

# Toward an Energy Function for the Contact Map Representation of Proteins

Kibeom Park, Michele Vendruscolo, and Eytan Domany\*

*Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot, Israel*

**ABSTRACT** We analyzed several energy functions for predicting the native state of proteins from an energy minimization procedure. We derived the parameters of a given energy function by imposing the basic requirement that the energy of the native conformation of a protein is lower than that of any conformation chosen from a set of decoys. Our work is motivated by a recent result which proved that the simple pairwise contact approximation of the energy is insufficient to satisfy simultaneously such a basic requirement for all the proteins in a database. Here, we investigate the reasons of such negative results and show how to improve the predictive power of methods based on energy minimization. We generated decoys by gapless threading, and we derive energy parameters by perceptron learning. We first considered hydrophobic contributions to the energy, defined in several ways, and showed that the additional hydrophobic terms enlarge slightly the number of proteins that can be stabilized together. Next, we performed various modifications of the pairwise energy term. We introduced (1) a distinction between inter-residue contacts on the surface and in the core of a protein and (2) a simple distance-dependent pairwise interaction in which a two-tier definition of contact replaces the original (single-tier) one. Our results suggest that a detailed treatment of the pairwise potential is likely to be more relevant than the consideration of other forces. *Proteins* 2000;40:237–248. © 2000 Wiley-Liss, Inc.

**Key words:** protein folding; contact maps; pairwise contact potential; hydrophobicity

## INTRODUCTION

Proteins play a central role in nearly all biological processes at the cellular level; they are responsible for catalyzing and regulating biochemical reactions, transporting material and information, and form the basic structures such as skin, hair, and tendon.<sup>1</sup> The biologic function of a protein is determined by its chemical composition of the molecule and by its spatial structure.<sup>2</sup> A protein is a linear chain of covalently bound amino acids, whose length ranges from tens to thousands of these basic monomers. The number of proteins whose sequences have been determined is about 200,000.<sup>3</sup>

Most proteins are believed to fold, under physiologic conditions, into a unique compact structure that is determined by its amino acid sequence.<sup>4</sup> Although the determi-

nation of the three-dimensional structure from the sequence has been one of the central problems in molecular biology for several decades, it has met, so far, with limited success.<sup>5</sup>

One of the most widely used methods aimed at solving the protein folding problem and making reliable structure predictions is that of energy minimization. Assuming that the molecule is at thermal equilibrium when in its native state, one actually has to minimize a *free energy*.<sup>4,6–8</sup> Clearly, to find the correct structure by energy minimization, it is a matter of paramount importance to determine the correct energetics. The use of rigorous or empirical calculations to determine the forces acting between amino acids in a vacuum or in solution encounters formidable computation difficulties.<sup>9,10</sup> Furthermore, it can be realized only within the framework of an atomistic description of amino acids, which, despite recent progresses,<sup>11,12</sup> is still not practically feasible for folding simulations.

An alternative to the atomistic description is to introduce simplified or *reduced representations* of a protein's structure.<sup>13–16</sup> In these studies, amino acids are usually represented by one or more interacting units which may also have some internal degrees of freedom. Usage of such simplified representations of proteins raises, however, a very serious problem, namely how to determine the proper (free) energy function. The question is: What kinds of "interactions" between these simplified coarse-grained units could make native structures correspond to energy minima?

Several groups have contributed to the effort of giving an answer to such questions, by using different kinds of reduced representations,<sup>17–24</sup> but no satisfactory solution has been found yet. In previous work, we have used contact maps as the representation of a protein's structure.<sup>6,8,25,26</sup> The natural guess for an energy function is based on *pairwise interactions* between a pair of amino acids that *are in contact*. Such an energy function is made specific by choosing a definition of contact and by a set of *contact energy parameters*. Once the energy function is specified,

---

Grant sponsors: The Minerva Foundation; Germany-Israel Science Foundation (GIF); US-Israel Science Foundation (BSF); European Molecular Biology Organization (EMBO); Israeli Ministry of Science.

Dr. Vendruscolo's present address is Oxford Centre for Molecular Sciences, New Chemistry Laboratory, South Parks Road, OX1 3QT Oxford, United Kingdom.

\*Correspondence to: Eytan Domany, Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel. E-mail: fedomany@weizmann.weizmann.ac.il

Received 7 December 1999; Accepted 2 March 2000

one tries to find the map whose energy is the lowest among an ensemble of alternative structures, called *decoys*.

The possible approaches to the problem differ at this juncture by the manner in which the decoys are generated. One possibility is to execute a search in the entire space of physical contact maps, looking for the maps of lowest energy.<sup>27</sup> Because it is difficult to identify physical maps, an alternative is to construct decoys by *threading* the given sequence through the known structures of some longer proteins.<sup>28–31</sup> With either approach, a relevant question to ask is whether it is possible at all to find a set of contact energy parameters for which the native map of one protein (or more) is that of lowest energy.

The answer to this question depends on the decoys against which the native map is tested. For decoys obtained by gapless threading, the answer is positive for a single protein. On the other hand, when one generates decoys in a more painstaking and computationally expensive way, the answer is negative. It has been proved that the simple pairwise contact energy function is not capable of assigning the lowest energy to the experimentally determined structure, when tested against a sufficiently large set of carefully selected decoys.<sup>25,31</sup> The only structural information that was used about the decoys is their contact map. By “carefully selected” decoys we mean, first and foremost, that their contact maps correspond to real chain configurations, i.e., the distances between consecutive units are within a narrow window, the chain does not self-intersect (i.e., the distance between two nonconsecutive units exceeds a threshold), etc. Throughout the protein literature, decoys produced by threading are considered to be physical, because the conformation of the  $C_\alpha$  atoms of the sequence that is being tested coincides with the conformation of the backbone of a real polypeptide chain, taken from the data bank of known structures. This backbone-based definition of a physical decoy was adopted in<sup>25</sup> and in<sup>32</sup>; only decoys whose  $C_\alpha$  atoms constitute a physically realizable backbone were admitted, and the corresponding contact map was referred to as a “physical map.” To prove that it is not possible to stabilize the native map of a single protein within the contact approximation, a large number of decoy maps had to be generated,<sup>25</sup> which were all physical *and* of low (contact) energy.

It should be noted that there may be a problem with this generally accepted definition of “physical” structures. One can easily imagine a situation in which threading the backbone of sequence *A* into the experimentally determined backbone of sequence *B* produces a conformation that cannot be realized. This can happen when we thread the  $C_\alpha$  atom of a large residue of *A* (say Tyrosine) into a position that was occupied by the  $C_\alpha$  of a small residue (say Glycine) of *B*. In this case, even though the backbone conformation *is* physical, an attempt to replace the side chains of *B* by those of *A* may violate *steric constraints* associated with the side chains. Decoys obtained by threading, as well as by the method that was used to generate physical maps,<sup>25</sup> suffer from this problem.

The proof mentioned above established that within the widely accepted, backbone-based definition of physical structures one cannot stabilize the native map by any choice of the contact energies. In principle, one cannot rule out the possibility that if the decoy maps are modified to satisfy also the steric constraints associated with the side chains, stabilization of the native map will become possible. We are planning to show that this is not the case and believe that, even when all steric constraints are satisfied, the native map cannot be stabilized by a pairwise contact energy; such an energy function is too simple to stabilize proteins.

Many forces have been studied and suggested to be of central importance for the formation of a stable structure, such as the long-range electrostatic interactions, hydrogen bonding, van der Waals interactions, and hydrophobicity.<sup>2</sup> Even though we are trying to construct a phenomenologic *free* energy function, some reflection of all these may be needed to be taken account to obtain minimal energy maps that approximate better the native one.

Our long-term aim is to test systematically which of these terms is more important to stabilize native maps as those of minimal energy. To do this, we add one by one various terms, described in the next section, to the energy function and study the improvement induced by each. The manner in which we measure the improvement induced follows a methodology introduced in an earlier publication<sup>26</sup> and will be explained in the Learning Energy Parameters by Perception section. The basic idea is to require that the (known) native maps of a set of proteins have lower energy than all the respective decoy maps one produces. This *basic requirement* takes the form of a large set of *inequalities* that need to be satisfied.<sup>26,33,34</sup> For better energy functions and associated parameters one expects to be able to stabilize more proteins against larger sets of decoys.

The first extra terms that are added to the energy function in this study represent hydrophobicity. Many studies have reported the central role of hydrophobic effects in protein folding.<sup>2,20,34–36</sup> The general rule is that hydrophobic residues tend to be closely packed in the interior of proteins, whereas hydrophilic ones are mostly exposed to the solvent. However, the exact functional form of the hydrophobic interactions is still unknown and under debate.<sup>37,38</sup> We have proposed several ways of introducing hydrophobicity, and present, in the Effects of Hydrophobic Energy section, their effect on the size of the set that can be stabilized.

Next, we turned to investigate the effect of improving and generalizing the pairwise contact energy function. First, the original pairwise term gives the same value of energy regardless of whether the interaction occurs inside the core of the globular proteins or on their surface. We considered the difference by dividing the pairwise term into two classes; occurring inside and on the surface. This consideration reflects the fact that when a pair of amino acids resides on the protein’s surface it is surrounded by water molecules, whose presence affects the residue-residue interaction.

Another obvious modification that we studied is to allow the energy to depend on the distance between the two residues. This is done by differentiating the definition of *contact*, replacing the all-or-none nature of the widely-used definition by a two-tiered one. These energy terms are introduced in the next section. The results for the modified pairwise energy function are presented in the Modifications of Pairwise Energy Function section. Finally, we briefly summarize our results in the Discussion section.

### REPRESENTATION OF STRUCTURE AND ENERGY FUNCTION

The contact map of a protein with  $N$  amino acids is an  $N \times N$  matrix  $\mathbf{S}$ , whose elements are defined as

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

The contact between two residues can be defined in different ways. In particular, we will consider the “ $C_\alpha$ ” definition, in which two amino acids are considered in contact when their  $C_\alpha$  atoms are closer than some threshold distance  $R_c$ . Denoting the positions of the  $C_\alpha$  atoms of residues  $i$  and  $j$  by  $\mathbf{r}_i$  and  $\mathbf{r}_j$ , the definition (1) becomes

$$S_{ij} = \begin{cases} 1 & \text{if } |\mathbf{r}_i - \mathbf{r}_j| < R_c \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

The effect of varying  $R_c$  has been investigated previously.<sup>31,33,39</sup> This definition of contact can be modified in various ways. For example, one may use a different  $R_c$  for every pair of amino acids, or use for an interacting pair of residues  $a, b$  a value  $R_c(a, b) = R_c(a) + R_c(b)$ . Work on these generalizations is under way and will be published elsewhere (Park K, Park C-W, Domany E, unpublished results).

The generalization we studied here is described below; first, we define the contact energy of a conformation. Denote the amino acids sequence of a protein of length  $N$  by

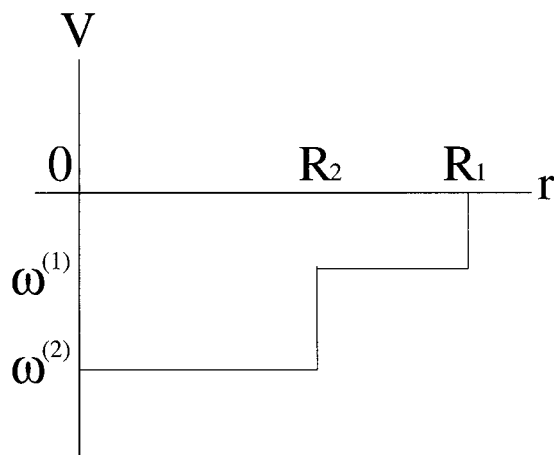
$$\mathbf{A} = (a_1, a_2, \dots, a_N).$$

The simplest, most widely used pairwise contact energy is given by

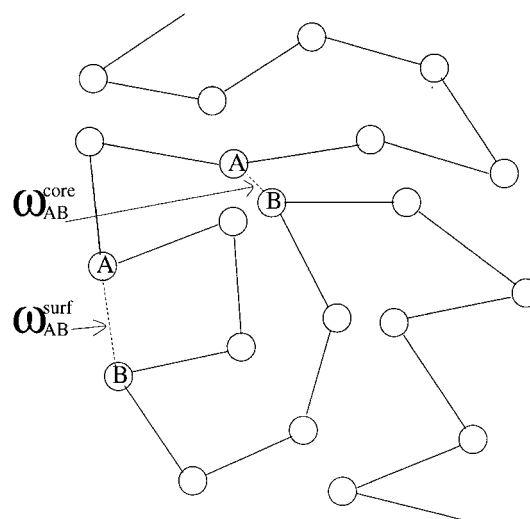
$$E^{\text{pair}}(\mathbf{A}, \mathbf{S}, \omega) = \sum_{i < j} \omega_{a_i a_j} S_{ij}, \quad (3)$$

where the 210 parameters,  $\omega_{a_i a_j}$ , which constitute a  $20 \times 20$  symmetric matrix, represent the contact energy between amino acids of type  $a_i$  and  $a_j$ . It should be stressed that according to our philosophy, Equation (3) is a lowest order approximation to a *phenomenologic free energy* for sequence  $\mathbf{A}$  having contact map  $\mathbf{S}$ . A less coarse-grained representation of the structure is in terms of a two-tiered contact map:  $\mathbf{S}^{(\mu)}$  with  $\mu = 1, 2$ :

$$S_{ij}^{(1)} = \begin{cases} 1 & \text{if } |\mathbf{r}_i - \mathbf{r}_j| < R_2 \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$



(a)



(b)

Fig. 1. **a:** Two-tiered pairwise potential  $V$  as a function of distance  $r$ . **b:** Schematic illustration of  $\omega_{AB}^{\text{core}}$  and  $\omega_{AB}^{\text{surf}}$  for a specific pair of amino acids A and B.

$$S_{ij}^{(2)} = \begin{cases} 1 & \text{if } R_2 < |\mathbf{r}_i - \mathbf{r}_j| < R_1 \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

The pairwise (free) energy function associated with this representation is

$$E(\mathbf{A}, \mathbf{S}, \omega) = \sum_{i < j} (S_{ij}^{(1)} \omega_{a_i a_j}^{(1)} + S_{ij}^{(2)} \omega_{a_i a_j}^{(2)}). \quad (6)$$

This form of the energy can be viewed as an approximation to a distance-dependent continuous function. With this form, one uses 420 contact energy parameters. Clearly, the energy function (3) is a particular case of (6). We have drawn schematically a two-tiered pairwise potential  $V$  as a function of the distance  $r$  in Figure 1a.

An entirely different modification of the contact energy, which also uses 420 parameters, takes into account the *environment* in which the pair of amino acids resides. In particular, it allows the same pair to interact differently if it is *inside the core* of the protein and if it is *on the surface*. In the latter case, the pair of residues is surrounded mainly by water molecules, whose presence may affect the interaction. The resulting energy function is

$$\omega_{a_i a_j} = \begin{cases} \omega_{a_i a_j}^{\text{core}} & \text{if amino acids } a_i \text{ and } a_j \text{ lie in the core,} \\ \omega_{a_i a_j}^{\text{surf}} & \text{otherwise,} \end{cases} \quad (7)$$

as is illustrated in Figure 1b. To implement this idea one has to decide, on the basis of the contact map, whether the pair  $(a_i, a_j)$  is on the surface or buried in the core. This position is decided on the basis of the number of contacts these amino acids have, as explained in detail in the Modifications of Pairwise Energy Function section.

One can use either of these three pairwise energy functions. In addition, as also discussed by other authors,<sup>36,34,20</sup> *hydrophobicity* terms were introduced as well. Whereas the pairwise contact terms depend on the identity of a pair of amino acids, the hydrophobic energy depends on the identity of each single amino acid. It has the form

$$E^{\text{hydro}}(\mathbf{A}, \mathbf{S}, \beta) = \sum_i^N \beta_{a_i} \left[ \sum_{k=1}^N S_{ik} - n_{a_i} \right]^2. \quad (8)$$

The parameters  $\beta_{a_i}$  and  $n_{a_i}$  depend on the identity of the  $i$ th amino acid. Clearly,  $\sum_k S_{ik}$  is the number of contacts which the  $i$ th amino acid has on a given contact map. The contribution of any amino acid  $u$  to  $E^{\text{hydro}}$  is to be minimal when the number of its contact is equal to the optimal value  $n_u$ . Hydrophobic amino acids are expected to have relatively larger values of  $n_u$  than hydrophilic ones. It is also important to note here that the parameter  $\beta_{a_i}$  should be positive.

### LEARNING ENERGY PARAMETERS BY PERCEPTRON

We want to determine energy parameters that satisfy the basic requirement presented in the Introduction section. That is, we express the energy as a linear function of various parameters and require that each one of a set of native maps has lower energy than all its corresponding decoys. This set of requirements reduces to a set of inequalities, and we try to find solutions for these by the perceptron learning rule. Even though this method was reviewed elsewhere,<sup>31</sup> we repeat it here because the hydrophobic term introduces a subtlety that was not dealt with before.

For concreteness, we describe the method for a combination of the simple pair energy of Equation (3) and the hydrophobic term Equation (8). Generalization to the other energy functions discussed above is straightforward.

#### Generating Linear Inequalities

First, we express the energy for any map  $\mathbf{S}^\mu$  as a linear function of the energy parameters, written as

$$\begin{aligned} E &= E^{\text{pair}}(\mathbf{A}, \mathbf{S}^\mu, \omega) + E^{\text{hydro}}(\mathbf{A}, \mathbf{S}^\mu, \beta) \\ &= \sum_{c=1}^{210} \omega_c N_c(\mathbf{S}^\mu) + \sum_{d=1}^{20} \beta_d M_d(\mathbf{S}^\mu), \end{aligned} \quad (9)$$

where  $N_c(\mathbf{S}^\mu)$  is the total number of contacts of type  $c$  and

$$M_d(\mathbf{S}^\mu) \equiv \sum_{i=1}^N \left[ \sum_{j \neq i} S_{ik}^{\mu} - n_d \right]^2 \delta(a_i, d). \quad (10)$$

Here the matrix  $\omega_{ij}$  has been rewritten as the corresponding 210-component vector  $\omega_c$ , and  $\delta(a_i, d)$  takes the value 1 if  $a_i = d$  and 0 otherwise. Note that the energy of a map is a linear function of the parameters  $\omega$  and  $\beta$ .

Suppose that we choose one protein with a known native map  $\mathbf{S}^0$  and generate a large set of decoys for it. In this work, as in Ref. 31, decoys were generated by gapless threading (see below).

The energy difference between  $\mathbf{S}^0$  and one particular decoy  $\mathbf{S}^\mu$  can be written in the form

$$\begin{aligned} \Delta E_\mu &= \sum_{c=1}^{210} \omega_c x_c^\mu + \sum_{c=1}^{20} \beta_c y_c^\mu \\ &\equiv \omega \cdot \mathbf{x}^\mu + \beta \cdot \mathbf{y}^\mu \\ &\equiv \mathbf{u} \cdot \mathbf{z}^\mu, \end{aligned} \quad (11)$$

where we have used the notation

$$\begin{aligned} x_c^\mu &\equiv N_c(\mathbf{S}^\mu) - N_c(\mathbf{S}^0), \\ y_c^\mu &\equiv M_c(\mathbf{S}^\mu) - M_c(\mathbf{S}^0), \end{aligned}$$

and introduced the following 230-component vectors

$$\begin{aligned} \mathbf{z} &= (x_1, x_2, \dots, x_{210}, y_1, y_2, \dots, y_{20}), \\ \mathbf{u} &= (\omega_1, \omega_2, \dots, \omega_{210}, \beta_1, \beta_2, \dots, \beta_{20}). \end{aligned}$$

Note that the basic requirement, of all native maps having a lower energy than their respective decoys  $\mu$ , takes the form

$$h_\mu \equiv \mathbf{u} \cdot \mathbf{z}^\mu > 0. \quad (12)$$

This linear inequality should be satisfied for all the decoy-native fold combinations used. The vector  $\mathbf{u}$  contains the parameters of the energy function that are to be found, and the (known) vector  $\mathbf{z}^\mu$  is determined by the native and decoy contact maps.

#### Perceptron Learning

Perceptron learning is a simple method to look for a solution  $\mathbf{u}$  of these  $\mu = 1, 2, \dots, P$  inequalities.<sup>40,41</sup> The vectors  $\mathbf{z}^\mu$  are referred to as “patterns” in the perceptron literature. The vector of energy parameters  $\mathbf{u}$  is *learned* in the course of a training session. The  $P$  patterns are presented cyclically; after presentation of the  $\mu$ th pattern, the weights  $\mathbf{u}$  are updated according to the following rule:

$$\mathbf{u}'_{\bar{0}} \begin{cases} (\mathbf{u} + \eta \mathbf{z}^\mu) / |\mathbf{u} + \eta \mathbf{z}^\mu| & \text{if } \mathbf{u} \cdot \mathbf{z}^\mu < 0, \\ \text{otherwise.} & \end{cases} \quad (13)$$

Different choices are possible for the parameter  $\eta$ . We use the learning rule introduced in Ref. 42, in which the parameter  $\eta$  is updated as

$$\eta = \frac{-h_\mu + 1/d}{1 - h_\mu/d}, \quad (14)$$

together with the *despair* parameter  $d$ , which evolves as

$$d^{\text{new}} = \frac{d + \eta}{\sqrt{1 + 2\eta h_\mu + \eta^2}}, \quad (15)$$

where both the vectors  $\mathbf{u}$  and  $\mathbf{z}^\mu$  are normalized (see below for details on normalization).

The training session terminates with two possible outcomes: if the set of inequalities has a solution, the algorithm finds one in a finite number of steps; if there is no solution, unlearnability is detected. Denote the magnitude of the vector  $\mathbf{z}^\mu$  by

$$\begin{aligned} Z^\mu &= \sum_{c=1}^{230} (z_c^\mu)^2 \\ &= \sum_{c=1}^{210} (x_c^\mu)^2 + \sum_{c=1}^{20} (y_c^\mu)^2 \\ &\equiv X^\mu + Y^\mu. \end{aligned}$$

It was shown in Nabutovsky and Domany<sup>42</sup> that the problem is unlearnable, if the *despair* parameter  $d$  exceeds a critical threshold

$$d_c = \sqrt{M(2Z^{\text{max}})^{M/2}},$$

where  $M$  is the number of components of  $\mathbf{u}$  and  $Z^{\text{max}}$  is the maximal value of  $Z^\mu$ .

It is trivial to see that if  $\mathbf{u}^* = (\omega^*, \beta^*)$  is a solution for the training set  $(\mathbf{x}^\mu, \mathbf{y}^\mu)$ , then  $\mathbf{u}_\lambda^* = (\omega^*, \lambda\beta^*)$  is a solution for the training patterns  $(\mathbf{x}^\mu, \mathbf{y}^\mu/\lambda)$ . The outcome of the learning process does not depend on the choice of  $\lambda$ , whereas the learning time does depend strongly on  $\lambda$ . We have checked the dependence of learning time on  $\lambda$ , and found that the most effective value of  $\lambda$  is such that makes  $X^\mu$  and  $Y^\mu$  be the same order. We have then divided both  $\mathbf{x}^\mu$  and  $\mathbf{y}^\mu$  by the average values of  $X^\mu$  and  $Y^\mu$ , respectively,

$$\begin{aligned} \mathbf{x}_{\text{res}}^\mu &= \frac{1}{\sqrt{X_{\text{ave}}^\mu}} \mathbf{x}^\mu, \\ \mathbf{y}_{\text{res}}^\mu &= \frac{1}{\sqrt{Y_{\text{ave}}^\mu}} \mathbf{y}^\mu, \end{aligned}$$

for each example, and take the normalized  $\mathbf{z}^\mu$  as

$$\mathbf{z}_{\text{norm}}^\mu = \left( \sqrt{21/23} \mathbf{x}_{\text{res}}^\mu, \sqrt{2/23} \mathbf{y}_{\text{res}}^\mu \right) / \left( \frac{21}{23} \left| \mathbf{x}_{\text{res}}^\mu \right|^2 + \frac{2}{23} \left| \mathbf{y}_{\text{res}}^\mu \right|^2 \right)^{1/2}.$$

In addition, we introduce the fictitious examples  $\tilde{\mathbf{z}}^\nu$ , ( $\nu = 1, 2, \dots, 20$ ), which are vectors of zeros except at one component:  $\tilde{z}_i^\nu = \delta_{i,\nu}$ . Then the added conditions

$$h_\nu = \mathbf{u} \cdot \tilde{\mathbf{z}}^\nu > 0 \quad (16)$$

guarantee that  $\beta_\nu$  takes a positive value.

Although perceptron learning is suitable for the present purpose, it is not the only way to prove unlearnability. Other techniques, such as linear programming<sup>43,44</sup> and in particular the simplex method,<sup>45</sup> are in principle available to solve this problem (see also the method used by Settanni and coworkers<sup>34</sup>).

### Generating Decoys from Threading Databases

The decoys were generated by means of gapless threading, which is done straightforwardly by using the contact map representation as follows. First, a database of proteins of known structure is compiled. Once the contact maps of the proteins in the database are obtained, we generate decoys for a given sequence of length  $N$  from the structures of all proteins of length  $N' (> N)$ , by selecting submaps of size  $N \times N$  along the main diagonal of the contact map of the longer proteins. Thus, the total number of decoys  $P$  generated by means of gapless threading is set by the number  $M_p$  of proteins in the database and by their lengths  $N_i$ , ( $i = 1, 2, \dots, M_p$ ), given by

$$P = \sum_{N_j \geq N_i} (N_j - N_i + 1).$$

Clearly, the more proteins we have in our database, we will have more conditions to satisfy and the learning problem becomes more difficult.

In this work, we used the datasets of Ref. 31, which we review here briefly. We start with a list of 312 proteins obtained by WHATCHECK.<sup>46</sup> From these, we select SET<sub>154</sub> by eliminating proteins that have any of following properties: 1) The  $C_\alpha$  distance between consecutive residues lies outside the interval of four standard deviations  $\sigma$  from the mean  $d$  ( $d = 3.81$ ,  $\sigma = 0.05$ ). 2) Any residue that does not match the 20 standard amino acid types. When the first or the last residue is undefined, we remove the residue, not the protein. 3) Any chain for which the  $C_\alpha$  or the backbone  $N$  atoms' coordinates are not present. 4) Any unexplained mismatch between the sequence of amino acids presented in SEQRES and the actual sequence appearing in the coordinates section.

### EFFECTS OF HYDROPHOBIC ENERGY

We turn now to describe our results when the hydrophobic term is added to the simple pairwise contact energy function, e.g., to Equation (3). The parameters  $n_{a_i}$  that enter in Equation (10) were derived from a statistical analysis of the known set of native maps, for every value of  $R_c$ . For each amino acid type, we calculated the frequency of the number of contacts in a set of native structures. Some of these histograms are presented in Figure 2. The mean values and standard deviations for the number of contacts of different amino acids, as obtained from the  $C_\alpha$  definition with  $R_c = 8 \text{ \AA}$ , are presented in Table I. Note

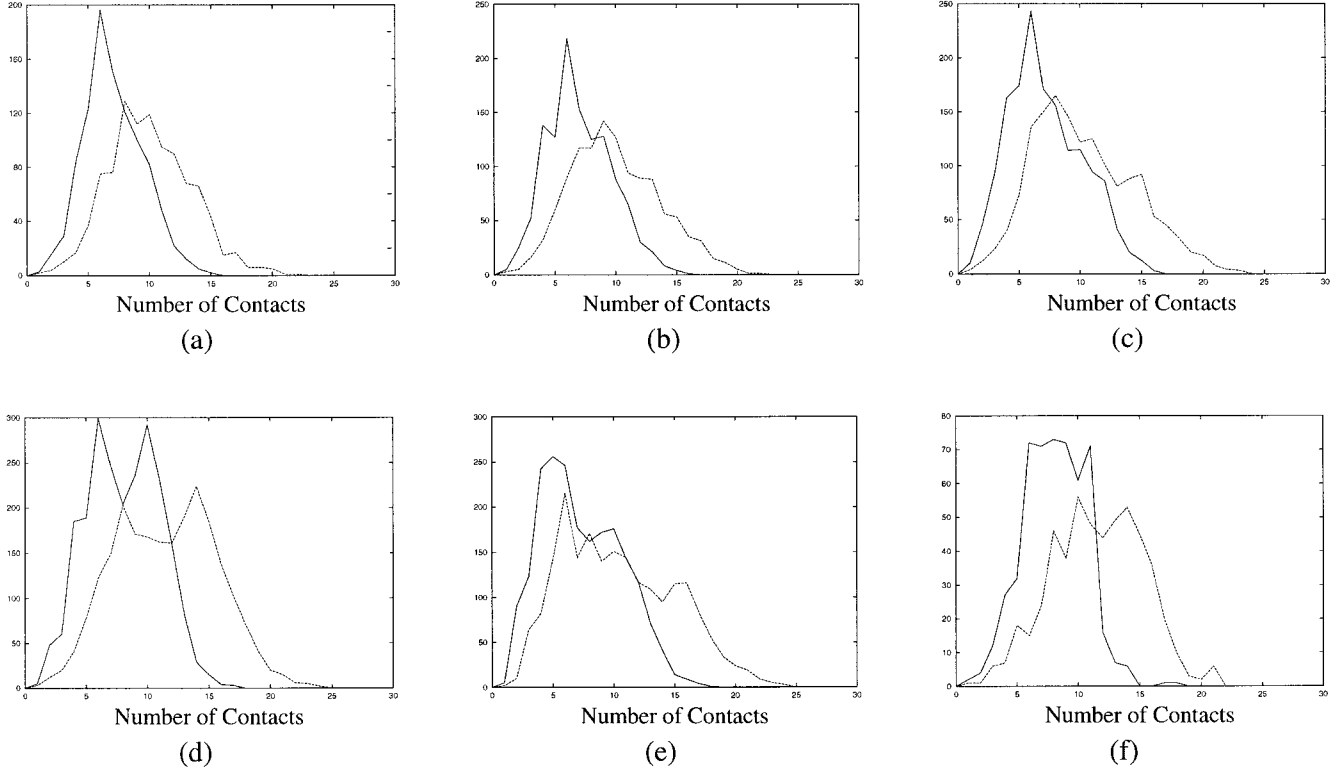


Fig. 2. Number density distributions of the number of contacts for amino acids (a) GLN, (b) ASN, (c) SER, (d) ALA, (e) GLY, and (f) MET. Data are obtained from the whole PDB files of 154 proteins. Solid and dashed lines represent the cases  $R_c = 8 \text{ \AA}$  and  $9 \text{ \AA}$ , respectively.

**TABLE I. Means and Standard Deviations of the Number of Contacts of Amino Acids**

Amino Acid	ALA	GLU	GLN	ASP	ASN	LEU	GLY	LYS	SER	VAL
Mean	8.1	6.4	7.1	6.4	7.0	8.5	7.3	6.7	7.2	8.8
SD	2.9	2.3	2.4	2.5	2.6	2.4	3.3	2.3	3.0	2.6
Amino Acid	ARG	THR	PRO	ILE	MET	PHE	TYR	CYS	TRP	HIS
Mean	7.2	7.6	6.8	8.8	8.2	8.4	8.4	9.6	8.2	7.5
SD	2.4	2.8	2.9	2.5	2.5	2.4	2.7	2.7	2.6	2.6

that a similar analysis, performed in Ref. 8, was based on an all-atom definition of contact (see below).

As expected, because hydrophobic residues tend to be closely packed in the interior of globular proteins, whereas hydrophilic ones are mostly exposed to the exterior, the average number of contacts is higher for the hydrophobic residues. To test quantitatively this statement, we calculated the Pearson correlation coefficients between the mean value of contacts for each amino acid and the hydrophobic indices used by other authors, which is presented in Table II.

The mean value of the number of contacts shows a reasonable correlation with any of the other, previously used hydrophobicity indices, and we turned to determine the energy parameters  $\omega_{u,v}$  and  $\beta_u$  by learning training sets of increasing number of proteins,  $M_p$ , and varying  $R_c$  as well. For the pairwise contact energy only (i.e., without the hydrophobic term) the dependence of learnability on  $M_p$  and  $R_c$  has been studied in Ref. 31. The main result of

that study is that there exists a finite region in the  $(R_c, M_p)$  plane in which the problem is learnable. Our purpose in this work is to investigate how the size of this region is affected by addition of the hydrophobic energy term. We present in Figure 3 both “phase boundaries,” obtained with and without the hydrophobic term. We denote by diamonds learnable cases and by crosses the unlearnable cases.  $SET_{154}$  is found unlearnable at  $R_c = 11, 12,$  and  $13 \text{ \AA}$ . A subset  $SET_{141}$ , obtained by removing the 13 *worst* proteins, which are responsible for the majority of mismatches, is learnable in the region  $11 \leq R_c \leq 13 \text{ \AA}$ , where we have put diamonds. Next, we selected a subset  $SET_{123}$  in the same way, and have found that it is learnable in the region  $8 \leq R_c \leq 16 \text{ \AA}$ . Finally, we have drawn a solid line, connecting the mid-points between crosses and diamonds, which represents the border of the region of learnability. The previous result<sup>31</sup> without the hydrophobic term, is given, for comparison, by the dashed line.

TABLE II. Pearson Correlation Coefficients Among Several Hydrophobic Indices<sup>a</sup>

	CS	KD	OMH	SE	Eng	MD	$C_{\alpha}^{\dagger}$	$C_{\alpha}^{\ddagger}$	AA
CS		0.74	0.74	0.73	-0.74	0.77	0.93	0.95	0.81
KD			0.75	0.88	-0.85	0.50	0.83	0.82	0.54
OMH				0.72	-0.68	0.83	0.79	0.80	0.85
SE					-0.93	0.51	0.71	0.72	0.54
Eng						-0.45	-0.73	-0.74	-0.49
MD							0.70	0.73	0.99
$C_{\alpha}^{\dagger}$								0.99	0.76
$C_{\alpha}^{\ddagger}$									0.79

<sup>a</sup>Our indices are obtained from the  $C_{\alpha}$  definition with  $R_c = 8 \text{ \AA}$  ( $C_{\alpha}^{\dagger}$ ),  $R_c = 9 \text{ \AA}$  ( $C_{\alpha}^{\ddagger}$ ), and the All-Atom definition with  $R_c = 5 \text{ \AA}$  (AA). Details of others are as follows: CS,<sup>47</sup> KD,<sup>48</sup> OMH and SE,<sup>49</sup> Eng,<sup>50</sup> and MD.<sup>8</sup> Note that the index Eng is anticorrelated with others, reflecting *hydrophilicity*.

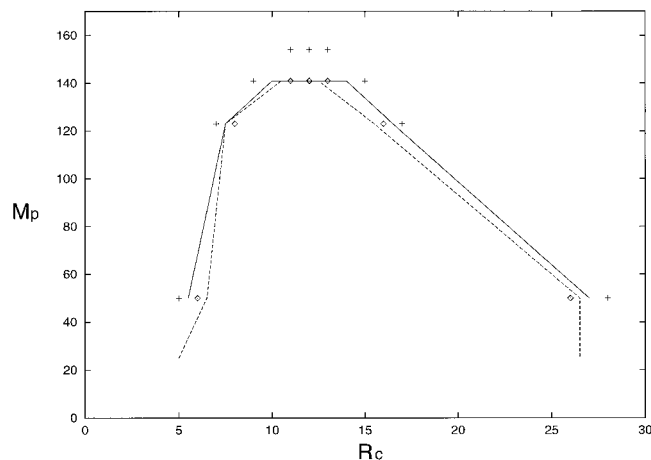


Fig. 3. Region of learnability on the plane of the number of proteins and the value of  $R_c$  with the hydrophobic energy function given in Equation (8). The previous result with the simple pairwise energy function is given by the dashed line.

It is evident that the inclusion of the hydrophobic energy enlarges the learnable region only marginally. This indicates that the hydrophobic energy of the form of Equation (8) does not improve significantly the quality of the phenomenologic free energy we try to derive. In the remainder of this section, we present our attempts to understand this result.

One possible reason could be that our form of the hydrophobic part of the free energy is too restrictive. Note in particular that we added 20 conditions (see Eq. 16) to ensure positivity of the parameters  $\beta_u > 0$  for all amino acids. The basis for this requirement is clear—we assumed that each amino acid prefers a certain number of contacts and the “energy” takes its minimal value when this number is attained by the map. This assumption was reflected in the original work that introduced this hydrophobic energy<sup>8</sup> by approximating the the frequency distributions  $f^{\mu}(n)$  by having  $n$  contacts, by Gaussians with different centers and widths for different amino acid types, e.g.,

$$f^{\mu}(n) \propto \exp\left[-\frac{(n-n_{\mu})^2}{2\sigma_{\mu}^2}\right], \quad (17)$$

which naturally led to  $\beta_u = 1/(2\sigma_u^2)$ .

We looked carefully at the distributions of contacts as obtained from known native structures taken from the PDB and, to our surprise, found that for certain amino acids the distribution is *bimodal*, as can be seen in Figure 2d, e, and f. One might think that this is an artifact of the  $C_{\alpha}$  definition of contact, which lacks information on side chains. However, when we adopted the “All-Atom” (AA) definition of contact (in which two amino acids are considered in contact when any two heavy atoms (excluding hydrogens) that belong to the two residues are at a distance lower than  $R_c$ ), we still found bimodality, although it is not as pronounced as in the case of the  $C_{\alpha}$  definition.

For amino acids having a bimodal distribution, the most probable values for the number of contacts is *not* the average number; there is, therefore, no reason to expect that in these cases the “hydrophobic energy” is minimal when the number of contacts is equal to the average value. Keeping the form of Equation (8), we can accommodate the fact that energy increases as the number of contacts approaches the mean value by having a negative value of  $\beta$ . Having a negative  $\beta_i$  in Equation (8) appears, at the first sight a dangerous predicament, since a configuration may gain energy by increasing the contacts of amino acid  $i$  beyond reason. However, we are protected from such an instability by the fact that all our maps are taken from segments of real, physical chains; hence, no configurations with an unruly large number of contacts will ever enter our calculations of the energy. Hence, we treated the effect of removing the positivity constraint from the  $\beta_i$ .

Our results indicated that the sign of the parameters  $\beta$  is *not* crucial. We found that the phase boundary, obtained by using the same process as before, did not change much by omitting the extra conditions Equation (16). The main change was that by allowing negative values of  $\beta$  the SET<sub>141</sub> became learnable for  $R_c = 8$  and  $9 \text{ \AA}$ .

At this point, we considered another possibility, namely that our entire notion of approximating the contact free energy as a *sum of two terms* in Equation (9) is at fault. Our calculations of the parameters of the hydrophobic contribution to the energy are based on a statistical analysis of known structures. In turn, the analysis is based on the assumption that the probability  $P$  of a microscopic

conformation follows the Boltzmann distribution. In the contact map representation, this fact is expressed as

$$P(\mathbf{S}, \mathbf{A}) \propto \exp[-H_{\text{eff}}(\mathbf{S}, \mathbf{A})] \quad (18)$$

where the dependence on the *temperature*  $T$  is absorbed in the parameters of the phenomenologic free energy or effective Hamiltonian  $H_{\text{eff}}(\mathbf{S}, \mathbf{A})$ . The exact form of the Hamiltonian  $H_{\text{eff}}$  is unknown and we have approximated it as  $E$  in Equation (9). The problem arises from the fact that  $E^{\text{pair}}$  is also dependent on the variable  $n$ . As a consequence, Equation (17) may be inadequate, because the hydrophobic effects may be double counted in this approach (or already included, to a large extent, in the pairwise terms). If so, even if there exists a hydrophobic energy term of the form of Equation (8), one should *not* determine the parameters  $n_{a_i}$  using simple statistics on PDB-based contact maps. Rather, one should relax also the restriction of externally fixed  $n_{a_i}$  and to rewrite the hydrophobic energy function in Equation (8) as follows:

$$\begin{aligned} E^{\text{hydro}} &= \sum_{i=1}^N \beta_{a_i} [N_i^2 - 2n_{a_i} N_i + n_{a_i}^2] \\ &= \sum_{i=1}^N [\gamma_1(a_i) N_i^2 + \gamma_2(a_i) N_i + \gamma_3(a_i)], \end{aligned} \quad (19)$$

where  $N_i \equiv \sum_j S_{ij}$ . By this step we relaxed restrictions on the hydrophobic energy and also added 20 more parameters to the energy function, making solution of the inequalities Equation (12) easier. We note that since we are dealing with differences in free energies between the native state and alternative conformations, the parameters  $\gamma_3$ , which do not depend on the conformation, do not appear in the basic requirement. The new form of  $E^{\text{hydro}}$

$$E^{\text{hydro}} = \sum_{i=1}^N [\gamma_1(a_i) N_i^2 + \gamma_2(a_i) N_i], \quad (20)$$

does not necessarily have a minimum when  $N_i = n_{a_i}$ ; therefore, we have freed the parameters  $n_{a_i}$  from their original meaning of being the average number of contacts of a residue of species  $a_i$ .

However, after carrying out the calculations, also in this case we found that replacing the  $\beta_i$  by the two independent 20-component vectors  $\gamma_1$  and  $\gamma_2$  *does not change* significantly the region of learnability!

The bottom line of these attempts is that a phenomenologic free energy term of the form (8) does not represent in a useful way the hydrophobic part of the free energy. Nevertheless, it is interesting to compare the obtained parameters  $(\beta_i, n_i)$  and  $(\gamma_1(i), \gamma_2(i))$ . Without imposing any conditions on  $\gamma_1$  and  $\gamma_2$ , it is found that some of the  $\gamma_1$  are negative and some of the  $\gamma_2$  are positive. The corresponding amino acid types are strongly related to those with bimodal distributions of the numbers of their contacts, even though not exactly coincide. We checked the correlations between  $(\beta_i, n_i)$  obtained without sign restriction (but

**TABLE III. Optimal Values of the Threshold  $N_{\text{th}}$  for Each Set of Proteins**

$R_c$ (Å)	6	6.5	7	10	12	13	15	17	18	19	20	22
SET <sub>123</sub>	3	5	6								84	100
SET <sub>141</sub>	3	5	6					60	67	75	83	
SET <sub>154</sub>				15	26	31	45					

with fixed values of  $N_i$ ) and  $(\gamma_1(i), \gamma_2(i))$ . We found that the parameters  $\gamma_1(i)$  and  $\beta_i$  are highly correlated with each other, whereas  $\gamma_2(i)$  and  $n_i$  are not. The correlation coefficient between  $\gamma_1(i)$  and  $\beta_i$  is 0.85 for  $(R_c, M_p) = (8 \text{ Å}, 123)$  and 0.69 for  $(11 \text{ Å}, 141)$ , whereas that between  $\gamma_2(i)$  and  $-2n_i\beta_i$  is 0.059 and 0.36 for  $(8 \text{ Å}, 123)$  and  $(11 \text{ Å}, 141)$ , respectively.

### MODIFICATIONS OF PAIRWISE ENERGY FUNCTION

We repeated the same procedure used for the hydrophobic energy to study the effect of introducing modified pairwise energy parameters, as given in Equations (6) and (7).

#### Differentiating Between Interactions in the Core Versus Surface

First, we used the modified contact energy given above, in Equation (7). The physical motivation for introducing this modification is obvious, i.e., the interaction between two residues is different when they are buried inside the protein's core and when the pair is on the macromolecule's surface. In the latter case, the two interacting amino acids are in contact also with water molecules, whose presence may modify the contact energy. Clearly, once we have a contact map we can determine whether any particular residue  $i$  is in the core or on the surface, just by *counting* the total number of contacts it has,  $n_i = \sum_k S_{ik}$ . Allowing the value of the energy to depend on this number amounts to taking into account terms beyond the simple pairwise contact approximation.

We have assumed that the amino acids lying on the surface have relatively smaller number of contacts than those in the core and regarded the contact between amino acids  $i$  and  $j$  as "in the core" if  $\sum_k (S_{ik} + S_{jk})$  is larger than a certain critical threshold  $N_{\text{th}}$ . We counted for many native structures the number of contacts on the molecule's surface and in its core and verified that indeed the average number of contacts that a surface residue has is smaller than that of a residue from the core.

We first thought that the precise value used for  $N_{\text{th}}$  may be of importance, but upon further investigation found that learnability, which is the main concern of this work, is affected very weakly by varying  $N_{\text{th}}$ , unless it is chosen to be too excessive. Thus, we have adopted a choice for which the total number of native contacts on the surface is about equal to the total number of native contacts in the core, because it gives rise to the fastest learning time. Detailed values of those  $N_{\text{th}}$ 's as a function of  $R_c$ , are given in Table III for each set of proteins.

Hence, the energy function for map  $\mathbf{S}$  and sequence  $\mathbf{A}$  takes the form

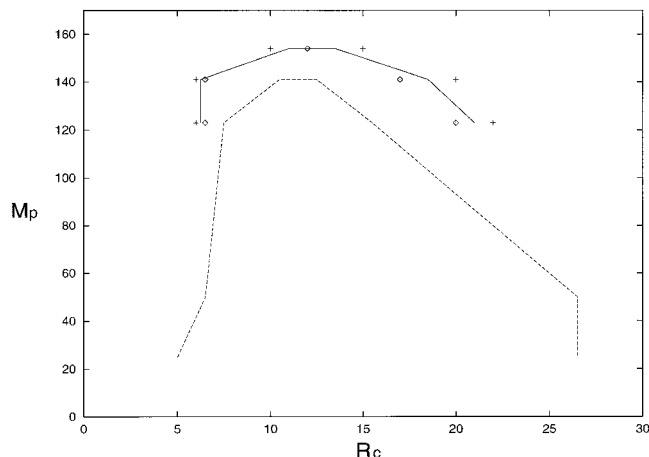


Fig. 4. Region of learnability on the plane of the number of proteins and the value of  $R_c$  with the hydrophobic energy function given in Equation (22). The previous result with the simple pairwise energy function is given by the dashed line.

$$E(\mathbf{A}, \mathbf{S}, \omega) = \sum_{i < j}^N [S_{ij} \theta(\sum_k (S_{ik} + S_{jk}) - N_{th}) \omega_{a_i a_j}^{core} + S_{ij} \theta(N_{th} - \sum_k (S_{ik} + S_{jk})) \omega_{a_i a_j}^{surf}], \quad (21)$$

where  $\theta(x)$  is the Heaviside theta function, which is 0 for  $x < 0$  and 1 otherwise.

As was done above, we can express the energy difference between any decoy map  $\mathbf{S}_\mu$  and the native map  $\mathbf{S}_0$  as a linear function of the energy parameters:

$$\Delta E_\mu = \sum_{c=1}^{210} \{ \omega_c^{core} [N_c^{core}(\mathbf{S}^\mu) - N_c^{core}(\mathbf{S}^0)] + \omega_c^{surf} [N_c^{surf}(\mathbf{S}^\mu) - N_c^{surf}(\mathbf{S}^0)] \}, \quad (22)$$

where  $N_c^{core}(\mathbf{S})$  and  $N_c^{surf}(\mathbf{S})$  denote the numbers of contacts of type  $c$  for a given map  $\mathbf{S}$ , which occur in the core and on the surface, respectively.

With this modified energy function, the region of learnability is remarkably enlarged. We found that the whole set of proteins  $SET_{154}$  becomes learnable at  $R_c = 12 \text{ \AA}$ . The windows of learnability are also extended to  $6.5 \leq R_c \leq 17$  in  $SET_{141}$ , and  $6.5 \leq R_c \leq 20$  in  $SET_{123}$  (see Fig. 4). The effect on learnability is much stronger than what we observed when the hydrophobic terms were added, see the Effects of Hydrophobic Energy section.

We have analyzed the correlation between two sets of 210 parameters  $\omega_c^{core}$  and  $\omega_c^{surf}$ , for two values of  $R_c$ . For both cases, the correlation is found to be very weak.

Further investigation of the obtained parameters would be an interesting topic for future study. It is possible that classifying the position of an interacting pair using only the number of total contacts is not sufficient, because the number of contacts is very sensitive to  $R_c$  and to the internal structure of the protein. In some cases, a buried

pair may have a smaller value of the number of total contacts due to the peculiar structure of the protein. Moreover, different values of the threshold  $N_{th}$  may be needed for residues of different sizes. Still, ambiguity may exist for pairs in the intermediate region between core and surface. The main point we demonstrated in this section is that, by differentiating the pairwise interactions into two classes, we get a wider region of learnability than before.

One has to remember though that when hydrophobicity was added, the number of parameters in the energy increased from 210 to 230 (or, at most, 240), whereas with the present energy function we went from 210 to 420 possible parameters. Clearly, for more parameters, we expect to be able to satisfy a larger number of inequalities and, hence, expect to have a larger solvable region in the  $(R_c, M_p)$  plane. To test the extent to which the improvement in learnability was due just to the increased number of parameters, we performed a ‘‘control experiment’’: considered the two-tiered contact maps and associated energy function of Equation (6).

### Distance-Dependent Contact Energies

Several studies have pointed out the difficulties arising from the all-or-none definition of a contact implicit in Equation (1), in which even a slight change in interatomic distances between two homologous structures may cause different lists of contacts.<sup>51–53</sup> The ‘‘true’’ underlying interaction term in the free energy may well be a continuous function of the distance between the  $C_\alpha$  atoms. The simplest generalization in this direction is to identify *two* kinds of contact, depending on the distance between the  $C_\alpha$  atoms, as given in Equations (4-5).

We generated, from the databank of known structures, the two (distance dependent) contact maps (5), for native structures and for their corresponding decoys (obtained by threading). For each two-tiered map and sequence the energy, as defined in Equation (6), can be evaluated. As before, the energy difference between a native state and the  $\mu$ th decoy can be written as a linear function of the 420 contact energy parameters:

$$\Delta E_\mu = \sum_{c=1}^{210} \{ \omega_c^{(1)} [N_c^{(1)}(\mathbf{S}^{(1),\mu}) - N_c^{(1)}(\mathbf{S}^{(1),0})] + \omega_c^{(2)} [N_c^{(2)}(\mathbf{S}^{(2),\mu}) - N_c^{(2)}(\mathbf{S}^{(2),0})] \}, \quad (23)$$

where  $N_c^{(1)}(\mathbf{S}^{(1)})$  and  $N_c^{(2)}(\mathbf{S}^{(2)})$  denote the numbers of contacts of type  $c$ , each determined from its corresponding distance-dependent contact map.

In principle, we should determine the region of learnability in a three-dimensional parameter space, of  $(R_1, R_2, M_p)$ . However, in this study, we do not give a complete phase diagram in the whole space. Rather, we investigated the  $(R_1, R_2)$  plane for two values of  $M_p$ , namely, the sets  $SET_{141}$  and  $SET_{154}$ . We show in Figure 5 the points where the set is learnable and those where we have proved that it is unlearnable. Clearly, only  $R_1 \geq R_2$  is relevant.

On the dashed line  $R_1 = R_2$ , we marked by solid squares the points  $R_c = 6.5$  and  $17 \text{ \AA}$ ; these were the

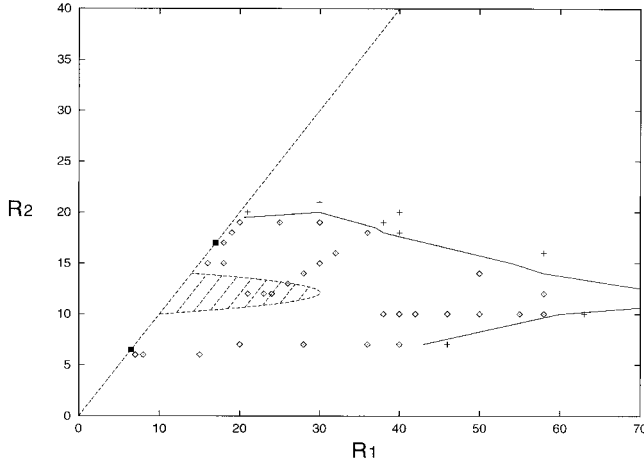


Fig. 5. The area between the two solid lines is the region of learnability of  $SET_{141}$  on the  $(R_1, R_2)$  plane. Diamonds indicate that for a particular value of  $(R_1, R_2)$  the set is learnable, whereas crosses represent unlearnable cases. The dashed line  $R_1 = R_2$  is for guidance, representing the case of a single-step potential. The two points marked by solid squares on this line are the boundaries of the range of learnability for the previously treated potential of Equation (21). The shaded area is an approximate region of learnability for  $SET_{154}$ .

boundaries of learnability of  $SET_{141}$ , by using the 420-parameter pairwise energy function (21) studied above. Another limit of interest is  $R_1 \rightarrow \infty$ , which recovers the original 210-parameter simple pairwise energy function (3). Indeed, the learnable region of  $R_2$  converges to [11 Å, 12 Å] in this limit.<sup>31</sup> Between these two limits, we found that  $SET_{141}$  becomes learnable for the relatively wide range  $6 \text{ Å} \leq R_2 \leq 19 \text{ Å}$ , provided  $R_1$  is not too large.

We have also analyzed the correlations between the values of the 420 energy parameters that were obtained for various values of  $R_1$  and  $R_2$ . Each pair of points in the learnable region of Figure 5, gives two sets of 420 parameters  $(\omega^{(1)}, \omega^{(2)})$ . For each pair, we have calculated three types of correlation: between 210  $\omega^{(1)}$  parameters, 210  $\omega^{(2)}$  parameters, and 420 total parameters. The results show an interesting behavior. For a fixed value of  $R_1$ , the correlation coefficients between  $\omega^{(1)}$  are very high ( $> 0.88$  for all cases), regardless of the value of  $R_2$ , whereas the coefficient between  $\omega^{(2)}$  decreases rapidly as the difference between  $R_2$  grows. Similar behavior is found when we fix the value of  $R_2$  varying  $R_1$ . The correlation coefficient between 420 parameters shows monotonic decrease as the distance between the two points grows. It naturally leads to the following approximations:

$$\omega_i^{(1)}(R_1, R_2) \approx f(R_1, R_2)\omega_i(R_1) + g(R_2),$$

$$\omega_i^{(2)}(R_1, R_2) \approx f'(R_1, R_2)\omega_i(R_2) + g'(R_1).$$

By using the least-square method, we have fitted the obtained parameters  $\omega^{(1)}$  and  $\omega^{(2)}$ , and found that the scalar function  $f$  (or  $f'$ ) is not a simple function of  $R_1$  (or  $R_2$ ), nor of  $|R_1 - R_2|$ . Within a statistical error,  $g$  and  $g'$  are found to be negligible,  $O(10^{-3})$ .

Evidently, the distance-dependent extension of our energy function performs as well or even better on  $SET_{141}$  as the core-surface interaction; by using both descriptions, we were able to stabilize all proteins of this set for a nearly similar range of definitions of contact, e.g.,  $R_c$  (or  $R_2$ ) ranging from 6 to 17–19 Å. We also tried to stabilize  $SET_{154}$ , which was learnable for the core-surface extension (with  $R_c = 12 \text{ Å}$ ). We used the distance-dependent contact energies and found that  $SET_{154}$  is again learnable in a rather wide range of  $R_1$  and  $R_2$ ; for  $10 \text{ Å} < R_2 < 14 \text{ Å}$ , the region of learnability extends to  $R_1 = 30 \text{ Å}$  or more. Corresponding region of learnability for  $SET_{154}$  is naively demonstrated by shaded region in Figure 5.

## DISCUSSION

The aim of this work was to understand which features of protein are more relevant and must be considered in the approximation of the energy function. We have investigated two kinds of approximations for the free energy. The first was to include an explicit hydrophobic term and the second to describe the pairwise interaction between residues in a more detailed way than the simplest contact interaction that we discussed in previous work.<sup>31</sup> We have shown how the modified approximations affect the learnability and which of the representations of the energy function is more suitable to satisfy our demand of stabilizing native contact maps.

In considering hydrophobicity, we found that the region of learnability drawn on the  $(R_c, M_p)$  plane was only slightly increased compared with the case without hydrophobicity. When the hydrophobicity is considered in a different manner, we found a small but interesting effect on learnability. In considering pairwise interactions, we have observed equivalent effects when the definition of contact is slightly modified without considering hydrophobic energy. This poses a question on the contact map representation itself. Even a subtle modification of the definition of contact (corresponding to the region where  $R_1$  and  $R_2$  differ slightly) gives rise to significant effects, making the whole set of proteins learnable. Therefore, we argue that the definition of the contact is, at least, as important as the specific assignment of the energy function. However, one should be wary that, although supplementing the hydrophobic interaction required 20 additional parameters, in modifying the pairwise contact energy function we introduced 210 new parameters. It is possible that the number of proteins that can be stabilized by a given form of the energy depends also on the number of energy parameters involved. This possibility is supported by our finding that with both the modifications of the pairwise contact energy that we tried we were able to stabilize all the proteins in database that we used. It is not easy to clarify unambiguously such a matter. The difficulty is that all the different forms of the energy that we discussed are phenomenologic approximations to the free energy of a certain sequence in the space of contact maps. As a consequence, it is possible that for example the hydrophobic interaction is already partially included in the pairwise contact energy. Therefore, that the improve-

ment induced by explicitly including a hydrophobic term was only marginal might be due both to the small number of new energy parameters introduced and to a double-counting of the same physical interactions. Although double-counting is allowed within a phenomenologic approximation, it could not be expected to produce a significant improvement of the performance of the energy function. However, it is outside the scope of the present work to attempt to quantify the extent of the double-counting in a particular form of the energy.

Although the whole set of proteins considered in this work, has been found learnable with the distance-dependent pairwise energy, from our previous results,<sup>31</sup> we are inclined to think that there should be a maximal number of proteins which can be stabilized together if one increases the number of proteins beyond 154. To stabilize larger numbers of proteins, we expect that one should go further on in modifying the the energy function, as it was already suggested by other authors.<sup>13,17,18,23,24,34,54</sup> For such purpose, this work constitutes a first step and a guideline.

Other factors to be pointed out are the definition of contact and the way of obtaining decoys. In this work, we have used mainly the  $C_\alpha$  definition. With other definitions of contact, which demand more information, it is possible to stabilize a larger number of proteins.<sup>31</sup> However, one cannot conclude from this that the  $C_\alpha$  definition is worse than others. To clarify whether the definition of contact is more relevant than the modifications of energy, is beyond the purpose of our work. Nonetheless, based on the results presented here, we expect that the modification of the pairwise potential could be more relevant than the consideration of other forces. We have also used the method of gapless threading, because it is a efficient way to obtain decoys. However, our conclusion is not limited to threading, and the main results will hold for other ways of generating decoys.

### ACKNOWLEDGMENTS

We are grateful to Luis Serrano for the suggestion to investigate the different roles played by surface residues and buried ones, to Ron Elber for discussing with us a similar approach, based on Equation 11 (unpublished), and to Ido Kanter and Gaddy Getz for discussions.

### REFERENCES

- Creighton TE. Protein folding. New York: WH Freeman; 1992.
- Fersht AR. Structure and mechanism in protein science. New York: WH Freeman; 1992.
- Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res* 1999;27:49–54.
- Anfinsen C. Principles that govern the folding of protein chains. *Science* 1973;181:223–230.
- Lattman EE. Third meeting on the critical assessment of techniques for protein structure prediction. *Proteins* 1999;37:1.
- Vendruscolo M, Domany E. Protein folding using contact maps. *Vitam Horm* (in press).
- Scheraga HA. Calculation of stable conformations of polypeptides, proteins, and protein complexes. *Chem Scr* 1989;29A:3.
- Mirny L, Domany E. Protein fold recognition and dynamics in the space of contact maps. *Proteins* 1996;26:391–410.
- Vasquez G, Nemethy M, Scheraga HA. Conformational energy calculations on polypeptides and proteins. *Chem Rev* 1994;94:2183–2239.
- Brooks CL, Karplus M, Pettitt BM. Proteins: a theoretical perspective of dynamics, structure and thermodynamics. *Adv Chem Phys* 1988;71:1–259.
- Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* 1998;282:740–744.
- Lazaridis T, Karplus M. “New view” of protein folding reconciled with the old through multiple unfolding simulations. *Science* 1997;278:1928–1931.
- Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol* 1976;104:59–107.
- Skolnick J, Kolinski A. Monte Carlo approaches to the protein folding problem. *Adv Chem Phys* 1999;105:203–242.
- De Witte RS, Shakhnovich EI. Pseudodihedrals: simplified protein backbone representation with knowledge-based energy. *Protein Sci* 1994;3:1570–1581.
- Zhou YQ, Karplus M. Interpreting the folding kinetics of helical proteins. *Nature* 1999;401:400–403.
- Hao MH, Scheraga HA. How optimization of potential function affects protein folding. *Proc Natl Acad Sci USA* 1996;93:4984–4989.
- Mirny L, Shakhnovich EI. How to derive a protein folding potential? A new approach to an old problem. *J Mol Biol* 1996;264:1164–1179.
- Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. *Curr Opin Struct Biol* 1996;6:195–209.
- Park BH, Huang ES, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J Mol Biol* 1997;266:831–846.
- Seno F, Maritan A, Banavar JR. Interaction potentials for protein folding. *Proteins* 1998;30:244–248.
- Thomas PD, Dill KA. An iterative method for extracting energy-like quantities from protein structures. *Proc Natl Acad Sci USA* 1996;93:11628–11633.
- Koretke KK, Luthey-Schulten Z, Wolynes PG. Self-consistently optimized energy functions for protein structure prediction by molecular dynamics. *Proc Natl Acad Sci USA* 1998;95:2932–2937.
- Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. When is the quasicheical approximation correct? *Protein Sci* 1997;6:676–688.
- Vendruscolo M, Domany E. Efficient dynamics in the space of contact maps. *Fold Des* 1998;3:329–336.
- Vendruscolo M, Najmanovich R, Domany E. Protein folding in contact map space. *Phys Rev Lett* 1999;82:656–659.
- Vendruscolo M, Domany E. Pairwise contact potentials are unsuitable for protein folding. *J Chem Phys* 1998;109:11101–11108.
- Bowie D, Luthy JU, Eisenberg D. A method to identify protein sequences that fold into a known 3-dimensional structure. *Science* 1991;253:164–170.
- Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. *Nature* 1992;358:86–89.
- Fisher D, Rice D, Bowie JU, Eisenberg D. Assigning amino acid sequences to 3-dimensional protein folds. *FASEB J* 1996;10:126–136.
- Vendruscolo M, Najmanovich R, Domany E. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? *Proteins* 2000;38:134–148.
- Vendruscolo M, Kussell E, Domany E. Recovery of protein structure from contact maps. *Fold Des* 1997;2:295–306.
- Maierov VN, Crippen GM. Contact potential that recognizes the correct folding of globular proteins. *J Mol Biol* 1992;227:876–888.
- Settanni G, Micheletti C, Banavar JR, Maritan A. Determination of optimal effective interactions between amino acids in globular proteins. <http://xxx.lanl.gov/abs/condmat/9902364> 1999.
- Dill KA, Bromberg S, Yue KZ, Fiebig KM, Yee DP, Thomas PD, Chan HS. Principles of protein folding: a perspective from simple exact models. *Protein Sci* 1995;4:561–602.
- Hao MH, Scheraga HA. On foldable protein-like models; a statistical-mechanical study with Monte Carlo simulations. *Physica A* 1997;244:124–146.
- Silverstein KAT, Haymet ADJ, Dill KA. Molecular model of hydrophobic solvation. *J Chem Phys* 1999;111:8000–8009.
- Lum K, Chandler D, Weeks JD. Hydrophobicity at small and large length scales. *J Phys Chem B* 1999;103:4570–4577.

39. Crippen GM. Prediction of protein folding from amino-acid-sequence over discrete conformation space. *Biochemistry* 1991;30:4232–4237.
40. Hertz J, Krogh A, Palmer RG. Introduction to the Theory of Neural Computation. Santa Fe Institute Studies in the Science of Complexity. Lecture Notes v.1 (Computation and Neural Systems Series). Addison-Wesley Publishing Company; New York, 1991.
41. Watkin TLH, Rau A, Biehl M. The statistical mechanics of learning a rule. *Rev Mod Phys* 1993;65:499–556.
42. Nabutovsky D, Domany E. Learning the unlearnable. *Neural Comput* 1991;3:604–616.
43. Jurs PC. Computer software applications in chemistry. New York: John Wiley; 1986.
44. Karmarkar N. A new polynomial time algorithm for linear programming. *Combinatorica* 1984;4:373–395.
45. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes in Fortran: the art of scientific computing. New York: Cambridge University Press; 1992.
46. Hooft RW, Sander C, Vriend G. Verification of protein structures: side-chain planarity. *J Appl Cryst* 1996;29:714–716.
47. Casari G, Sippl MJ. Structure-derived hydrophobic potential—Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J Mol Biol* 1992;224:725–732.
48. Kyte J, Doolittle RF. A simple method for displaying the hydrophobic character of a protein. *J Mol Biol* 1982;157:105–132.
49. Sweet RM, Eisenberg D. Correlation of sequence hydrophobicities measures similarity in 3-dimensional protein structure. *J Mol Biol* 1983;171:479–488.
50. Engelman DM, Steitz TA, Goldman A. Identifying nonpolar transbilayer helices in amino-acid-sequences of membrane-proteins. *Ann Rev Biophys Biomol Struct* 1986;15:321–353.
51. Finkelstein AV. Protein structure: what is it possible to predict now? *Curr Opin Struct Biol* 1997;7:60–71.
52. Koehler JPA, Rooman MJ, Wodak SJ. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J Mol Biol* 1994;235:1598–1613.
53. Bahar I, Jernigan RL. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J Mol Biol* 1997;266:195–214.
54. Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci* 1999;8:361–369.