, S_{ij}, of this pair; (if $S_{ij} = 1$, then change to $S'_{ij} = 0$; if $S_{ij} = 0$, then to $S'_{ij} = 1$), and the difference of energy $\delta E = H(S') - H(S)$ was calculated (ignoring $H^{constraints}$). Only for negative δE was the new state of the pair accepted. After that a new pair i and j was picked, an so on. This procedure was applied till the system reached a state, such that any step taken from it would increase the energy. This state is clearly a local minimum of the energy function (under the dynamics of changing one contact at a time). In order to check this, after every 100 steps all the pairs i and j were tested according to the procedure described above. If at least one pair with $\Delta H < 0$ was found, then the minimization procedure was continued; for each final state of the map its energy and distance d_{AB} [see Eq. (30)] from the native map were calculated.

The "naive" energy minimization described above gave the following results:

1. The pattern of the contact map was completely destroyed after about 1000 steps (Fig. 3).
2. The energy decreased rapidly and went to a limiting constant value, well below that of the native state (Fig. 4a).
3. The distance of the final contact map from the native one is large (Fig. 4b).

Similar results were obtained even when the "naive" procedure was modified in a manner that preserves the total number of contacts. These results mean that the native state is far from the local energy minima that are reached by the naive minimization algorithm. The final states that are reached do not correspond to any realistic structure. These are serious flaws, and they are due to the fact that our "naive" minimization procedure ignored $H^{constraints}$. In the process of energy minimization a single new contact between *any* pair of amino acids was allowed to be created. Real physical contacts can be created, however, only when the corresponding pair of amino acids is separated by a small distance. This means that by choosing an amino acid pair randomly and trying to form a contact for this pair, we are likely to violate some geometric con-

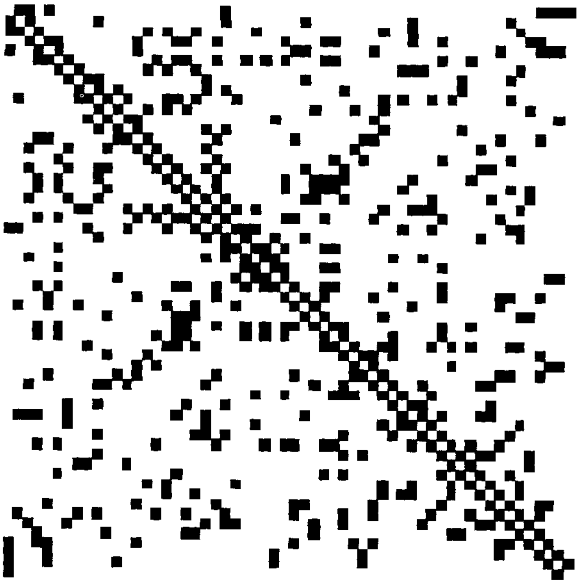

Fig. 3. Final contact map corresponding to local minimum reached by "naive" Monte-Carlo energy minimization that ignored $H^{constraints}$. The minimization started from the native contact map (see Fig. 1).

straints of the structure. Had we had in Equation (3) an explicit expression for $H^{constraints}$, assigning infinite energy to unphysical contact maps, such Monte Carlo moves would have been rejected. In order to take into account the constraints implicitly, we introduce restrictive dynamic rules.

Dynamic Rules

Dynamic rules restrict the changes that a contact map can undergo in a single Monte Carlo step. The basic idea is that if one starts from an allowed contact map (that corresponds to a physically possible structure), making moves according to these rules will guarantee that the resulting new map is also physically allowed.

The dynamic rules define which states of the contact map S^{t+1} are accessible from the current state S^t by changing only one cell on the map. In particular, our dynamic rules forbid the appearance and removal of some contacts. At each step of the Monte-Carlo procedure one chooses a cell on the map randomly, and then checks whether the state of this cell can be changed, according to the dynamic rules. In other words, the dynamic rules determine whether the particular pair of amino acids can move in order to form or to destroy a contact between them. It is clear that the possibility of forming a contact between a given pair of amino acids depends on the spatial distance between them, and on the flexibility of the protein in the particular conformation. Although the contact map does not contain explicit information about the distances or conformations of particular amino acids, one can partially retrieve such information from the sates of neighboring pairs. Here and later we refer to the amino acid pairs $(i \pm 1, j \pm 1)$; $(i, j \pm 1)$; $(i \pm 1, j)$ as the neighborhood of pair (i, j). Whether the state of contact of pair (i, j) can be changed or not is determined by the contacts in its neighborhood and by the distance $|i - j|$ along the polypeptide chain. For nearby amino acids (with $|i - j| \leq 4$) we distinguish between two main conformations: helical (turned) and nonhelical (extended). Hence changing the contact of an amino acid pair (i, j) with $|i - j| \leq 4$ can be considered as part of a transition between turned and extended conformations. The influence of chain connectivity is less important for pairs that are well separated along the chain, that is, $|i - j| > 4$, and the possibility to change the state of contact of such a pair depends mainly on its spatial distance and environment. Consequently, we distinguish between two regions of the contact map: the diagonal region, with $|i - j| \leq 4$, and off-diagonal contacts, $|\mathrm{i} - \mathrm{j}| > 4$. Different dynamic rules are applied to pairs belonging to the two different regions of a contact map .

Dynamic rules for the off-diagonal region

To construct these dynamic rules we analyzed possible states of a pair's neighborhood, according to the number of neighboring contacts P. Any pair of amino acids in the protein structure can be classified into one of the following categories, according to the state of the corresponding cell on the contact map: *distant pair*—corresponds to a cell that has less than P_1 contacts in its neighborhood; *proximate pair*—the cell has P_1 to P_2 neighboring contacts; and *packed pair*—the cell has more than P_2 neighboring contacts. Here P_1 and P_2 are two parameters of our dynamic rules. After considerable testing, we found that the values

$$P_1 = 4, P_2 = 6 \qquad (32)$$

were optimal. Now we formulate dynamic rules for the off-diagonal region describing possible dynamics of amino acid pairs in each of the three categories.

Creation of a contact for a distant pair of amino acids is very improbable (or even impossible), since it requires significant conformational changes in the whole protein structure. Consequently we prohibit creation of contacts between distant pairs. Prohibiting creation of contacts between distant amino acids requires to prohibit also annihilation of the existing contacts of this type, in order to preserve detailed balance in the simulation. The only possible dynamics for distant amino acids is 'sliding'; breaking of the i, j contact and creation of one at $i, j \pm 1$ or $i \pm 1, j$. In terms of the contact map this rule can be stated as follows: *if the number of neighboring contacts is less than* P_1, *then an existing contact can move to one of the neighboring cells.* The only one possible asymmetry that may occur happens when a distant cell is the neighbor of a nondistant cell, only

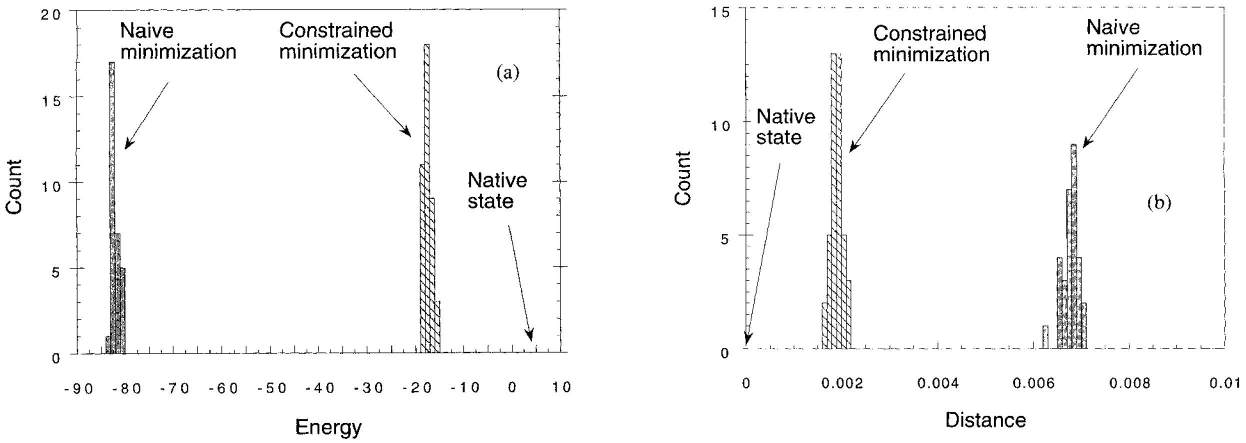

Fig. 4. Histograms of **(a)** energies of contact maps and **(b)** distances to the native map, obtained by constrained (*dashed bars*) and 'naive' (*solid bars*) energy minimizations, starting from the native state.

one of the two is in contact, and contact motion was generated between these sites. This is very rare.

Considering packed pairs, one can see that both disappearance and sliding of existing contacts are very improbable, mainly because of excluded volume interactions with nearest amino acids. We prohibit the annihilation of a contact for this pair. Conservation of detailed balance requires prohibiting also the creation of new contacts for packed pairs. Thus, for a packed region the dynamic rules are: *if the number of neighboring contacts for a given cell is larger than* P_2, *then no changes of the cell state are possible.*

We turn, finally, to a cell belonging to the intermediate category, that is, cells whose number of contacts m is such that $P_1 \leq m \leq P_2$. These cells correspond to a pair of amino acids at some intermediate distance and are free to form a contact or to break an existing contact. According to the introduced dynamic rules, contacts of the intermediate category make the most significant contribution to the dynamics of the contact maps. The rule is: *if m, the number of neighboring contacts of the cell satisfies* $P_1 \leq m \leq P_2$, *then the state of the cell can be changed.*

Dynamic rules for the diagonal region

The diagonal region of a contact map contains contacts between close-by amino acids along the polypeptide chain. Since the most important structural elements displayed in the diagonal region are turns and helices, we consider dynamics in this region as changes of these elements. Here we distinguish between the following events: helix extension, helix shrinking, turn appearance and turn disappearance. Actually, we consider a turn as the main "building block" of helices. A turn is a set of three contacts $(i, i + 4)$, $(i, i + 3)$, $(i + 1, i + 4)$. It looks like a right-angle-shaped group of three contacts on a contact map (Fig. 5). A helix can be considered as a set of overlapping turns, and helix extension (shrinking) is an addition (loss) of the end turn. Dynamic rules in the diagonal region are formulated in terms of turns (not single contacts) and a single event is a change of the turn state.

Since a turn consists of three cells, there are $2^3 = 8$ possible states of each turn (Fig. 5). The a priori probability of a transition between states a and b of a turn is given by a transition matrix W_{ab}, normalized to have $\Sigma_b W_{ab} = 1$. To preserve detailed balance we set $W_{ab} = W_{ba}$. At each step of the Monte-Carlo procedure for the diagonal region one turn is chosen randomly on a contact map. A new state of this turn is then selected according to the matrix W_{ab}. The elements of the matrix depend on whether the chosen turn is in the body of a helix, at the end of a helix or not in a helix. The matrix W_{ab} is constructed in a way that prohibits rare events (breaking of a helix in the middle, or appearance of a new turn in an extended region) and to increase the probability of helix extension and shrinking at the ends (see Table VI). From the physical point of view, cooperative creation and annihilation of contacts in the diagonal region reflects cooperative helix formation in proteins, mentioned as an important feature of protein folding.[22,53]

Elimination of unrealistic contacts

The aim of our dynamic rules is to provide realistic dynamics of contact maps and to suppress emergence of maps that have geometrically impossible contacts. We identified two main classes of such contacts; both correspond to interaction of helices with other structural elements. Since our rules take into account only local neighborhoods of the contact that is changed, formation of the unphysical contacts described below is not eliminated by the dynamic rules described above. Therefore, we introduced additional dynamic rules, whose task is to prohibit appearance of these unrealistic contacts.

Contact (i, j) is classified as unrealistic of the first

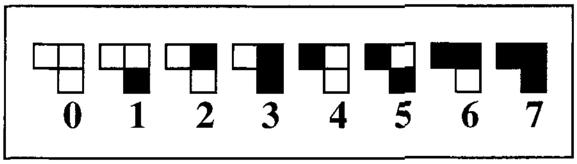

Fig. 5. Right-angle-shaped block that corresponds to a turn in a helix on a contact map. All eight possible states of the block are shown.

class if i and j belong to the *same* α helix and $|i - j| > 4$. Existence of such an (i, j) contact would bend the helix in a way that has never been observed in protein structures. Contact (i, j) belongs to the second class if any $(i, j \pm 2)$ contacts exist, when j and $j + 2$ belong to the same α helix and $|i - j| > 4$. In fact, if i is in contact with $j + 2$, then it can not form a contact with j because j is on *the other side* of the helix. To eliminate unrealistic contacts, we extend our dynamic rules in a way that prohibit their creation in the off-diagonal region of a map, as well as the extension of helices in a manner that makes some existing contacts unrealistic.

Our dynamic rules are supplemented by the standard Metropolis procedure to simulate dynamics at a non zero temperature T. If the chosen cell can change its state, the energy difference δE associated with this change is calculated, and if $\delta E \leq 0$, the new state is accepted. When $\delta E > 0$, we accept the new state with probability $e^{-\delta E/T}$. This is achieved by selecting a random number ρ, uniformly distributed in the interval $0 \leq \rho \leq 1$; if $\rho < e^{-\delta E/T}$, the new state is accepted. The procedure is guaranteed to lead the system to thermal equilibrium at temperature T. At $T = 0$, this procedure corresponds to energy minimization.

DYNAMICS AND FOLD RECOGNITION

We now describe some applications of our energy function and dynamic rules to different tasks of protein science. We consider and discuss (a) simulation of protein dynamics in the vicinity of the native state, (b) melting of the structure at high temperatures followed by slow cooling and refolding, and (c) searching for the lowest energy map of a fixed sequence, using a combination of screening allowed possible maps with our dynamic procedure.

Energy minimization from the native state

Starting from a native contact map a cell (i, j) was randomly chosen. If the state of contact of (i, j) was allowed to change (according to the dynamic rules introduced above) we calculated the energy of the new state and accepted only steps that decreased the energy. After a relatively small number of such steps ($\sim 10^2$) the system reached a local minimum, that is, such a state that any step permitted by our dynamic rules increased the energy. The procedure was repeated several times starting from the same native map, but for different sets of random numbers (that govern the choice of the contact to be changed at each step). Histograms of energy values and distances from the native map are shown on Figure 4. The results demonstrate that (1) contact maps corresponding to local minima, found by the constrained Monte-Carlo procedure, are very similar to the native one; (2) the energy values of these local minima are all in the vicinity (not significantly below) the native state's energy. Hence the native state has a set of local minima at it's vicinity. The contact maps of the final states had very similar secondary structures to those of the native initial map. The α helices and β sheets are in similar locations, and the off diagonal groups of contacts also appear in the same regions as on the native map (see Figure 1). A typical final map of our dynamics has about the same number of contacts as the native map ($N_c \approx 180$). The distance $d \approx 0.002$ corresponds to about 65 mismatches. These mismatches arise from about 30 contacts (out of the 180) being *slightly* displaced, leaving the general appearance of the final map very close to the native one. These numbers are to be compared with the typical distances $d \approx 0.007$ obtained from naive unconstrained minimization, which yielded final maps (see Figure 3) of typically 280 contacts and 350 mismatches with the native map. We conclude that our energy function, when supplemented by our dynamic procedure, captures some important properties of the real protein energy landscape and dynamics.

Low- and high-temperature dynamics

In the next computer experiments we heated the protein (BPTI) to a low temperature (again starting from the native state, but accepting steps that raise the energy, with probability $e^{-\delta E/kT}$). The energy increases and fluctuates around $E \approx 40$, while the distance to the native state is $2.8 \cdot 10^{-3}$. Then the system was quenched suddenly to $T = 0$, and the ensuing dynamics took it to a local minimum of the energy. We found that the native structure is relatively stable at low temperatures ($T \sim 1$). Analysis of contact map 'snapshots' made during simulation shows stability of secondary structure elements as well. Events of contact appearing/disappearing and contact moving happen mainly in the off-diagonal region ($|i - j| > 5$) of the contact map. In a real protein this low-temperature dynamics corresponds to side chain reorientations and slight conformational changes of the main chain. Our next experiment was to submit the protein to a "high temperature shock," allowing it to interact with a heat bath (at $T = 7$) for *short times.* We observed now melting of secondary structure elements, with β sheets melting before α helices. These observations are in good agreement with molecular dynamic simulations of this protein.[54] This partial melting of protein structure is reversible for short heating durations; de-

TABLE VI. A Priori Transition Probabilities W_{ab} Between States *a* and *b* of a turn.*

Conformation	0	1	2	3	4	5	6	7
0	0.00	0.15	0.00	0.15	0.00	0.00	0.00	0.70
1	0.15	0.35	0.00	0.40	0.00	0.00	0.00	0.10
2	0.00	0.00	0.95	0.05	0.00	0.00	0.00	0.00
3	0.15	0.40	0.05	0.25	0.00	0.00	0.00	0.15
4	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.95	0.00	0.05
6	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
7	0.70	0.10	0.00	0.15	0.00	0.05	0.00	0.00

*The eight states are shown in Figure 6.

creasing the temperature slowly (annealing) leads to 'rebuilding' of secondary structures. We note also that the system returns to the initial range of the energy and to short distance to the native state. On the other hand, long heating significantly corrupted the contact map, and then annealing did not succeed to return the system to the vicinity of the native state. A plot, presenting energy and distance from the native state, shown at a sequence of time steps, measured during relaxation (annealing), is shown in Figure 6, for two Monte Carlo runs. Successful relaxation returns the system to the region near the native state and with the same value of the energy. For unsuccessful refolding, that started from a state significantly different from the native one, we observe trapping of the system far from the native state.

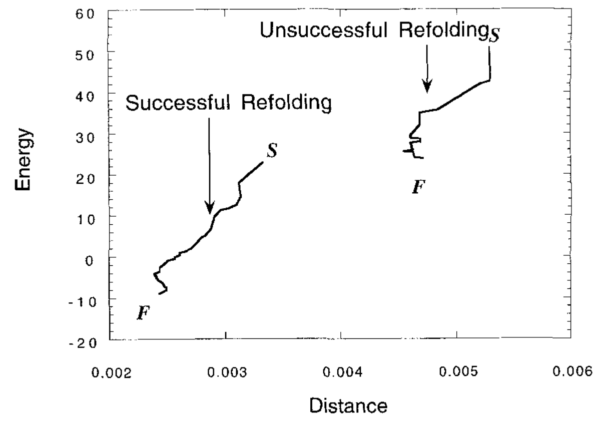

Fig. 6. Energy and distance from the native state, shown at a sequence of (Monte Carlo) times for successful and unsuccessful refolding of a protein, starting from two partially unfolded states. The starting state for refolding is denoted by *S* and the final state by *F*.

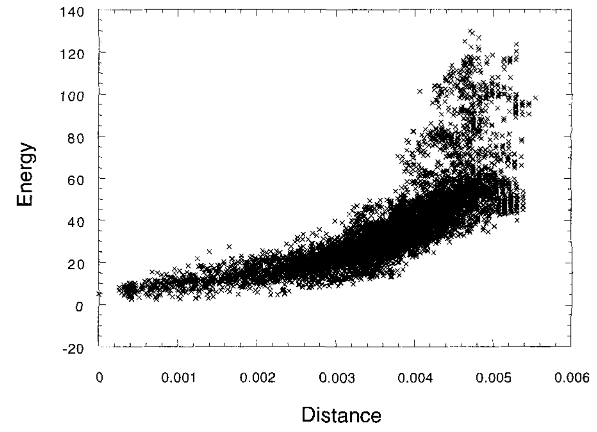

Fig. 7. Energy landscape of a protein in the vicinity of the native state. Energy is plotted as a function of distance from the native state when the system is heated in several Monte-Carlo runs.

Protein energy landscape

The failure to refold significantly unfolded proteins can be explained by the roughness of the protein energetic landscape. When cooled, a significantly unfolded protein goes down in energy and misfolds. The energy values of these misfolded structures are higher than those of the correctly folded ones. Consequently, the energy function is able to identify misfolded structures, but our dynamic procedure is unable to follow the correct folding pathway.

To investigate the protein energy landscape in the vicinity of the native state we recorded the energy values after every 10th step of our Monte-Carlo heating procedure at T = ∞. Figure 7 shows the energy as a function of distance from the native state, as obtained during several heating runs. There is a region of similar energy values near the native state, with $d < d_c \approx 4 \cdot 10^{-3}$, beyond which the energy increases rapidly. From states within this region the protein refolds to the vicinity of its native state, whereas states with $d > d_c$ lie in other "valleys" of the energy landscape.

Combining dynamics with fold recognition

As discussed in detail in the section on fold recognition, when a given protein sequence is threaded through a large set of candidate maps obtained from the database of known structures, maps that correspond to similar structures to the native one were recognized in many cases among those with low energy. However, the energies of these nativelike structures were not always the lowest (i.e., right next in rank to the native structure). We present now initial investigations aimed at elucidating whether dynamics can help to identify which of the low-lying submaps have indeed nativelike structure and which are "false positives."

We have chosen several cases to test whether dynamics can improve results of fold recognition. For each fixed sequence we first performed regular fold recognition and selected the 50 structures (maps) whose energies were the lowest. Then, to let the map adapt to the fixed threaded sequence we run dynamics, using each selected map as the staring point of our dynamics. We present in Table VII results obtained from dynamics at $T = 0$ and at $T = 0.05$, for four different fixed protein sequences. The initial and final ranks (in energy) of known strands of high structural homology are presented and discussed.

For the *D-ribose-binding protein (2dri) sequence* four homologs were found among the 50 maps of lowest energy. As can be seen, dynamics moved all four homologs to one of the first six positions. On the other hand, a false-positive map (2tbv), whose structure differs from the native one, has, even after dynamics, the lowest energy among nonnative structures.

Among the four homologs of the *interleukin 4 (1rcb) sequence* dynamics improves recognition of 1bgc, but also increases the number of false positives and moves the other nativelike structures (1mrr, 1col, 1vsg) far from the top of the list.

For the *cytochrome c′ (2ccy) sequence* dynamics improve recognition for some nativelike structures, but also raises the rank of some. It should be noted that computational limitations forced us to perform only relatively short runs of dynamics, that is, 16 N^2 MC steps, where N is the length of the protein. Results of the combined fold recognition and dynamics procedure depend on the temperature used for dynamics and on the length of the Monte Carlo runs (data not shown). This leads us to the conclusion that longer runs are needed to allow the homologous contact maps to adapt to the one optimal for the threaded sequence. Acceleration procedures such as simulated annealing may enhance the contribution of dynamics to fold recognition.

Such an experiment was performed for another well-known pair of proteins with very low sequence identity and high structural similarity: *plastocyanin and pseudoazurin.*[30] Both proteins adopt eight strand β barrel folds. Using the sequence of plastocyanin, the native fold was recognized as the best one, as expected, but the very similar pseudoazurin structure had a relatively high energy and rank 6, with five false-positive results with slightly lower energies. Again using the structures of lowest energy as our starting point, we performed extensive dynamics, slightly heating and cooling these maps. As a result, the pseudoazurin structure acquired the closest energy to that of the native fold of plastocyanin (see Table VIII).

TABLE VII. Effect of Dynamics on the Ranking of Homologous Structures by (Increasing) Values of their Energies

Threaded sequence	Homolog	Rank before dynamics	Rank after dynamics ($T=0$)	Rank after dynamics ($T=0.05$)
2dri	2liv	3	2	6
	1pfk	4	3	3
	2ada	10	4	4
	1phh	>20	5	2
1rcb	1bgc	13	7	5
	1mrr	1	19	16
	1col	18	15	17
	1vsg	7	27	27
2ccy	1prc	5	13	5
	1brd	11	8	6
	1mrr	1	5	9

SUMMARY AND DISCUSSION

We introduced and calculated the parameters of an energy function for contact maps of proteins. This function contains, in addition to the widely used interaction term between amino acids that are in contact, a new term, that takes represents the variation in the hydrophobicities of different amino acids. For the first term we had to estimate the contact energies for any pair of amino acids, and for the second the desired number of contacts for each amino acid, and the widths of these distributions. These parameters were obtained from a statistical analysis of the data bank of known protein structures. Once these parameters were determined, the value of the energy can be calculated for any sequence folded into any physically realizable structure.

This energy function was tested extensively. We tested first the sequence specifity of the energy function, by using the contact map of a known structure, and substituting in it various sequences. Sequence specifity was demonstrated by the very low energy of the true, native sequence that corresponds to the template map. The second set of rather extensive tests involved fold recognition. This was done by keeping a template sequence fixed, and threading *it* through a large number of candidate maps, obtained from known structures (with longer sequences than the template). In an overwhelming fraction of cases the true map was recognized as that of lowest energy. Even in some cases known to be difficult for fold recognition, our potential did succeed. Careful comparisons were made between performances of various potentials with those of our energy function.

Other protein strands, whose structure is known to be similar to the native one, were contained (for most cases studied) in the subset of low-energy maps. Such homologous pairs were identified even

TABLE VIII. Energies of various contact maps through which the sequence of plastocyanin was threaded, obtained before and after dynamics*

	Before dynamics			After dynamics		
Protein	H	H^{pair}	H^{hyd}	H	H^{pair}	H^{hyd}
Plastocyanin	7.717	−28.602	36.319	−24.554	−41.608	17.054
L-Lactate dehydrogenase	62.731	5.488	57.243			
Eglin C	63.027	−0.703	63.730	15.124	−19.973	35.097
Apolipoprotein E3	63.940	14.423	49.517	33.675	5.435	28.240
Heat shock cognate 70	64.432	14.510	49.922	21.172	−2.856	24.028
Mesentericopeptidase	64.863	−0.540	65.403			
Pseudoazurin	66.948	7.659	59.289	5.204	−17.113	22.317
Apolipoprotein E4	69.246	17.912	51.334			
Colicin A	69.344	17.508	51.836			
Apolipoprotein E2	69.542	15.507	54.035			
Myoglobin	69.839	20.349	49.490			
Metmyoglobin	70.236	19.532	50.704			
Adenylate kinase	70.395	17.821	52.574	37.742	5.990	31.752
Myoglobin	70.588	19.413	51.175	38.151	5.000	33.151
Uteroglobin	71.641	19.419	52.222			
Hemoglobin F	71.799	15.664	56.135	37.918	5.439	32.479
Apolipoprotein E4	71.939	18.326	53.613			
Bira bifunctional protein	72.248	5.457	66.791	38.357	−9.587	47.944
Apolipoprotein E4	72.583	22.916	49.667			

*Dynamics improves the ranking of pseudodazurin significantly.

when the sequence homology was rather low. In addition, families of related structures were successfully identified by this screening test.

Another novel feature introduced here is that of Monte Carlo dynamics in the space of contact maps. This way of simulating protein dynamics, using our energy function to accept or reject moves, runs into a rather difficult problem; in general, changing a contact map may produce a new map that cannot possibly correspond to any physically realizable structure. This occurrence was controlled indirectly by restricting the changes one is allowed to make on a contact map. Since the restrictions were based on our understanding of the possible motions a real polypeptide chain, we believe that they exclude generation of nonphysical structures.

With these dynamic rules we performed simulations of protein dynamics at low and high temperatures. At low temperatures we found local minima of the energy function near the native map, with similar secondary structure elements β sheets and α helices). High temperature dynamics led to melting of the structure, with β sheets melting before α helices. When heated from the native state for a short duration, followed by slow cooling, the partially melted protein refolded easily to a structure very close to the native one. For longer heating the ensuing relaxation took the protein to a misfolded state.

Finally, our fold-recognition procedure was combined with dynamics, using the low-energy maps obtained for a given sequence as the starting points for dynamics, with encouraging preliminary results. We plan to develop this methodology further by development of more reliable dynamic rules, and using acceleration methods (such as simulated annealing) for the simulations.

ACKNOWLEDGMENTS

We thank M. Levitt and J. L. Sussman for most helpful discussions, and R. Elber and C. Keasar for making unpublished data available. The research of E.D. was partially supported by US-Israel Binational Science Foundation (BSF) and the Germany-Israel Science Foundation (GIF) and the Minerva Foundation.

APPENDIX

We start the derivation of Equation (15) from the expression for the energy of pairwise interactions.

$$H = \sum_{i<j} S_{ij} E_{a_i,a_j} = \sum_{u,v} E_{uv} \sum_{i<j} S_{ij} \delta_{wa_i} \delta_{va_i} = \sum_{u,v} E_{uv} N_{uv} \tag{A1}$$

We went from summation over all contacts to summation over all classes of contacts.

N_{uv} is the number of contact in uv class, and E_{uv} is energy of each contact in this class. In this form the energy function is the same as one of a system consisting of different compartments uv, each with chemical potential E_{uv}. Our purpose is to derive an expression for the average number of contacts in a class, when the chemical potential is given. Expressing the chemical potential as a function of countable quantities, one will be able to estimate E_{uv} for each class uv for any given contact map.

We fix the total number of contacts N_o in the system. The size of the system, M_o, is the number of all possible locations for a contact: $M_o = N_p(N_p - 3)/2$. The size of a class uv is denoted as M_{uv} and means the number of possible locations of a contact in a class. There are 20 different amino acids and consequently $(20 \times 21)/2 = 210$ different pairs or classes of contacts.

Almost all of them are represented on the contact map of a real protein. We note that in general the number of contacts in each class is much smaller than the total number of contacts and the size of each class is much smaller than the size of the whole map:

$$\frac{N_{uv}}{N_o} \approx 10^{-2} \qquad \frac{M_{uv}}{M_o} \approx 10^{-2} \tag{A2}$$

This enables us to assume that the average number of contacts in each class does not depend on the number of contacts in the other classes. Hence a class is in equilibrium with the whole map and we can write down the expression for the partition function of a class:

$$Z = \sum_{n=0}^{m} f(n)\, e^{-\varepsilon n} \tag{A3}$$

We omitted the index uv of a class. Here $f(n)$ is the weight of each conformation, proportional to the probability to find precisely n contacts located in the class of size m, when N_o contacts are randomly distributed on a map of size M_o. Assuming that contacts are distributed independently, this probability can be written as

$$f(n) = p^n(1 - p)^{N_o - n}\, C^n_{N_o} \qquad p = \frac{m}{M_o} \tag{A4}$$

where p is the probability to find one contact randomly dropped on a map of size M_o in the class of size m. For the case of large N_o and small p this distribution can be approximated[55] by the Poisson distribution with average $\xi = N_o p$. Since for a protein of length $N_p \approx 10^2, N_o \approx 70$, and $p = m/M_o \approx 10^{-2}$, we can apply the Poisson approximation for $f(n)$:

$$f(n) = e^{-\zeta} \frac{\zeta^n}{n!} \tag{A5}$$

where

$$\zeta = N_o \frac{m}{M_o} \tag{A6}$$

Now one can easily calculate the partition function of the class.

$$Z = \sum_{n=0}^{m} f(n) e^{-\varepsilon n} = e^{-\zeta} \sum_{n=0}^{m} \frac{\zeta^n}{n!} e^{-\varepsilon n} \approx$$
$$e^{-\zeta} \sum_{n=0}^{\infty} \frac{(\zeta e^{-\varepsilon})^n}{n!} = e^{-\zeta} \exp(-\zeta e^{-\varepsilon}) \tag{A7}$$

We performed summation to infinity, because m is much larger than the average value of n. To derive an expression for average number of contact in a class we recall that ϵ is a chemical potential. Hence

$$<n> = \frac{\partial logZ}{\partial \varepsilon} = \frac{\partial(-\zeta e^{-\varepsilon})}{\partial \varepsilon} = \zeta e^{-\varepsilon} = N \frac{M}{M_o} e^{-\varepsilon} \tag{A8}$$

For a given contact map and each class of contacts uv one can easily count the quantities N_o, M_o, N_{uv}, M_{uv}. The chemical potential of the class is estimated by

$$E_{uv} = -log \left[\frac{N_{uv} M_o}{N_o M_{uv}}\right] \tag{A9}$$

REFERENCES

1. Creighton, T.E. "Protein Folding." New York: W.H. Freeman, 1993.
2. Levitt, M. Protein folding. Curr. Opin. Struct. Biol. 1:224–229,1991.
3. Dill, K.A.: Folding proteins: Finding a needle in a haystack. Curr. Opin. Struct. Biol. 3:99–103, 1993.
4. Frauenfelder, H., Wolynes, P.G. Biomolecules: Where the physics of complexity and simplicity meet. Phys. Today 47:58–64, 1994.
5. Karplus, M., Shakhnovich, E. Protein folding: Theoretical studies of thermodynamics and dynamics. In "Protein Folding." Creighton, T.E. (ed.). New York: W.H.Freeman, 1992.
6. Protein Data Bank 1993 Release 67, Brookhaven National Laboratory.
7. Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wycoff, H., Phillips, D.C. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature, 181:662-666, 1958.
8. Perutz, M.F., Rossman, M.G., Cullis, A.F., Muirhead, H. Will, G., North, A.C.T. Structure of haemoglobin. A three-dimensional Fourier synthesis at 5.5 Å resolution obtained by x-ray analysis. Nature 185:416–422, 1960.
9. Scheraga, H. Calculation of stable conformations of polypeptides, proteins, and protein complexes. Chem. Scripta 29A:3–13, 1989.
10. Sali, A., Shakhnovich, E., Karplus, M. Kinetics of protein folding: A lattice model study of the requirements for folding to the native state. J. Mol. Biol. 235:1614–1636, 1994.
11. Anfinsen, C.B. Principles that govern the folding of protein chains. Science 181:223–230, 1973.
12. Kim, P., Baldwin R. Intermediates in the folding reactions of small proteins and the mechanism of protein folding. Annu. Rev. Biochem. 51:459–489,1990.
13. Karplus, M., Weaver, D.L. Protein folding dynamics: The diffusion-collsion model and experimental data. Prot. Sci. 3:650–668, 1994.
14. Levinthal, C. Are there pathways for protein folding? Chim. Phys. 65:44–45, 1968.
15. Sali, A., Shakhnovich, E. I., Karplus, M. How does a protein fold? Nature 369:248–251, 1994.
16. Brooks, C.L.I., Karplus, M., Pettit, B.M. "Proteins: A Theoretical Perespective of Dynamics, Structure and Thermodynamics." New York: John Wiley and Sons, 1988.
17. Bouzida, D., Kumar, S., Swendsen, R.H. Efficient Monte Carlo methods for computer simulation of biological molecule. Phys. Rev. A45:8894–8901, 1992.
18. Wilson, C., Doniach, S. A computer model to dynamically simulate protein folding: Studies with crambin. Proteins 6:193–209, 1989.
19. Kolinski, A., Skolnick, J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins 18:338–352, 1994.
20. Hinds, D.A., Levitt, M. A lattice model for protein structure prediction at low resolution. Proc. Natl. Acad. Sci. U.S.A. 89:2536–2540, 1992.

21. Hinds, D.A., Levitt, M. Exploring conformational space with a simple lattice model for protein structure. J. Mol. Biol. 243:668–682, 1994.
22. Dill, K.A., Feibig, K.M., Chan, H.S. Cooperativity in protein-folding kinetics. PNAS 90:1942–1946, 1993.
23. Shakhnovich, E.I., Gutin, A.M. A new approach to the design of stable proteins. Prot. Eng. 6:793–800, 1993.
24. Elber, R., Karplus, M. Multiple conformational states of proteins: Molecular dynamics of myoglobin. Science 235: 318–321, 1987.
25. Park, B.H., Levitt, M., The complexity and accuracy of discrete state model of protein structure. J. Mol. Biol. 249: 493–507, 1995.
26. Wodak S.J., Rooman M.J. Generating and testing protein folds, Curr. Opin. Struct. Biol. 3:247–259, 1993.
27. Rooman, M.J., Kocher, J.P., Wodak, S.J. Prediction of protein backbone conformation based on 7 structure assignments: Influence of local interactions. J. Mol. Biol. 221: 961–979, 1991.
28. Levitt, M. A Simplified representation of protein conformations for rapid simulations of protein folding. J. Mol. Biol. 104:59–107, 1976.
29. Ouzounis, C., Sander, C., Scharf, M., Schneider R. Prediction of protein structure by evaluation of sequence-structure fitness. J. Mol. Biol. 232:805–825, 1993.
30. Casari, G., Sippl, M.J. Structure derived hydrophobic potential. J. Mol. Biol. 224:725–732, 1992.
31. Bowie, J.U., Luthy, R., Eisenberg, D.A. Method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170, 1991.
32. Sippl, M.J. Calculation of conformational ensembles from potentials of mean force. J. Mol. Biol. 213:859–883, 1990.
33. Lifson, S., Sander, C. Antiparallel and parallel beta-strands differ in amino acid residue preferences. Nature 282:109–111, 1979.
34. Yee, D.P., Dill, K.A. Families and the structural relatedness among globular proteins. Prot. Sci. 2:884–899, 1993.
35. Godzik, A., Skolnick, J. Sequence-structure matching in globular proteins: Application to supersecondary and thretary structure determination. PNAS 89:12098–12102, 1992.
36. Holm, L., Sander, C. Protein structure comparison by alignment of distance matrices. J. Mol. Biol. 233:123–132, 1993.
37. Keasar, C., Elber, R. Homology as a tool in optimization problems: Structure determination of 2D heteropolymers. J. Phys. Chem. 99:11550–11556, 1995.
38. Monge, A., Lathrop, E.J., Gunn, J.R., Shenkin, P.S., Friesner, R.A. Computer modeling of protein folding: Conformational and energetic analysis of reduced and detailed protein models. J. Mol. Biol. 247:995–1012, 1995.
39. Chotia, C. Evolution of proteins formed by beta-sheets. J. Mol. Biol. 160:309–303, 1982.
40. Havel, T.F., Crippen, G.M., Kuntz, I.D. Effects of distance constraints on macromolecular conformation. II. Simulation of experimental results and theoretical predictions. Biopolymers 18:73–81, 1979.
41. Crippen, G.M. A novel approach to calculation of conformation: Distance geometry. J. Comput. Phys. 24:96–100, 1977.
42. Saitoh, S., Nakai, T., Nishikawa, K. A geometrical constraint approach for reproducing the native backbone conformation of a protein. Proteins 15:191–204, 1993.
43. Bohr, J., Bohr, H., Brunak, S., Cotterill, R.M., Fredholm, H., Lautrup, B., Petersen, S.B. Protein structures from distance inequalities. J. Mol. Biol. 231:861–869, 1993.
44. Holm, L., Sander, C. The FSSP database of structurally aligned protein fold families. Nucleic Acids Res. 22:3600–3609, 1994.
45. Orengo, C.A., Flores, T.P., Jones, D.T., Taylor, W.R., Thornton, J.M. Recurring structural motifs in proteins with different functions. Curr. Biol. 3:131–139, 1993.
46. Orengo, C. Classification of protein folds. Curr. Opin. Struct. Biol. 4:429–440, 1994.
47. Subbiah S., Laurents D.V., Levitt M. Structural similarity of DNA-binding domains of bacteriophage repressor and the globin core. Curr. Biology 3:141–148, 1993.
48. Jaritz, M., Sippl, M.J. Reported in Sippl, M.J. Knowledge-based potentials for proteins. Curr. Opin. Struct. Biol. 5:229–235, 1995.
49. Huang, E.S., Subbiah, S., Levitt, M. Recognizing native folds by arrangement of hydrophobic and polar residues. J. Mol. Biol. 252:709–720, 1995.
50. Sippl, M.J., Weitckus, S. Prediction of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. Proteins 13:258–271, 1992.
51. Crippen, G.M. Prediction of protein folding from amino acid sequence over discrete conformational space. Biochemistry 30:4232–4237, 1991.
52. Miyazawa, S., Jernigan, R. Estimation of effective inter-residue contact energies from protein crystal structures: Quasi-chemical approximation. Macromolecules 18:534–552, 1985.
53. Horovitz, A., Fersht, A.R. Co-operative interactions during protein folding. J. Mol. Biol. 224:733–740, 1992.
54. Daggett, V., Levitt, M. A model of the molten globule state from molecular dynamics simulation. PNAS 89:5142–5246, 1992.
55. Feller, W. "An Introduction to Probability Theory." New York: Wiley, 1968.