Building protein lattice models using self-consistent mean field theory

Patrice Koehl[a)]
Department of Structural Biology, Stanford University Medical School, Stanford, California 94305

Marc Delarue[b)]
Laboratoire d'Immunologie Structurale, Institut Pasteur, 15 rue du Docteur Roux, 75015 Paris, France

(Received 25 November 1997; accepted 2 March 1998)

An optimization protocol for modeling protein structures on lattice is proposed which is based on self-consistent mean field (SCMF) theory. In this procedure, the protein residues are supposed to be independent, and their possible positions are given by a list of lattice sites. To do this, an effective larger system is considered, in which each residue i is supposed to occupy all possible sites j, each with a weight $V(i,j)$ stored in the so-called lattice probability matrix **V**. The effective energy of the system is computed, and iteratively minimized with respect to the weights **V**, the lattice sites being fixed in space. The final self-consistent **V** matrix describes the conformational space available to the protein, based on the energy function implemented. This energy function contains two types of terms, namely simple geometric terms which ensure bond connectivity and prevent chain intersection, and energy terms specific to the problem of interest. The application of the above protocol to building a lattice model of a protein, given its three dimensional structure, is discussed and compared with other lattice fitting procedures. © *1998 American Institute of Physics.* [S0021-9606(98)50122-5]

I. INTRODUCTION

It is the ability of proteins to fold into unique three-dimensional (3D) structures that allows them to display their biological function. Hence defining a relationship between the amino acid sequence and the 3D conformation of a native protein is essential to understanding biological processes. Should the laws of folding be known, protein structure prediction, and, more generally, *de novo* design of even complicated protein folds, would become tractable. The attainment of this ultimate goal is sought through two main roads, the folding problem (i.e., finding a protein structure from its sequence)[1] and the inverse protein folding problem (i.e., defining which sequences are compatible with a given protein fold).[2] Both problems can be formulated as the search for minima of thermodynamic functions. However, the search for these minima suffers from huge combinatorial problems, in structure space as well as in sequence space. Attempts to get around these problems have focused on three different directions. First, the computational procedure can include empirical information on protein structures. For example, in homology building, the conformation of the main chain is copied from an homologous template, the new sidechains are built and the full structure is then subjected to energy minimization.[3–6] Second, the dimension of the problem can be drastically reduced: in structure space, this has led to united residue representations, in which each residue of a protein is represented by one or two atoms. Furthermore, a much simpler representation of the conformational space can be used, namely lattice models, which allows complete enumeration as well as simpler dynamic simulation, since the moves are discrete. In sequence space, amino acids have been grouped into categories, such as the hydrophobic (H) and polar (H) (for review on simple models, see Ref. 7). Third, the global search for the minimum of the energy function can be simplified either by performing a nondeterministic search of an approximate solution, as in Monte Carlo methods,[8] or by replacing and modifying the potential energy function itself, as in the diffusion equation method of Scheraga and co-workers,[9] or by using mean field theory (MFT) approaches.[10] The three routes defined above are obviously not exclusive of each other. This paper will be concerned with the projection of three dimensional protein structures onto a lattice, with as little distortion as possible.

In the first application of MFT in protein simulation, presented by Elber and Karplus,[11] approximate mean field treatment of protein and ligand dynamics enabled detailed studies of the diffusion pathways of carbon monoxide in myoglobin. The basic idea there was to create a system containing the protein and multiple copies of the ligand which do not see each other, and minimize the total free energy of this effective system. The major advantage of this approach is that N alternative configurations of the ligand can be examined using a *single simulation* of the effective system. This gain in computing time became even more apparent in the major subsequent application of MFT, namely prediction of protein sidechain positions based on a known backbone.[12] In this case, an N residue protein is divided into $N+1$ subsystems, consisting of the backbone and the N sidechains. The effective system contains the backbone and multicopies of each sidechain subsystem. Assuming that each sidechain has K possible conformations, a systematic search of the

a)Author to whom correspondence should be addressed. On leave of absence from UPR9003 du CNRS, Boulevard Sebastien Brant, 67400 Illkirch Graffenstaden, France. Electronic mail: koehl@hyper.stanford.edu
b)Electronic mail: delarue@pasteur.fr

energy minimum sidechain combination would require examining K^N configurations. Within the MFT approximation, the same problem is solved by considering a single effective system, which requires only $N \times K$ pairwise energy functions to be evaluated. Various methods based on MFT have been described, each using the same underlying concept described above. Applications include sidechain modeling,[12–15] loop design in homology modeling,[16] modeling of ligand–protein interactions,[17] specific protein folding studies,[18–21] and sequence design.[22,23]

We have developed our own variant of MFT, which we refer to as self-consistent mean field (SCMF) optimization.[13] SCMF optimization is based on a discrete, fixed ensemble of conformations for each subentity of the system considered; each of these conformations is weighted by a probability, which is refined by the SCMF procedure. As such, the method requires knowledge of the total conformational space available to the system under study, and, for that reason, its applications to proteins have been limited to cases in which the global geometry of the molecule is known. It should be noted that this is the case for most of the MFT procedures proposed so far. An extension of the SCMF procedure to the direct search of the 3D conformation of a protein (based on experimental data or solely on semiempirical energy functions), is possible if the conformational space is mapped by a 3D lattice model. In a simple lattice model, each amino acid of a protein is represented as a bead, and connecting bonds are represented by lines, which follow the geometry of the chosen background lattice. These models have proven useful to test the assumptions and approximations in analytical models for protein folding as well as protein dynamics studies.[7] For this study, their major advantage is to propose a discrete representation of the conformational space available to a protein. The purpose of this paper is to present a general formulation of the SCMF procedure for protein conformational studies on a lattice, followed by an application to the problem of fitting a given protein structure to a given lattice.

Applications of MFT to protein folding problems on lattice have already been described, in particular in the lattice neural network minimisation (LNNM) procedure proposed by Rabow and Scheraga.[20] The SCMF procedure developed here is in fact a different formulation of the LNNM procedure, with variations in its computer implementation. What is new here is that we show how this procedure can be extended to handle large protein systems while keeping the computing time reasonable, as well as a detailed description of its application to fitting a protein on a lattice.

Lattice models are a major tool for protein folding studies. Consequently, the problem of defining the closest lattice representation of a given protein has attracted significant interest recently. Many techniques have been proposed to solve this problem,[24–29] which have been based on the minimization of the overall distance of the lattice model to the exact structure, as measured by the coordinate root mean squared deviation (cRMS). A second measure of the similarity of two structures exists, namely the distance root mean squared deviation (dRMS). dRMS can be understood also as the number of native contacts conserved in the projection onto a lattice. Here we show that the SCMF procedure can be adapted to minimize either the cRMS or the dRMS when fitting a protein to a lattice. A fit based on the cRMS will provide a model whose overall shape has been optimized to resemble the true protein, while a fit based on the dRMS will provide a model which tries to preserve the internal geometry of the protein. It should be noted that experimental data on protein structure are usually provided as internal distances, as for instance, in NMR.

This paper is organized as follows: In the following section, a brief description of the SCMF is given. We further detail its application to lattice studies, with emphasis on the specific case of fitting a protein to a lattice. The computational protocol itself is given in detail. In the next section, applications of the SCMF procedure to fitting a large set of proteins varying in size from 26 to 474 residues on lattices are presented. Fits based on cRMS and dRMS are presented, and the differences are discussed. Conclusions and final remarks are given in the last section.

II. METHODS

A. The self-consistent mean field (SCMF) procedure

Several descriptions of the MFT protocols have been published (see, for example, Refs. 10, 13 and 21). A basic outline of the SCMF procedure pertinent to protein studies on lattice is given here, with emphasis on the underlying assumptions.

We define the coordinate vector of all $C\alpha$ in a protein as $\mathbf{X}$. The energy E of the protein is given by a potential function U applied on these coordinates $\mathbf{X}$,

$$E = U(\mathbf{X}). \tag{1}$$

The native conformation of the protein is derived from the global minimum of E. The search for this global minimum is hindered by the presence of numerous local minima. One way to alleviate this problem is to consider an effective, larger system in which multicopies of $\mathbf{X}$ within a conformational space Ω are considered. The probability of finding the coordinates between $\mathbf{X}$ and $\mathbf{X}+d\mathbf{X}$ is denoted by $\rho(\mathbf{X})d\mathbf{X}$, where $\rho(\mathbf{X})$ is normalized to 1. The total effective energy of the system within Ω is given by

$$E_{\text{eff}} = \int_{\Omega} U(\mathbf{X})\rho(\mathbf{X})d\mathbf{X}. \tag{2}$$

The mean field approach to study this effective system is based on the following approximations of the probability distribution function ρ:

(1) It is assumed that ρ can be described by a Hartree *product* of *independent* probability densities of different subsystems:[30]

$$\rho(\mathbf{X}) = \prod_{j=1}^{J} \rho_j(\mathbf{X}_j). \tag{3}$$

Here the subsystems correspond to the $C\alpha$s, and J is the number of residues in the protein. This basically defines each $C\alpha$ as independent of the others. Bond connectivity will then have to be imposed as a constraint, and included in the energy function U.

(2) Each subsystem j can adopt a *finite* number of copies, and the ρ_j are expanded in delta functions,

$$\rho_j(\mathbf{X}_j)=\sum_{k_j=1}^{K_j} V(j,k_j)\,\delta(\mathbf{X}_j-\mathbf{X}_{k_j}^0), \tag{4}$$

where k_j runs over all copies of the subsystem j and $\mathbf{X}_{k_j}^0$ contains the coordinates of the k_j possible conformation for subsystem j. $V(j,k_j)$ are normalization factors satisfying,

$$\sum_{k_j=1}^{K_j} V(j,k_j)=1. \tag{5}$$

Here K_j are the K lattice sites accessible by the jth $C\alpha$ of the protein of interest, and the weights V are discrete spatial probability distribution functions for the positions of the $C\alpha$ on the lattice.

In the following, we also assume that the potential function U only contains one ($U^{(1)}$) and two body ($U^{(2)}$) interaction terms,

$$U(\mathbf{X})=\sum_{j=1}^{J} U_j^{(1)}(\mathbf{X}_j)+\frac{1}{2}\sum_{j=1}^{J}\sum_{i\neq j} U_{ij}^{(2)}(\mathbf{X}_i,\mathbf{X}_j) \tag{6}$$

(interactions between residue i and j are counted twice, hence the 1/2 factor). Substituting Eqs. (3), (4), and (6) into Eq. (1) and integrating over the spatial variables leads to

$$E_{\text{eff}}=\sum_{j=1}^{J}\sum_{k_j=1}^{K_j} V(j,k_j)U_j^{(1)}(\mathbf{X}_{k_j}^0) + \frac{1}{2}\sum_{j=1}^{J}\sum_{k_j=1}^{K_j}\sum_{i\neq j}^{J}\sum_{l_i=1}^{K_i} V(j,k_j)V(i,l_i)U_{ij}^{(2)}(\mathbf{X}_{k_j}^0,\mathbf{X}_{l_i}^0), \tag{7}$$

where $\mathbf{X}_{k_j}^0$ is the coordinate vector for the k_j possible conformation for subsystem j.

In our formulation of the MFT (i.e., the SCMF procedure[13]), the positions of the various copies of the subsystems are supposed to be known and fixed in space (they correspond to lattice sites in the application described here); the effective system is then described by an array $\mathbf{V}$, whose current element $V(j,k_j)$ is the probability that subsystem j is described by its possible state k_j. The problem of finding the global minimum energy for the true system (i.e., the protein) in the total conformational space is then mapped into the problem of finding the minimum of the free energy of the ''effective'' system, defined as

$$F=E_{\text{eff}}-TS, \tag{8}$$

where E_{eff} is the effective energy given in Eq. (7), which is in fact the sum of real potential energies calculated at different points [obtained from the delta function expansions in Eq. (4)] and multiplied by normalization factors, and T and S are the temperature and the entropy of the system, respectively.

Since all $C\alpha$ and all lattice sites are assumed to be independent, the entropy S is given by

$$S=-k\sum_{j=1}^{J}\sum_{k_j=1}^{K_j} V(j,k_j)\log[V(j,k_j)]. \tag{9}$$

The minimum of F is obtained by setting all its derivatives with respect to $V(j,k_j)$ to zero,

$$\frac{\partial F}{\partial V(j,k_j)}=\frac{\partial E_{\text{eff}}}{\partial V(j,k_j)}-T\,\frac{\partial S}{\partial V(j,k_j)}=0. \tag{10}$$

Substituting Eq. (9) in Eq. (10), and using the fact that the probabilities V are normalized [Eq. (5)], the solutions of Eqs. (10) are given by

$$V(j,k_j)=\frac{\exp\left(-\dfrac{W(j,k_j)}{kT}\right)}{\displaystyle\sum_{l_j=1}^{K_j}\exp\left(-\frac{W(j,l_j)}{kT}\right)}, \tag{11}$$

where $W(j,k_j)$ is the mean field applied to the conformation k_j of subsystem j,

$$W(j,k_j)=\frac{\partial E_{\text{eff}}}{V(j,k_j)} =U_j^{(1)}(X_{k_j}^0)+\sum_{i\neq j}\sum_{l_i=1}^{K_i} V(i,l_i)U_{ji}^{(2)}(X_{k_j}^0,X_{l_i}^0). \tag{12}$$

For given values of V, the mean field seen by residue j for all its possible lattice position kj can be evaluated using Eq. (12), and the spatial probabilities V can then be updated using Eq. (11). The optimization then proceeds as follows: the probability matrix V is initialized, for example with equiprobable values [i.e., $V(j,k_j)=1/K_j$, where K_j is the number of copies for subsystem j], and the system of Eqs. (11) and (12) is iterated until convergence, i.e., until self consistency is achieved.

Interestingly, Eqs. (11) and (12) have been independently derived by direct evaluation of the partition function Z using the saddle-point approximation for neural network minimization.[31,32] In the latter case, E_{eff} and W are generalized energies and kT is a parameter. This alternative derivation is referred to as the lattice neural network minimisation (LNNM) by Rabow and Scheraga.[20]

B. The potential energy function for fitting a protein on a lattice

Let $\mathbf{L}_1$, for $l=1,...,N$ give the coordinate vectors for the L lattice sites sufficient to contain the full protein for which a lattice model is sought.

We denote M_j the subset of lattice points accessible to the j-th $C\alpha$ of the protein (the construction of the complete lattice as well as the derivation of the different M_j will be detailed below). According to the notation adopted above, M_j contains K_j elements. The current element k_j of M_j corresponds to lattice site $\text{lat}_j(k_j)$, such that

$$\mathbf{X}_{k_j}^0=\mathbf{L}_{\text{lat}_j(k_j)}. \tag{13}$$

Any application of MFT to modeling proteins on a lattice such that each lattice site can represent a $C\alpha$ requires at least that two energy terms be included in the potential functions $U^{(2)}$,

(1) Two sets M_j and M_i have no reason to be exclusive, in which case, if no specific precautions are taken, the MFT procedure might position residue j and i on the same lattice site. To prevent such an intersection of the protein chain, a potential function I is introduced, such that

$$I_{ji}^{(2)}(\mathbf{X}_{k_j}^0,\mathbf{X}_{k_i}^0)=\delta(\mathrm{lat}_j(k_j)-\mathrm{lat}_i(k_i)), \tag{14}$$

where δ is the Dirac function [$\delta(x)=0$ if $x\neq 0$ and $\delta(0)=1$]: I is positive, nonzero, only if the lattice sites corresponding to conformations k_j for j and k_i for i are identical.

(2) Since the MFT procedure assumes the residues to be independent of each other, a bond connectivity criterion must be included, in the form of a potential B,

$$B_{ji}^2(\mathbf{X}_{k_j}^0,\mathbf{X}_{k_i}^0)=(\delta(|j-i-1|)+\delta(|j-i+1|))\times\delta_D(\mathrm{lat}_j(k_j),\mathrm{lat}_i(k_i)), \tag{15}$$

where

$$\delta_D(a,b)=\begin{cases}0 & \text{if } 2.6<d_{ab}<4.7\\ 1 & \text{otherwise}\end{cases} \tag{16}$$

(d_{ab} is the distance between lattice sites a and b).

If the residues i and j are adjacent in sequence, and their lattice sites are not properly bonded (defined here by the condition that their distance separation is between 2.6 and 4.6 Å), B is equal to 1, which corresponds to a penalty term. In all other cases, $B=0$.

These two energy terms, which ensure correct connectivity of the chain on the lattice, are then supplemented by a potential energy term specific to the problem of interest. In this paper, we are concerned with fitting a known structure on a lattice, which requires an energy term that will drive the lattice chain to resemble the true structure. There are two possible measures of the similarity of two protein models, which can be used for that purpose: the coordinate root mean square deviation (cRMS), defined by,

$$\mathrm{cRMS}=\left(\frac{\Sigma_{j=1}^{J}|\mathbf{X}_{aj}-\mathbf{X}_{bj}|^2}{J}\right)^{1/2}, \tag{17}$$

where $\mathbf{X}_{aj}$ and $\mathbf{X}_{bj}$ are the coordinates of atom j of model a and model b, respectively, J the total number of atoms considered, and where models a and b have been optimally superimposed, and the distance root mean square deviation (dRMS), defined by

$$\mathrm{dRMS}=\left(\frac{2\Sigma_{j=1}^{J-1}\Sigma_{i=j+1}^{J}(|\mathbf{X}_{ai}-\mathbf{X}_{aj}|-|\mathbf{X}_{bi}-\mathbf{X}_{bj}|)^2}{J(J-1)}\right)^{1/2}, \tag{18}$$

in which case no optimal superposition is needed.

Both measures can be used to define the best lattice model for a given protein structure, and both fit within the framework developed above:

(1) Fit based on cRMS: for a *given* orientation of the lattice with respect to the known protein structure, a potential function C, which contains only a one body term [$C^{(2)}=0$], is introduced, such that,

$$C_j^{(1)}(\mathbf{X}_{k_j}^0)=(\mathbf{X}_{k_j}^0-\mathbf{X}_j^P)^2, \tag{19}$$

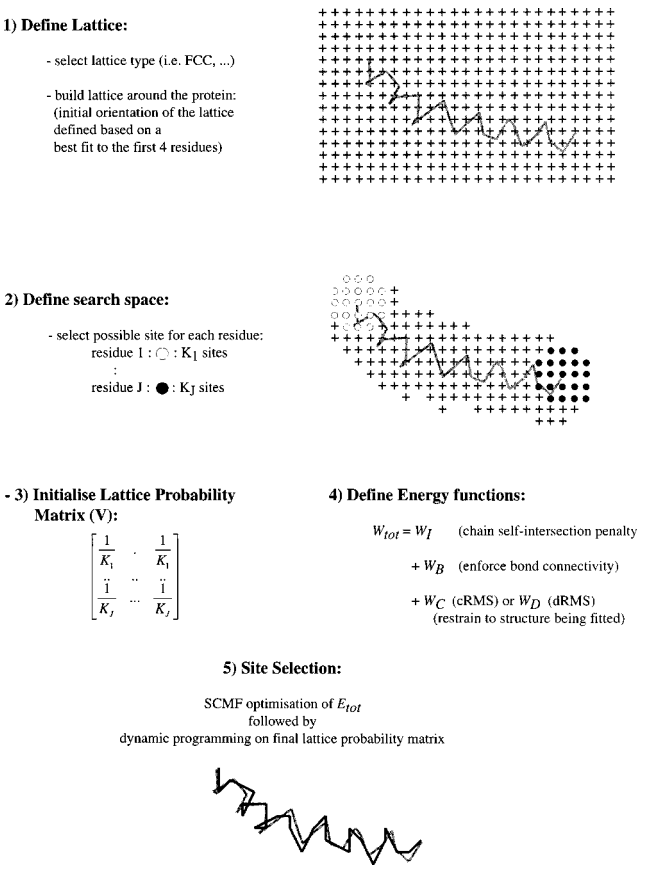

FIG. 1. Outline of the SCMF procedure for fitting a protein on a lattice.

where $\mathbf{X}_j^P$ is the position of the $C\alpha$ of residue j in the protein structure.

(2) Fit based on dRMS: a potential function D is introduced, such that,

$$D_{ji}^{(2)}(\mathbf{X}_{k_j}^0,\mathbf{X}_{k_i}^0)=(|\mathbf{X}_{k_j}^0-\mathbf{X}_{k_i}^0|-|\mathbf{X}_j^P-\mathbf{X}_i^P|)^2 \tag{20}$$

and $D^{(1)}=0$. It should be noted that D is independent of the orientation of the lattice.

The total potential function U is then defined as,

$$U=\omega_I I+\omega_B B+\omega_C C+\omega_D D, \tag{21}$$

where ω are weights: $\omega_D=0$ for a fit based on cRMS, and $\omega_C=0$ for a fit based on dRMS, respectively, referred to as SCMF-cRMS and SCMF-dRMS below.

Equation (21) includes both one body and two body terms.

C. The computational procedure

A general overview of the application of SCMF minimization to the problem of fitting a protein onto a lattice is given in Fig. 1. Most of the elements presented below are more general, however, and would apply to any application of MFT with lattice fitting procedures. Each step of Fig. 1 is commented on below.

1. Positioning the lattice

The SCMF procedure can be applied to any lattice type (an extensive listing of possible lattices is given by Park and Levitt[26]). For the chosen lattice, all possible conformations

of four sequentially bonded sites are generated, and optimally superimposed to the first four $C\alpha$ of the protein (four atoms being the minimum required to provide an unambiguous orientation in 3D space). The conformation with the lowest cRMS is chosen to define a reasonable orientation of the lattice with respect to the protein, and the remaining sites are generated based on the lattice definition, such as to surround the protein completely. (If the protein structure were not known, the lattice dimension would be deduced from the average behavior of proteins with similar sequence length.) For fitting based on dRMS, this initial orientation will not be important, while for fitting based on cRMS, other orientation will have to be tested.

2. Site selection for each residue

If the protein structure is not known, all lattice sites are allowed for all residues. The dimensionality of the problem can be drastically reduced, however, if the protein structure is known (even roughly). In this case a lattice is built around the protein, based on a reasonable orientation obtained using the procedure described in Sec. II C 1. Then for each residue j, only lattice sites surrounding j at a distance of 10 Å or less are included in M_j.

3. Initial lattice probability matrix

The lattice probability matrix ($\mathbf{V}$) describes the spatial distribution functions of the protein residues on the lattice: $V(i,l)$ is the probability that residue i occupies the lattice site l. Before refinement, no information on the position of the residues is provided, and all sites are supposed to be equiprobable,

$$V(j,k_j)=\frac{1}{K_j} \quad \text{for } k_j=1,...,K_j \quad \text{and } j=1,...,J. \tag{22}$$

Ideally, after convergence, each residue should occupy a single site on the lattice, hence the final matrix $\mathbf{V}$ should consist of 0s, with one 1 per row.

4. Calculation of the local mean fields W(j,kj)

The effective field $W(j,k_j)$ experienced by the k_j-th possible lattice site [$\mathrm{lat}_j(k_j)$ in the global ordering of the lattice sites] for residue j of the protein can be partitioned into

$$W(j,k_j)=\omega_I W_I(j,k_j)+\omega_B W_B(j,k_j)+\omega_C W_C(j,k_j)+\omega_D W_D(j,k_j), \tag{23}$$

where W_I is derived from the penalty function for chain intersection (I), W_B from the energy term for bond connectivity (B), and W_C and W_D from the energy terms for fitting the protein structure on the lattice based on cRMS and dRMS, respectively. All four terms are based on Eq. (12). The major advantage of time-independent MFT minimization is that its computer applications are usually fast, since all energy terms can be computed first and stored for subsequent use in the minimization cycles. However, this advantage might not be sufficient in the lattice applications described here. Assuming that all L lattice sites are accessible to all J residues in the protein, the computation of W_I, W_B, and W_D in Eq. (23) would generally require LJ multiplications (W_C only contains a one body interaction term, hence the operation count is 1). The overall order for each cycle of the SCMF procedure would then be $3L^2J^2$. The volume of a protein has been found to vary linearly with its number of residues,[33] hence L is proportional to J, which yields a computing time proportional to J^4, where J is the size of the protein. This order of operation can be significantly reduced, first by reducing the number of sites available to each residue (see above), second by rearranging the expressions for the W terms, and removing all known zero contributions.

a. Efficient computation of W_I. The penalty for chain intersection can be computed as follows. For a given lattice site s ($s=1,\cdots L$), we denote as $R(s)$ the list of protein residues for which s is allowed, and $idx(j,s)$ the position of s in the list of sites available to residue j of $R(s)$. We first compute

$$S=\sum_{j\in R(s)} V(j,idx(j,s)), \tag{24}$$

then

$$W_I(j,idx(j,s))=S-V(j,idx(j,s)) \quad \text{for all } j \text{ in } R(s). \tag{25}$$

With this device, the total number of operations for all W_I terms within a cycle has been reduced to $2LJ$ additions.

b. Efficient computation of W_B. The calculation of the energy term W_B which ensures bond connectivity can be similarly reduced, as initially shown by Rabow and Scheraga:[20] let s be the lattice site corresponding to the kj-th site available to residue j [$s=\mathrm{lat}_j(k_j)$ in the notation described above]. We denote as $N(s)$ the list of lattice sites which can be bonded to s. All lattice sites k available to residue $j-1$ which do not belong to $N(s)$ will contribute to the penalty term for incorrect bond length, yielding

$$P_{j-1}=\sum_{k\in M_{j-1},k\notin N(s)} V(j-1,k) =1-\sum_{k\in M_{j-1}\cap N(s)} V(j-1,k). \tag{26}$$

Including the term for residue $j+1$, we obtain

$$W_B(j,k_j)=2-\sum_{k\in M_{j-1}\cap N(s)} V(j-1,k) -\sum_{k\in M_{j+1}\cap N(s)} V(j+1,k), \tag{27}$$

so that the total number of operations for all W_B terms has been reduced to $2LJN$ additions, where N is the number of neighbors on the given lattice.

c. Computation of W_D. There is no simple reorganization for W_D. In this case, the major time-saving procedure is the reduction of the number of lattice sites available to a residue, as described in Sec. II C 2 above: if N_b is the average number of lattice sites selected for any given residue, the computation of W_D is then reduced from J^4 in general (see above) to $J^2N_b^2$ where J is the size of the protein. N_b is independent of protein length, and only depends on the lat-

tice type and the cutoff value chosen to define the possible lattice site for a given protein residue: for example, 138 is a typical value for N_b for an extended faced-centered cubic (eFCC) lattice, with a cutoff of 7.5 Å.

5. The SCMF optimization

Given the initial lattice probability matrix (**V**), all local mean fields can be calculated as described above. These effective local potentials are then converted back into probabilities, using Eq. (11), yielding a calculated lattice probability matrix $\mathbf{V}_{\text{calc.}}\mathbf{V}$ itself is updated according to

$$\mathbf{V}_{\text{new}}=\lambda\mathbf{V}_{\text{calc}}+(1-\lambda)\mathbf{V}_{\text{old}}, \tag{28}$$

in order to avoid oscillation in the system.[13,18] The procedure is then iterated till convergence, i.e., until **V** does not change any more (i.e., $\|\mathbf{V}_{\text{new}}-\mathbf{V}_{\text{old}}\|<0.0001$).

6. Building the lattice model from the optimized lattice probability matrix

The final lattice probability matrix can be seen as a profile scoring matrix for the ''alignment'' of the protein on the lattice. The optimal alignment is then obtained using a dynamic programming approach,[34] with the constraints that no gaps are allowed both in the protein and in the lattice (i.e., for two consecutive residues, only neighboring lattice sites are considered). The procedure generates a score matrix **S**, residue by residue, as follows: for the first residue, the score $S(1,i)$ of each possible lattice site i in M_1 is directly derived from the optimized probability matrix $\mathbf{V}_{\text{min}}$. Then the score $S(2,j)$ of each lattice site j in M_2 is built according to

$$S(2,j)=V_{\text{min}}(2,j)+\max\{S(1,k)\}_{k\in M_1\cap N(j)}, \tag{29}$$

where $N(j)$ is the list of lattice sites that can be bonded to j (see Sec. II C 4 above). For each j in M_2, the corresponding ''best'' lattice site k_b for residue 1 is stored as a pointer. The procedure is iterated over all residues until the score matrix is full. The lattice site for residue J with the highest score is selected, and the complete lattice chain is generated using backtracking.

This procedure can be further improved by introducing a penalty term for chain intersection: based on the score matrix derived up to residue i, the chain leading to site j for residue i is built, such that $k_b(l)$ is the lattice site occupied by residue $l-1$ for all l in $[2,i]$ (the k_b are derived from the pointers defined above). A penalty term $P(i,j)$ is added to $S(i,j)$, such that

$$P(i,j)=-\sum_{l=i}^{2} A\,\delta(k_b(l)-j), \tag{30}$$

where A is a parameter (set to 100) and δ the delta function: if none of the $k_b(l)$ is equal to j, $P(i,j)=0$.

Equations (29) and (30) ensure that the optimal chain is correctly bonded and nonintersecting. The introduction of Eq. (30) may lead, however, to a situation without any solution (i.e., in the case where all values in the final row of S are negative). This was not observed in the case of the eFCC lattice, when the cutoff to defined possible lattice site is set to 7.5 Å.

It should be noted that $\mathbf{V}_{\text{min}}$ is a probability matrix and should therefore first be transformed into energy terms (for sake of additivity) before dynamic programming can be performed. We have tested both approaches, and observed no difference in the resulting lattice models.

The final lattice model is compared to the true structure using both cRMS and dRMS measures.

7. Parameters

In all subsequent calculations, kT of Eq. (11) (for conversion of local mean fields into probabilities) has been set to 0.6, and maintained constant (which differs from the cooling procedure introduced both by Rabow and Scheraga[20] and Lee[14] in their MFT applications). The coefficients for intersection penalty (ω_I) and bond connectivity (ω_B) were both set to 50, while the coefficients ω_C and ω_D were set to 1 or 0, depending on the measure used for fitting the protein on the lattice. The parameter λ was set to 0.3. All optimizations required fewer than 100 cycles.

8. Protein structures database and lattice size and definition

To demonstrate the ability of the SCMF procedure to fit a protein on a lattice, it was applied to a database of 105 proteins varying in size from 26 residues (melittin; PDB code 2mlt) to 476 residues (glycosidase from fungus; PDB code 2aaa). The corresponding entries in the PDB[35] are: 156b, 2aaa, 8abp, 8adh, 3adk, a8atc, b8atc, a2aza, 3blm, 1bp2, 2ca2, 1cc5, 1ccr, a2ccy, 3cd4, 2cdv, 3cla, 3cna, a4cpa, 5cpa, 2cpp, 1cpv, 1crn, 2cro, e1cse, i1cse, 1ctf, 2cy3, 2cyp, 8dfr, a4dfr, a1dhf, 1eca, 4enl, e2er7, 1fba, 12fb4, a1fcb, 1fd2, 1fx1, 3fxc, 4fxn, a3gap, 2gbp, 1gcr, o1gd1, a1hhb, 1hip, a2hla, b2hla, 1hoe, 1ilb, 3icb, 7icd, 1101, 2lbp, 6ldh, 1lhl, 31lrd, a2ltn,1lz1, 5mba, 1mbd, a4mdh, 2mhr, 2mlt, 2mnr, 2ovo, a2pab, 9pap, 2paz, 1pcy, a1pfk, 3pgk, 3pgm, a1pii, b1pii, 1phh, 5pti, 4ptp, 1rhd, 2rhe, 2rnt, 7rsa, 5rxn, 2sga, 3sgb, 1sn3, 2sns, o2sod, 2ssi, 2stv, 1tim, i1tgs, 6tmn, 4tnc, a1tnf, 1ubq, 1utg, a9wga, r2wrp, b1wsy, a4xia, a1ypi.

We have chosen to use an extended face-centered cubic (eFCC) lattice [characteristic length: 1.9 Å; moves are permutations of the vectors $(\pm2,0,0)$, $(\pm2,\pm1,\pm1)$, and $(\pm1,\pm1,0)$, corresponding to a coordination number of 42] for all applications, since this lattice has been shown to provide adequate representation of proteins.[36,37]

III. RESULTS AND DISCUSSION

A. Protein lattice models based on cRMS minimization

Lattice models were constructed using the SCMF-cRMS procedure for all 105 proteins, using the first four residues of the protein to position the lattice (see Sec. II). Figure 2 shows the final corresponding cRMS as a function of the length of the protein. As already observed by Park and Levitt,[26] cRMS deviations rise with protein size up to a length of 200 residues approximately, after which they remain fairly constant, due to the partition of larger proteins into domains.

An important feature of deriving lattice models based on the cRMS measure is the orientation of the lattice with re-

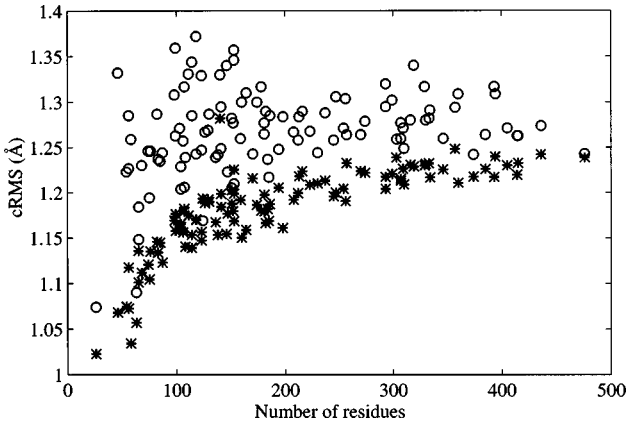

FIG. 2. The cRMS deviation between the lattice model and the true structure of each protein in our database is plotted as a function of the protein length. Lattice models were derived from SCMF minimizations based on cRMS, using a single orientation of the lattice (○), or a grid search of the best orientation of the lattice with respect to the true structure, with a angular step of 10° for each Euler angle (*). Note the dispersion in cRMS for the single orientation, which shows that the chosen orientation performs differently on different proteins.

spect to the protein. To assess this parameter, the calculations were repeated for each protein by testing different lattice orientations, obtained by systematic variations of 10° of each of the 3 Euler angles which defines the position of the lattice, for a total of 2916 different orientations. The models with the best cRMS are saved. Results are shown in Fig. 2. Orienting the lattice based on the first four residues is far from optimal, and better models could be generated for other orientations for all proteins in our database: the mean of the cRMS values between the lattice model and the true structure averaged over the 105 proteins is decreased from 1.28 to 1.18 Å (Table I). It is still unclear at this stage whether or not the models generated through the rotational grid search proposed here correspond to the global optimum fit (see discussion below).

TABLE I. Comparison of different techniques for generating lattice models of protein structures.[a]

Method	Energy function	cRMS (Å)	dRMS (Å)
SCMF-cRMS[b]	cRMS	1.27	1.04
SCMF-cRMS[c]	cRMS	1.18	0.97
SCMF-dRMS	dRMS	1.26	0.93
PL-cRMS	cRMS	1.14	0.94
PL-dRMS	dRMS	8.0[d]	0.98

[a]All methods were tested on extended face-centered cubic (eFCC) lattices. SCMF stands for the procedure based on mean field theory proposed in this paper, while PL stands for the Park and Levitt procedure (Ref. 26). Results are averaged over the 105 proteins in our database.
[b]Lattice orientation is derived from the first four residues of the protein.
[c]A coarse grid search (of 10° for each Euler angle) is performed to define the orientation of the lattice with respect to the protein.
[d]PL-dRMS builds a lattice model with minimal dRMS distance to the protein structure. Since no chirality constraint was introduced, the method cannot distinguish a structure from its mirror image, which explains its poor performance in terms of cRMS (see text for details).

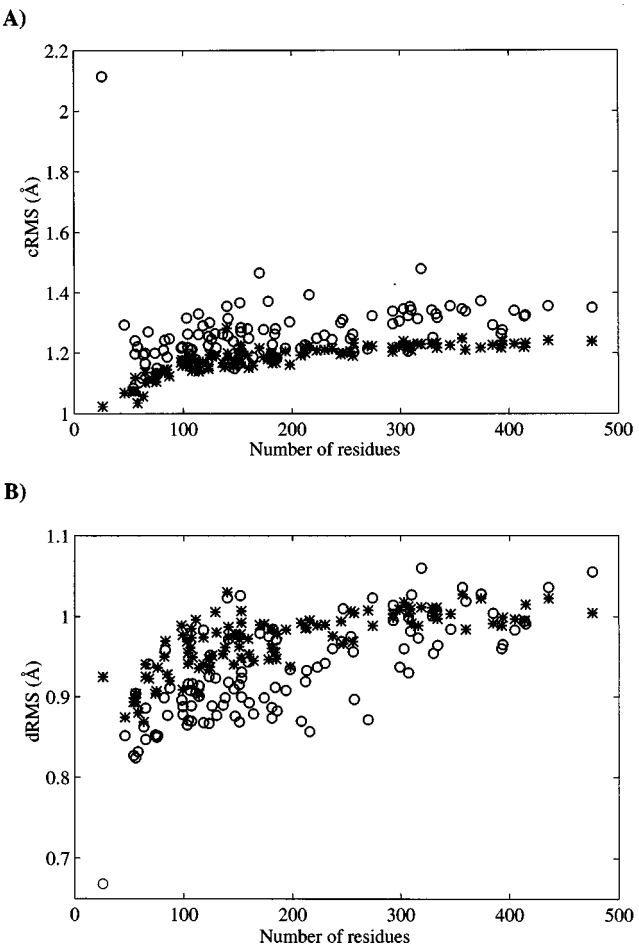

FIG. 3. The cRMS (A) and dRMS (B) deviations of the lattice model from the true structure of each protein in our database are plotted versus the number of residues in the protein. Two methods for generating the lattice models are compared, (1) SCMF-cRMS with grid search of the best orientation of the lattice with respect to the protein (*), and (2) SCMF-dRMS (○).

B. Comparing cRMS and dRMS as measures for generating protein lattice models

Parallel lattice model building for the 105 proteins was performed using SCMF-cRMS and SCMF-dRMS, and the results are compared in Fig. 3 and Table I (in the case of SCMF-cRMS, a systematic search of the lattice orientation is performed, as described above). As intuitively expected, minimization based on cRMS yields lattice models which fit to the protein with a better cRMS, while minimization using dRMS yields models with better dRMS to the protein structure, on average. dRMS is usually assumed to be a better measure of the overall similarity between two models of a protein, though it has one drawback in that the dRMS between a model and its mirror image is zero, while the corresponding cRMS can be quite large. This was observed here in the case of melittin (PDB code 2mlt), a 26-residue fully helical peptide. The lattice model for melittin generated using dRMS as a measure differed from the PDB structure 2mlt by 0.65 Å only using dRMS as a measure, but by 2.2 Å for the cRMS, which clearly indicates that this model is wrong (in fact, it corresponds to a left-handed helix). The model obtained using the cRMS measure did not show this

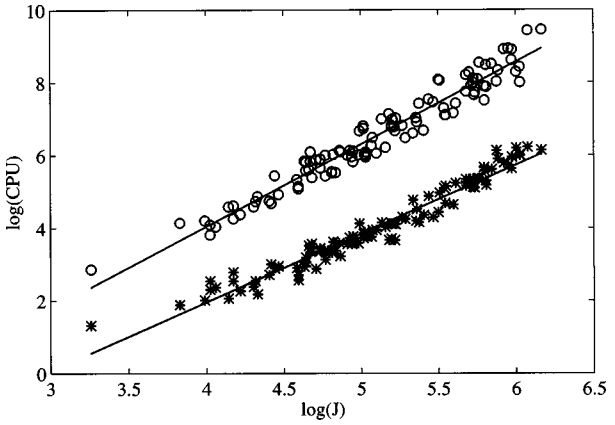

FIG. 4. The logarithm of the computing times (in seconds) required by the SCMF procedure to fit a protein on a lattice using an energy function based on cRMS (*) and dRMS (○) are plotted versus the log of the size (J, i.e., number of residues) of the protein. Computations were performed on each of the 105 proteins of our database. Linear least squares fit to the data are shown as continuous lines. A straight line fits the data for the SCMF-cRMS procedure with a correlation coefficient $R^2=0.97$, yielding: Log(CPU) $=1.88\,\mathrm{Log}(J)-5.58$, which is in agreement with a computing time for SCMF-cRMS proportional to J^2. Similarly, a straight line fits the data for the SCMF-dRMS procedure with $R2=0.97$, yielding Log(CPU) $=2.25\,\mathrm{Log}(J)-4.98$, which is in agreement with a computing time for SCMF-dRMS close to being proportional to J^2.

discrepancy (see Fig. 3). Interestingly, only melittin showed this behavior, and no models corresponding to a mirror image of the correct protein structure were obtained for larger proteins.

It should be mentioned that the lattice model generation procedure based on dRMS presented here is not completely independent of the orientation of the lattice because of the selection of the possible sites for each residue. The procedure was therefore tested with various initial orientations of the lattice, as well as with various cutoffs for the selection of the lattice sites (minimal value 7.5 Å). No differences were observed in the case of the eFCC lattice.

The total CPU time required for SCMF generation of lattice models based on both cRMS and dRMS measures varies as J^2 where J is the length of the protein (Fig. 4). This is in complete agreement with the estimation of times provided in the Sec. II: computation based on cRMS is dominated by the evaluation of the bond penalty term, which was shown to be proportional to JLN, where L is the total number of lattice sites and N the number of direct neighbors of a lattice site (for an eFCC lattice $N=42$). N being fixed, the computation order is reduced to J^2 since L varies linearly with respect to J. In the case of dRMS, the computation is dominated by the evaluation of W_D, which should be a J^4 process (see above). However, since the number of possible lattice sites for a given residue has been limited, the computation is more closely approximated by a J^2 process.

C. Comparison with other methods

Lattice models of protein have received increasing attention in the recent years, mainly because full atomic models do not presently allow the study of long range dynamic processes (such as the simulation of protein folding process) of even small proteins. However, to be useful and accurate, these models should fit as closely as possible to the off-lattice native protein structures. Several methods for fitting a protein chain on a lattice have been published recently. The simplest possible method to perform this task is to build a lattice around the protein of interest, and then to select the closest site for each residue which maintains connectivity with the lattice sites chosen for the neighboring residues. This is one of the three methods proposed by Godzik *et al.*[24] This algorithm, however, has no control on the quality of the overall fit, and may even fail to define a model if self-intersecting chains are excluded. Several more sophisticated methods have been proposed, including simulated annealing of an energy function based on cRMS,[24] dynamic programming,[25] a variant of MFT theory followed by dynamic programming,[28] as well as a chain growth method which conserves the lowest energy intermediate chains at each step.[26] All these methods attempt to generate lattice models with minimal cRMS deviations to the true protein structures, and are confronted with two major issues: (1) avoiding intersections or overlaps in the lattice chain, and (2) defining the orientation of the lattice with respect to the protein. The chain growth method developed by Park and Levitt[26] is designed to be self-avoiding, while all other methods introduce a penalty term for intersecting chains in the globally minimized energy function. The second problem related to the lattice orientation (as well as position) is either ignored, refined at each step of the chain growth, or solved through systematic or heuristic searches by building a lattice chain for a series of orientations and keeping the chain with the lowest cRMS to the true structure. The method proposed here completely avoids this problem by using the second measure of structural similarity, i.e., dRMS.

Though our initial goal was to derive a general, fast method for protein modeling on a lattice, and not to outperform other methods, it is important to test its usefulness by comparison with one of the more efficient other methods. We have chosen the method of Park and Levitt,[26] which is fast, reliable, and has an elegant solution to the problem of orienting the lattice with respect to the protein structure. This method can also be modified easily to use dRMS rather than cRMS as a minimization criterion. Briefly, all possible lattice chains of four residues are first generated, from which the Nkeep ones that are ''closest'' to the protein are selected (Nkeep is a parameter of the procedure). Then on each of these chains, all possible positions Np of the next residue are considered (eliminating those that result in a chain overlap), yielding a total of NkeepxNp chains. These chains are ranked according to their distances to the true structure, and the Nkeep best ones are kept to serve as starting points for generating the next residue. The procedure is iterated until the full chain is built. In their original paper, Park and Levitt used the cRMS (with optimal superposition of the models) to measure the distance or ''closeness'' of the lattice models to the protein. They have shown that the total running time of this method scales as J^2, where J is the number of residues in the protein, which makes it equivalent to one optimization based on SCMF with cRMS as a measure, using a single orientation of the lattice, and is faster than the SCMF proce-

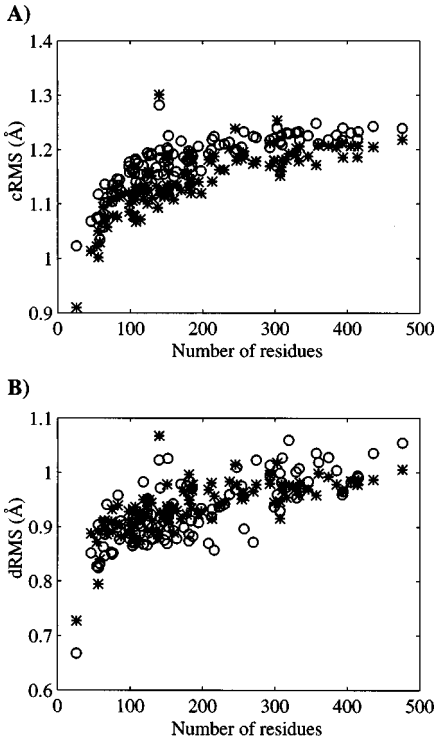

FIG. 5. The Park and Levitt (PL-cRMS) procedure (*) for fitting protein structures on lattice is compared to the SCMF-cRMS procedure (with grid search of the best orientation of the lattice with respect to the protein) (○) in (a), and to the SCMF-dRMS procedure (○) in (b). While PL usually provides better lattice models than SCMF-cRMS, it performs equally well as the SCMF-dRMS method.

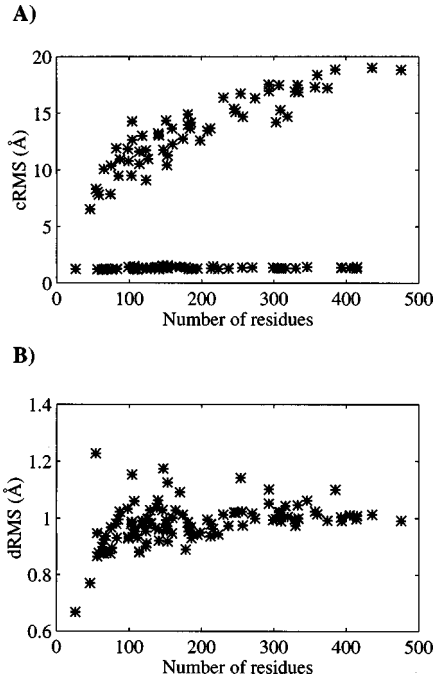

FIG. 6. The cRMS (A) and dRMS (B) deviations of the lattice model generated with the modified Park and Levitt method (PL-dRMS) from the true structure of each protein in our database are plotted versus the number of residues in the protein. Note that while all dRMS values are reasonable, 58 out of the 105 models have large cRMS values: in these cases, the mirror image of the protein structure was generated. Since information on the global chirality of the molecule was not introduced in the process of designing the lattice chain, it is expected that on average half of the lattice models built should correspond to a mirror image.

dure using dRMS for comparing models. It is straightforward to modify the Park and Levitt procedure such that it uses dRMS as a tool to compare the lattice model built with respect to the true structure. To minimize the computing time required by this modification, we use the recurrent equation

$$\frac{N(N-1)}{2}\,\mathrm{dRMS}_N^2=\frac{(N-1)(N-2)}{2}\,\mathrm{dRMS}_{N-1}^2+\sum_{i=1}^{N-1}\left(|\mathbf{r}_{ai}-\mathbf{r}_{aN}|-|\mathbf{r}_{bi}-\mathbf{r}_{bN}|\right)^2, \tag{31}$$

which simply states that the calculation of the dRMS related to a chain of length N (dRMS_N) can be estimated from the value of the dRMS for the same chain in which only the first $N-1$ residues are considered (dRMS_{N-1}). Both the original (PL-cRMS) and the modified (PL-dRMS) Park and Levitt methods were used to generate lattice models for the 105 proteins of our database, and Fig. 5 and Table I show the comparison of these results with those obtained by the SCMF procedures.

The original PL-cRMS method of Park and Levitt generates better models in terms of cRMS than the SCMF-cRMS procedure [Fig. 5(a)]. The models generated by the SCMF procedure could be improved by performing a finer grid search or by using an optimization protocol to determine the best orientation of the lattice with respect to the protein, but at a high computational cost. Interestingly, on average the models generated by the PL-cRMS procedure have similar dRMS to the true structures than those generated by the SCMF-dRMS procedure [Fig. 5(b) and Table I].

The minimizations based on the modified procedure of Park and Levitt (PL-dRMS) are good examples of the difficulties of using dRMS as a distance measure between two models: for 58 proteins out of the 105 considered, the minimized lattice model corresponds to a mirror image of the true structure, as detected by a low dRMS value [Fig. 6(a)], but a high cRMS value [Fig. 6(b)]. We did not encounter the same problem with the SCMF-dRMS method (except for one protein, melittin), not because of the SCMF method itself, but because of the selection of possible sites for each residue of the protein. This in fact defines the equivalent of a ''tube'' in which the protein fits, and this ''tube'' provides the global chirality. On a more general basis, this problem can be solved by including a torsion angle constraint in the energy function to be minimized. Another option would be to start the calculation with energy terms for both cRMS and dRMS, and slowly reducing the weight of the cRMS term to 0.

IV. CONCLUSION

An application of SCMF theory to modeling proteins on a lattice has been proposed. Under the usual mean field approximations, the protein residues are supposed to be independent, and their possible positions are given by a list of lattice sites; connectivity is maintained through an additional energetic term. An effective system is considered, in which

each residue i is supposed to occupy all its available sites j, each with a weight $V(i,j)$. The energy of the system is computed, and iteratively minimized with respect to the weights **V**, with the lattice sites fixed in space. The final corresponding **V** matrix describes the conformational space available to the protein, based on the energy function implemented. This energy function contains two types of terms, namely simple geometric terms required for all protein lattice simulations based on MFT, which ensure bond connectivity and limit the risks that the protein chain self-intersects, and energy terms specific to the problem of interest. The latter can include (knowledge-based) pair potentials for protein folding studies and/or energy terms based on experimental terms for protein modeling based on NMR or x-ray data.

The application of the above protocol to building a lattice model of a protein given its structure has been discussed in detail. Two measures of the similarity of two protein models have been considered, namely the coordinate root mean square deviation (cRMS), and the distance root mean square deviation (dRMS), yielding two different energy terms which could easily be incorporated within the SCMF protocol. This protocol proved to be efficient in both cases, yielding lattice models with low RMS, comparable to those obtained by other lattice fitting procedures. In the lattice model building procedure based on dRMS, we have shown that the problem of generating models which correspond to a mirror image of the true structure could be solved in most cases by limiting the conformational space available to each residue.

The unique feature of the SCMF technique described in this work is that it does not provide a single solution (i.e., a single protein conformation), rather it provides a description of the conformational space available to the protein under the constraint of the specified energetic terms. In the case of the generation of lattice models based on a known protein structure, the result of the procedure is a scoring matrix of correspondence between residues in the protein and lattice sites. The optimal lattice model based on this matrix is derived by dynamic programming. We are currently working on a extension of this idea to the protein structure comparison, which would allow the generation of a structural correspondence matrix (**SCM**) between the two proteins. Indeed, the alignment of two protein structures does not have a unique answer,[38,39] and the **SCM** matrix should be an efficient way to generate all alternative alignments.

ACKNOWLEDGMENTS

We acknowledge encouragement and support from J.-F. Lefèvre and M. Levitt, in whose laboratories this work was carried out. We also thank Dr. Jerry Tsai and Dr. Peter David for providing useful comments on the manuscript. Part of this work was conducted at Stanford, while P.K. was a recipient of an American Cancer Society (ACS) fellowship, managed by the Union Internationale Contre le Cancer (UICC, Geneva, Switzerland).

[1] C. Anfinsen, Science **181**, 223 (1973).
[2] C. Pabo, Nature (London) **301**, 200 (1983).
[3] T. L. Blundell, B. L. Sibanda, M. J. E. Sternberg, and J. M. Thornton, Nature (London) **326**, 347 (1987).
[4] N. Summers, W. Carson, and M. Karplus, J. Mol. Biol. **196**, 157 (1987).
[5] N. Summers and M. Karplus, J. Mol. Biol. **210**, 785 (1989).
[6] M. Levitt, J. Mol. Biol. **226**, 507 (1992).
[7] K. A. Dill, S. Bromberg, K. Z. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan, Protein Sci. **4**, 561 (1995).
[8] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[9] L. Piela, J. Kostrowicki, and H. A. Scheraga, J. Phys. Chem. **93**, 3339 (1989).
[10] P. Koehl and M. Delarue, Curr. Opin. Struct. Biol. **6**, 222 (1996).
[11] R. Elber and M. Karplus, J. Am. Chem. Soc. **112**, 9161 (1990).
[12] A. Roitberg and R. Elber, J. Chem. Phys. **95**, 9277 (1991).
[13] P. Koehl and M. Delarue, J. Mol. Biol. **239**, 249 (1994).
[14] C. Lee, J. Mol. Biol. **236**, 918 (1994).
[15] M. Vasquez, Biopolymers **36**, 53 (1995).
[16] P. Koehl and M. Delarue, Nature Struct. Biol. **2**, 163 (1995).
[17] G. M. Verkhivker and P. A. Rejto, Proc. Natl. Acad. Sci. USA **93**, 60 (1996).
[18] A. V. Finkelstein and B. A. Reva, Nature (London) **351**, 497 (1991).
[19] A. V. Finkelstein and B. A. Reva, Protein Eng. **7**, 617 (1992).
[20] A. A. Rabow and H. A. Scheraga, J. Mol. Biol. **232**, 1157 (1993).
[21] A. Finkelstein and B. Reva, Protein Eng. **9**, 387 (1996).
[22] B. A. Reva and A. V. Finkelstein, Protein Eng. **5**, 625 (1992).
[23] M. Delarue and P. Koehl, *Proceedings of the Pacific Symposium on Biocomputing*, edited by R. G. Altman, A. K. Dunker, L. Hunter, and T. Klein (World Scientific, Singapore, 1996).
[24] A. Godzik, A. Kolinski, and J. Skolnick, J. Comput. Chem. **14**, 1194 (1993).
[25] D. Rykunov, B. Reva, and A. Finkelstein, Proteins: Struct., Funct., Genet. **22**, 100 (1995).
[26] B. H. Park and M. Levitt, J. Mol. Biol. **249**, 493 (1995).
[27] B. Reva, D. Rykunov, A. Olson, and A. Finkelstein, J. Comp. Biol. **2**, 527 (1995).
[28] B. A. Reva, A. V. Finkelstein, D. S. Rykunov, and A. J. Ohlson, Proteins: Struct., Funct., Genet. **26**, 1 (1996).
[29] A. A. Rabow and H. A. Scheraga, Protein Sci. **5**, 1800 (1996).
[30] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics* (Pergamon, New York, 1958).
[31] D. E. Van den Bout and T. K. I. Miller, Biol. Cybern. **62**, 129 (1989).
[32] C. Peterson and B. Söderberg, Int. J. Neural Syst. **1**, 3 (1989).
[33] M. Hao, S. Rackovsky, A. Liwo, M. Pincus, and H. Scheraga, Proc. Natl. Acad. Sci. USA **89**, 6614 (1992).
[34] R. F. Bellman, *Dynamic Programming* (Princeton University Press, Princeton, 1957).
[35] F. C. Bernstein, T. F. Koetzle, G. Williams, D. J. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, J. Mol. Biol. **112**, 535 (1977).
[36] D. G. Covell and R. L. Jernigan, Biochemistry **29**, 3287 (1990).
[37] D. Covell, Proteins: Struct., Funct., Genet. **14**, 409 (1992).
[38] A. Godzik, Protein Sci. **5**, 1325 (1996).
[39] F. Zu-Kang and M. J. Sippl, Folding & Design **1**, 123 (1996).