

Optimization of Protein Structure on Lattices Using a Self-Consistent Field Approach

BORIS A. REVA,^{1,2} D.S. RYKUNOV,³ ALEXEI V. FINKELSTEIN,⁴ and JEFFREY SKOLNICK¹

ABSTRACT

Lattice modeling of proteins is commonly used to study the protein folding problem. The reduced number of possible conformations of lattice models enormously facilitates exploration of the conformational space. In this work, we suggest a method to search for the optimal lattice models that reproduced the off-lattice structures with minimal errors in geometry and energetics. The method is based on the self-consistent field optimization of a combined pseudoenergy function that includes two force fields: an “interaction field,” that drives the residues to optimize the chain energy, and a “geometrical field,” that attracts the residues towards their native positions. By varying the contributions of these force fields in the combined pseudoenergy, one can also test the accuracy of potentials: the better the potentials, i.e., the more accurate the “interaction field,” and the smaller the contribution of the “geometrical field” required for building accurate lattice models.

Key words: protein structure prediction, lattice modeling, self-consistent field optimization.

INTRODUCTION

LATTICE MODELING OF PROTEINS is widely used in the numerical investigations of protein folding kinetics and thermodynamics (Dashevskii, 1980; Covell and Jernigan, 1990; Hind and Levitt, 1994; Kolinski and Skolnick, 1996). The finite number of possible conformations of lattice models enormously facilitates exploration of the conformational space of a molecule. A very first problem in lattice modeling is to build a lattice model, given molecular coordinates and a lattice. This model has to be reasonably precise in two respects: it must reproduce the protein chain geometry and the protein energetics. This is not a trivial task because the model also has to satisfy the conditions of chain connectivity and self-avoidance (Godzik *et al.*, 1992; Rykunov *et al.*, 1995; Reva *et al.*, 1995; Reva *et al.*, 1996). A geometrically accurate lattice model that preserves chain connectivity can be built by dynamic programming (Rykunov *et al.*, 1995; Reva *et al.*, 1995; Reva *et al.*, 1996). Geometrically accurate self-avoiding models can be built by a self-consistent field (SCF)-based optimization of the error function that includes terms penalizing overlapping of the chain residues (Reva

¹Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037.

²On leave from Institute of Mathematical Problems of Biology, Russian Academy of Sciences, 142292, Pushchino, Moscow Region, Russian Federation.

³Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, 142292 Pushchino, Moscow Region, Russian Federation.

⁴Institute of Protein Research, Russian Academy of Sciences, 142292, Pushchino, Moscow Region, Russian Federation.

et al., 1996). In this work, we extend the approach of SCF-based optimization for building energetically and geometrically accurate models.

METHODS

(a) A combined energy function for lattice models

Let us consider a chain of N monomers (residues a_1, \dots, a_N) with 3D coordinates x_1, \dots, x_N (for simplicity, we will assume further that x_i is given by the C_α -atom coordinate of the i -th residue of a protein chain; a generalization for chains with side groups can be done according to Kolinski and Skolnick (1996) and let vectors $R_i, i = 1, \dots, N$, give the lattice points corresponding to these residues in the lattice chain model.

To optimize a lattice model of a protein chain with respect to both geometry and energy, we minimize a pseudoenergy function consisting of three terms: the terms maintaining the chain connectivity, geometrical error and chain energy.

The chain connectivity condition is included by the terms:

$$U_i(R_i, R_{i+1}) = \begin{cases} 0, & \text{if } ||R_i - R_{i+1}| - d_i| \leq \gamma \\ +\infty, & \text{otherwise} \end{cases}, \quad (1)$$

where $i = 1, \dots, N - 1$. In this expression, $d_i = |x_i - x_{i+1}|$ is the actual distance between residues i and $i + 1$ (for a protein α -carbon chain without cis-Pro residues $d = 3.8\text{\AA}$) and γ limits the allowed deviation of inter-residue distance in the lattice model from its actual value; in this work, $\gamma = \Delta/2$, Δ is a lattice spacing.

The geometrical error function (Rykunov *et al.*, 1995; Reva *et al.*, 1995; Reva *et al.*, 1996), presented as:

$$E_{err}(R_1, \dots, R_N) = \sum_{i=1}^N f_i(R_i), \quad (2)$$

where

$$f_i(R) = \begin{cases} (x_i - R)^2 & \text{if } R \text{ is one of the lattice points surrounding } x_i \\ +\infty, & \text{otherwise} \end{cases}, \quad (3)$$

gives the deviation of coordinates of the model from the actual 3D structure. The smaller the error function, the better the model. One can see that the standard root mean squared deviation (RMSD) of the lattice model with respect to the native structure is $(E_{err}/N)^{1/2}$.

The condition that R_i must be one of the lattice points surrounding x_i is specified for computational efficiency. (Theoretically, one can consider all lattice points as being allowed for each of the chain residues.)

In this study, we allow the points R_i to belong to only the first shell of 8 lattice points surrounding the point x_i . Our experiments on a simple cubic lattice of 3.8\AA spacing show (see below) that the first shell is sufficient for building continuous lattice models.

To simultaneously take into account both lattice model energy, $E\{a_1, R_1, \dots, a_N, R_N\}$, and geometrical accuracy, we suggest the following combined pseudoenergy function:

$$V\{a_1, R_1, \dots, a_N, R_N\} = \sum_{i=1}^{N-1} U_i(R_i, R_{i+1}) + AE\{a_1, R_1, \dots, a_N, R_N\} + (I - A) \sum_{i=1}^N f_i(R_i) \quad (4)$$

It is easy to see that by changing A from 0 to 1 , one can scan all the possible cases between the most geometrically accurate models and the lowest energy models.

Thus, the problem is to find the minimum of the pseudoenergy V :

$$\min_{R_1} \dots \min_{R_N} V(a_1, R_1, \dots, a_N, R_N) = V(a_1, R_1^*, \dots, a_N, R_N^*) \quad (5)$$

and to obtain the lattice model coordinates R_1^*, \dots, R_N^* corresponding to this minimum. Then one can find the model chain energy $E\{a_1, R_1^*, \dots, a_N, R_N^*\}$ and the geometrical error $E_{err} = E_{err}(R_1^*, \dots, R_N^*)$.

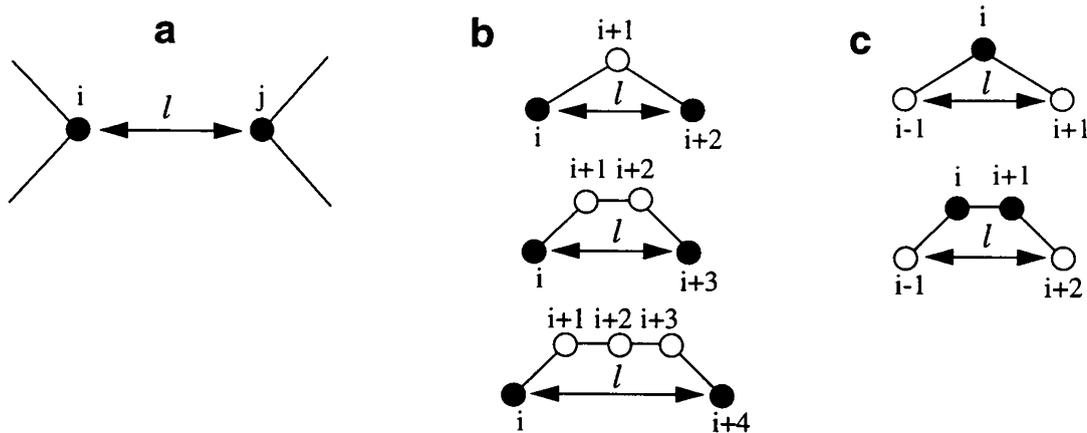


FIG. 1. Long-range and short-range interactions: residues for which potentials are derived are shown by filled circles. (a) long-range interactions: $|i - j| \geq 5$; the potential $\varepsilon_{a_i a_j}(l)$ depends on the distance l between remote chain residues and on the chemical identity of these residues a_i and a_j ; (b) short-range potentials $h_{a_i a_{i+2}}^{(2)}(|R_i - R_{i+2}|)$, $h_{a_i a_{i+3}}^{(3)}(|R_i - R_{i+3}|)$, $h_{a_i a_{i+4}}^{(4)}(|R_i - R_{i+4}|)$ depend on the distance between terminal residues and their chemical sorts; (c) short-range potentials $b_{a_i}^{(2)}(|R_{i-1} - R_{i+1}|)$ and $b_{a_i a_{i+1}}^{(3)}(|R_{i-1} - R_{i+2}|)$ depend on chain bending in the intervening residues i (or i and $i + 1$) that affects the distance between the terminal residues $i - 1, i + 1$.

In this study for energy calculations, we use our recently derived lattice-adapted potentials (Reva *et al.*, 1997), which take into account both long- and short-range interactions, as shown in Figure 1. With these energy functions, the lattice model energy is presented as:

$$\begin{aligned}
 E\{a_1, R_1, \dots, a_N, R_N\} = & \sum_{i=1}^{N-5} \sum_{j=i+5}^N \varepsilon_{a_i a_j}(|R_i - R_j|) + \sum_{i=1}^{N-2} h_{a_i a_{i+2}}^{(2)}(|R_i - R_{i+2}|) \\
 & + \sum_{i=1}^{N-3} h_{a_i a_{i+3}}^{(3)}(|R_i - R_{i+3}|) + \sum_{i=1}^{N-4} h_{a_i a_{i+4}}^{(4)}(|R_i - R_{i+4}|) \\
 & + \sum_{i=1}^{N-2} b_{a_i}^{(2)}(|R_{i-1} - R_{i+1}|) + \sum_{i=1}^{N-3} b_{a_i a_{i+1}}^{(3)}(|R_{i-1} - R_{i+2}|). \quad (6)
 \end{aligned}$$

The energy terms ε , $h^{(2)}$, $h^{(3)}$, $h^{(4)}$, $b^{(2)}$ and $b^{(3)}$ are all described in the legend to Figure 1.

(b) SCF-based optimization of the energy function

The energy function in the form given by Equation (4) cannot be minimized (as in Hind and Levitt (1994) and Kolinski and Skolnick (1996)) using dynamic programming because the “long-range” terms of (6) depend on coordinates of non-neighbor residues. However, one can use a self-consistent field (SCF) theory (Finkelstein and Reva, 1996; Reva *et al.*, 1996; Kubo, 1965) to minimize such a function.

The idea of the SCF approximation is to represent the result of the pairwise residue interactions as a modification of the 3D fields acting on the residues. When these 3D fields replace the long-range interactions of the residues, the effective chain energy has the form:

$$V^{eff}(R_1^*, \dots, R_N^*) = \sum_{i=1}^{N-1} U_i(R_i, R_{i+1}) + \sum_{i=1}^N \Psi_i(R_i). \quad (7)$$

Here

$$\Psi_i(R) = (I - A)f_i(R) + A\Delta\Psi_i(R) \quad (8)$$

is the potential acting on a residue i at a point R . The term $\Delta\Psi_i$ is the average potential “felt” by a residue i at a point R under a given distribution of the other residues in space.

The distribution of residues is given by functions $\{W_i(R)\}$, $(i = 1, \dots, N)$; $W_i(R)$ is a probability that residue i occupies lattice point R . The force field created by this distribution of residues in space is given by the potential:

$$\begin{aligned} \Delta\Psi_i(R) = & \sum_{\substack{j \neq i, \\ i-1, i+1}}^N \sum_r \varepsilon_{a_i a_j} (|R-r|) W_j(r) + \sum_{k=2}^4 \Theta(i-k) \sum_r h_{a_i-k a_i}^{(k)} (|R-r|) W_{i-k}(r) \\ & + \sum_{k=2}^4 \Theta(N-i-k+1) \sum_r h_{a_i a_{i+k}}^{(k)} (|R-r|) W_{i+k}(r) \\ & + \Theta(i-2) \sum_r b_{a_i-1}^{(2)} (|R-r|) W_{i-2}(r) + \Theta(N-i-1) \sum_r b_{a_{i+1}}^{(2)} (|R-r|) W_{i+2}(r) \\ & + \Theta(i-3) \sum_r b_{a_{i-2} a_{i-1}}^{(3)} (|R-r|) W_{i-3}(r) + \Theta(N-i-2) \sum_r b_{a_{i+1} a_{i+2}}^{(3)} (|R-r|) W_{i+3}(r). \quad (9) \end{aligned}$$

Here

$$\Theta(l) = \begin{cases} 1, & l > 0 \\ 0, & \text{otherwise} \end{cases}.$$

For the sake of computational efficiency, we treat short- and long-range interactions in (9) equally. (In principle, it is more correct to consider short-range interactions using "generalized links" (Finkelstein and Reva, 1992), including two or three neighbor residue each.)

(c) Computation of probabilities

The form of the energy function (7) enables the efficient computation of the probabilities $\{W_i^\Psi(R)\}$ provided the potentials $\{\Psi_i(R)\}$ are given (Finkelstein and Reva, 1992; Rykunov *et al.*, 1995).

For any link i

$$W_i^\Psi(R) = P_i^\Psi(R) Q_i^\Psi(\bar{R}) / Z^\Psi \quad (10)$$

where $P_i^\Psi(R)$ is the partition function of a chain fragment preceding residue i fixed in the position R , $Q_i^\Psi(R)$ is the partition function of the following chain fragment, and Z^Ψ is the total partition function of the chain. The advantage of the energy function in the form of (7) is that one can calculate the partition functions from the recursion relations:

$$\begin{cases} P_1^\Psi(\bar{R}_1) = 1 \\ P_i^\Psi(R_i) = \sum_{R_{i-1}} P_{i-1}^\Psi(R_{i-1}) \exp[-(\Psi_{i-1}(R_{i-1}) + U_{i-1}(R_{i-1}, R_i)) / T], \end{cases} \quad (11)$$

$$i = 2, \dots, N,$$

and

$$\begin{cases} Q_N^\Psi(R_N) = \exp[-\Psi_N(R_N) / T] \\ Q_i^\Psi(R_i) = \exp[-\Psi_i(R_i) / T] \sum_{R_{i+1}} Q_{i+1}^\Psi(R_{i+1}) \exp[-U_i(R_i, R_{i+1}) / T], \end{cases} \quad (12)$$

$$i = N-1, \dots, 1,$$

while

$$Z^\Psi = \sum_R P_1^\Psi(R) Q_1^\Psi(R). \quad (13)$$

The sums in (11) and (12) are taken over all positions of links $i-1$ and $i+1$, respectively; T is the temperature, the value of which is considered below (the temperature is introduced to avoid trapping the system in local energy minima).

(d) Iterative solution of the SCF equations

The self-consistent solution of (8)–(13) is obtained when the computed probabilities $\{W_i^{*\prime}(R)\} = \hat{W}\{\Psi_i^*\}$ coincide with the probabilities $\{W_i\}$ used to calculate the potentials $\{\Psi_i^*(R)\} = \hat{\Psi}\{W\}$.

(A) Initialization

$$(i) \left\{ \Delta\Psi_1^0, \dots, \Delta\Psi_N^0 \right\} \equiv 0$$

$$(ii) \left\{ W_1^0, \dots, W_N^0 \right\}_{\Psi^0} = \hat{W}\{\Psi^0\} \quad \text{Eqs.(10)-(13)}$$

$$(iii) \left\{ \Delta\tilde{\Psi}_1^0, \dots, \Delta\tilde{\Psi}_N^0 \right\}_{W^0} = \Delta\tilde{\Psi}\{W^0\} \quad \text{Eq.(9)}$$

$$(iv) F^0 = -T\text{Ln}(Z^0) + \frac{1}{2} \sum_{i=1}^N \sum_{R_i} \Delta\tilde{\Psi}_i^0 W_i^0 - \sum_{i=1}^N \sum_{R_i} \Delta\Psi_i^0 W_i^0$$

$$s = 1$$

(B) Iterative descent to the free energy minimum

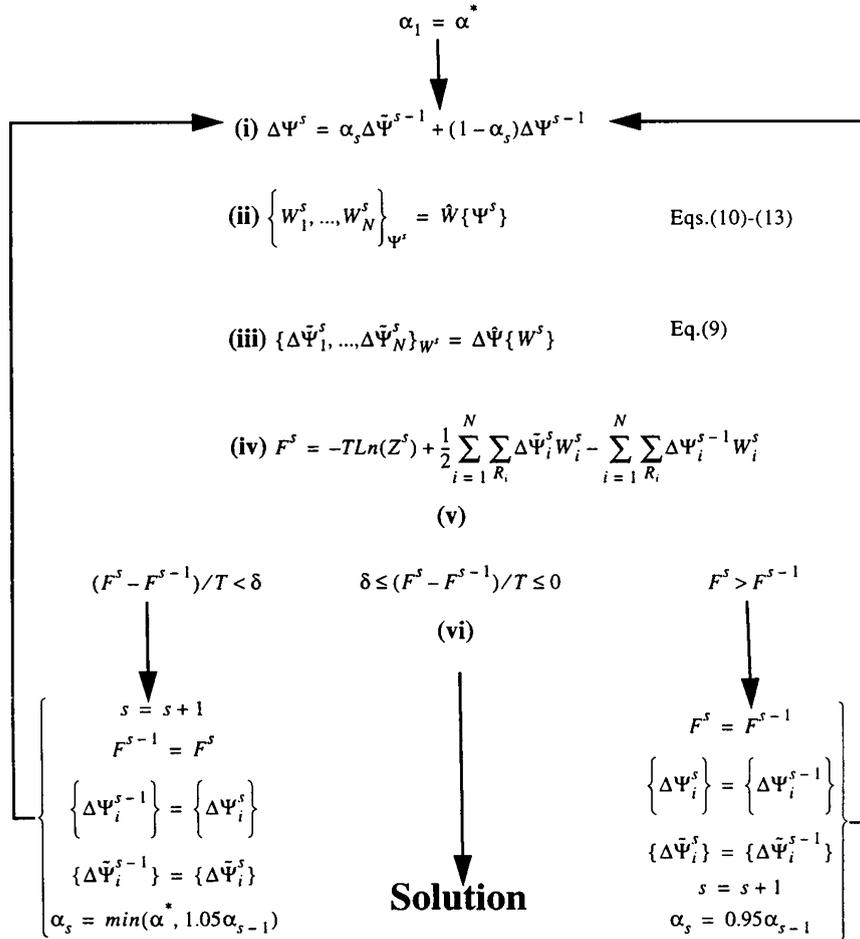


FIG. 2. The flow chart of SCF optimization. The iterative search for the SCF solution consists of (A) initialization and (B) iterative descending to the free energy minimum. Initialization is done in 4 steps: (i) one assigns starting potentials (typically we take starting potentials to be equal to zero, however, in the general case, they can be assigned arbitrary or, depending on the task, chosen more specifically); (ii) the probabilities $\{W_i^0, i = 1, \dots, N\}$ are computed by Equations (10)–(13); (iii) the new potential $\{\tilde{\Psi}_i^0, i = 1, \dots, N\}$ are computed according to Equation (9); and, finally, (iv) the initial free energy is computed (Finkelstein and Reva, 1996). The iterative descending procedure (B) is organized into 6 steps: (i) computation of the current iterative potentials (typically $\alpha^* = 0.9$); computation of (ii) the new probabilities; (iii) the new potentials, and (iv) the new free energy; (v) comparison of the free energies obtained in the current and in the previous iterations, and (vi) stopping of the procedure if the reduction of the free energy is smaller than the convergence parameter δ (typically $\delta = -0.05$ in the energy units) or reassigning of the potentials, the free energy and the iterative parameter α , if convergence of the free energy is not achieved.

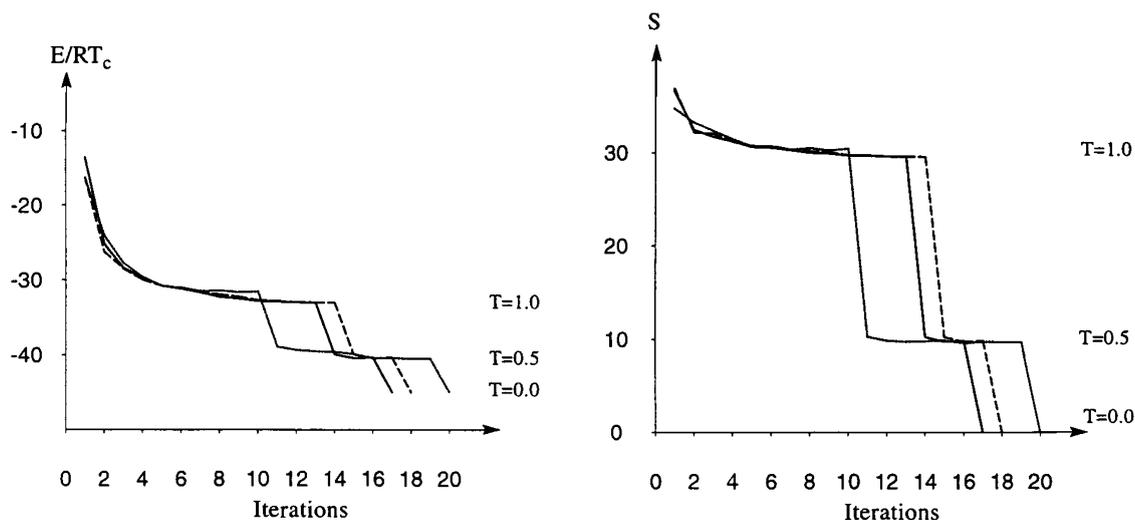


FIG. 3. Energy and entropy of a crambin molecule as a function of number of iterations in SCF-optimization. The calculations are done with the combined energy function (5) at $A = 0.9$ on a simple cubic lattice of spacing 3.8\AA . The protocols shown by filled, dashed and dotted lines correspond to three different randomly assigned starting fields; the temperature protocol used in annealing: $T = 1.0$ until the SCF solution is obtained; then $T = 0.5$ until the new SCF-solution is not obtained; and then $T = 0$.

According to the general theory (Finkelstein and Reva, 1996) the self-consistent solution corresponds to the minimum of the free energy F :

$$F^* = -T \ln(Z^*) - \frac{1}{2} \sum_{i=1}^N \sum_{R_i} \Delta \Psi_i^* W_i^* \quad (14)$$

The iterative algorithm used to search for the free energy minimum is given in Figure 2.

The self-consistent solution is obtained when the free energy no longer changes upon iteration; typically it takes ~ 20 – 40 iterations.

In principle, the SCF solution can depend on the starting field. Below, we show that this dependence is minor in our calculations.

(e) A temperature protocol for the lowest energy lattice models search

A solution of SCF-equations at non-zero temperatures is a set of probabilities $\{W_i(R)\}$, $i = 1, \dots, N$. To get a unique model, one needs to decrease the temperature to zero. Thus, one needs to use an annealing procedure: one starts at some $T \neq 0$, obtains the corresponding SCF solution, then decreases temperature, obtains a new solution, etc., until zero temperature is reached. To calculate the chain distribution at $T = 0$, we use the statistical mechanics of $1D$ systems especially adapted to zero temperature (Finkelstein and Reva, 1992). This approach finds all the lowest energy pathways while taking into account a possibility of ground state degeneracy. Finally, dynamic programming (Rykunov *et al.*, 1995) singles out only one of the lowest energy chain pathways.

RESULTS AND DISCUSSION

(a) Choice of the temperature protocol

In general, the chain models found by the SCF-optimization depend on the starting field and on the temperature protocol. This dependence results from using a self-consistent field approximation (Finkelstein and Reva, 1996; Kubo, 1965). Only the “interaction field” (see (9)) is the subject of this approximation. When $A = 0$, (4), this field does not act; in this case one can take $T = 0$ from the very beginning (or, as that it is absolutely the same, one uses dynamic programming) and come to the lowest pseudoenergy solution.

When $A \neq 0$, different temperature protocols have different qualities. For example, one should not start at too low a temperature because this will trap the molecule in local minima. Our experiments show that one has to start with a moderate temperature (e.g., with $T = 1$) and decrease this temperature gradually. Figure 3

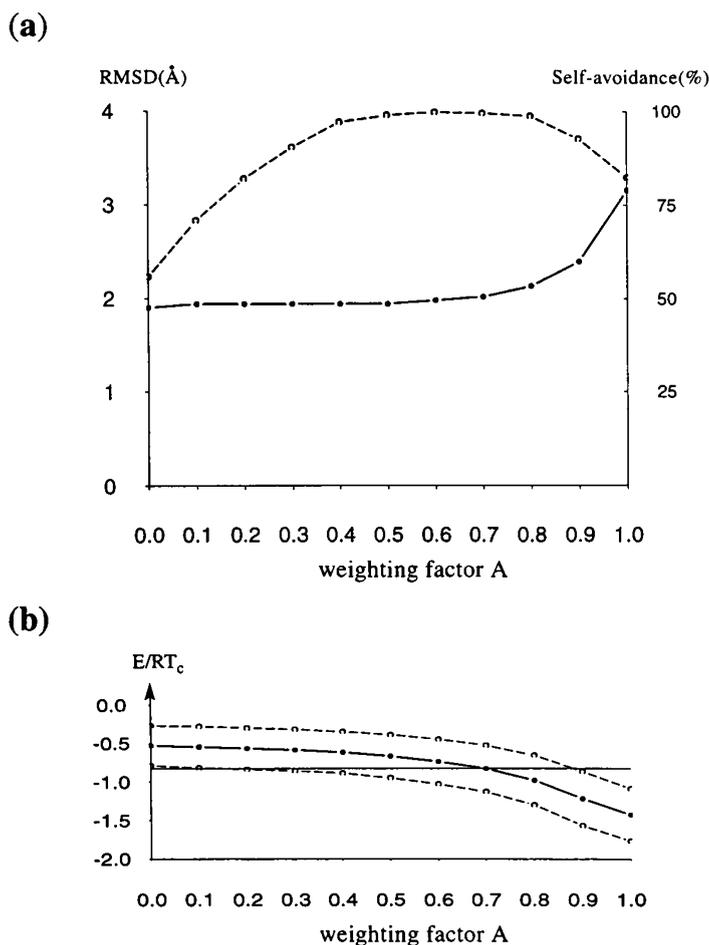


FIG. 4. (a) RMSD (filled circles) and percentage of self-avoiding models (open circles) as a function of the weighting factor A . The results are averaged over 10 proteins (PDB codes: 1crn (46); 1dtx (58); 1ptx (64); 2ctx (71); 1cks (78); 2bop (85); 7pcy (98); 3sic (107); 1bnd (109); 2rsl (120); chain lengths are given in parentheses); for each of the proteins 100 lattice models were built on a lattice of spacing 3.8\AA using the constraint of 8 allowed lattice points per residue; (b) Averaged per residue energy (filled circles) and the range of dispersion (shown by the dashed lines) as a function of the weighting factor A ; the mean energy of the actual off-lattice structures (-0.83) is given by the thin line.

shows energy and entropy changes in the course of iterations for the temperature protocol, which turned out to be one of the best.

We checked this temperature protocol using 100 different randomly chosen starting fields and found that for a crambin molecule at $A = 1$, the dispersion of the lattice model energies is ~ 1 (in RT_c units). This is smaller than the energy variation caused by different lattice-protein orientations ~ 4 .

(b) Search for the optimal models

In Figure 4, we present averaged geometrical accuracy (RMSD), average residue energy and dispersion, and also a fraction of self-avoiding models as functions of the weighting factor A of the pseudoenergy (see (4)).

One can see that at $A \approx 0.7$, the SCF-based optimization algorithm builds self-avoiding and rather accurate (geometrically and energetically) lattice models. However, one can see also that at $A \geq 0.8$ when the energy term dominates the pseudoenergy (4), the protein chain chooses an optimal lattice model that is rather far from the true off-lattice chain pathway, i.e., RMSD is large. However, the energy of such a lattice model is significantly lower than the true off-lattice energy. The RMSD of the lattice models built at $A \sim 1$ approaches the maximal possible deviation $3.2\text{\AA} \sim \Delta$ when 8 lattice points of the first shells are allowed per residue, and even greater when two ($4^3 = 64$ lattice points) or three shells ($6^3 = 216$ lattice points) are allowed. Figure 4 also shows the reduction in the number of self-avoiding models at $A \geq 0.8$, i.e., when the energy term dominates. One of the reasons for this reduction is the use of a too narrow "tube" (only 8 lattice points per residue in

width) allowed to the chain to search for the lowest energy conformation. However, a deeper reason is that the employed lattice potentials are not accurate enough to select the native structure, although they gave quite reasonable results in recognition of the native structure in threading (Reva *et al.*, 1995). Thus, the SCF-based optimization algorithm for lattice model building appears to provide a more severe test for lattice potentials.

CONCLUSIONS

In this work, we have suggested and tested the new approach for building lattice models of protein structure. The method builds lattice models using an SCF-based optimization of the combined pseudoenergy energy function, which includes both potential energy and geometrical constraint terms (error function).

We have found the optimal combination of the energy and the geometrical constraints and have shown that one can reproduce off-lattice structures with minimal errors in geometry and energetics. The obtained models can be used as target structures in protein folding simulations held on 3D lattices. The approach also allows one to measure the difference between geometrically accurate and lowest energy lattice models and tests the accuracy of potentials for protein structure prediction.

ACKNOWLEDGMENTS

This work was supported by NIH Grant GM48835 (to JS) and by NIH Fogarty Research Collaboration Grant No. TW00546 (to JS and AVF). AVF acknowledges support by an International Research Scholar's Award No. 75195-544702 from the Howard Hughes Medical Institute.

REFERENCES

- Covell, D., and Jernigan, R. 1990. Conformations of folded proteins in restricted spaces. *Biochemistry* 29, 3287–3294.
- Dashevskii, V.G. 1980. A lattice model for the protein three-dimensional structure. *Mol. Biol. (Moscow)* 14, 105–117.
- Finkelstein, A.V., and Reva, B.A. 1992. A search for the stable state of a short chain in a molecular field. *Prot. Eng.* 5, 617–624.
- Finkelstein, A.V., and Reva, B.A. 1996. A search for the most stable folds of protein chains: Application of a self-consistent molecular field theory to a problem of protein three-dimensional structure prediction. *Prot. Eng.* 9, 387–397.
- Godzik, A., Kolinski, A., and Skolnick, J. 1992. Lattice representation of globular proteins: How good are they? *J. Com. Chem.* 14, 1194–1202.
- Hind, D., and Levitt, M. 1994. Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.* 243, 668–682.
- Kolinski, A., and Skolnick, J. 1996. *Lattice Models of Protein Folding, Dynamics and Thermodynamics*. R.G. Landes Co., Austin.
- Kubo, R. 1965. *Statistical Physics*. North-Holland Publishing, Amsterdam.
- Reva, B.A., Rykunov, D.S., Olson, A.J., and Finkelstein, A.V. 1995. Constructing of lattice models of protein chains with side groups. *J. Com. Biol.* 2, 527–535.
- Reva, B.A., Finkelstein, A.V., Rykunov, D.S., and Olson, A.J. 1996. Building self-avoiding lattice models of proteins using a self-consistent field optimization. *Proteins* 26, 1–8.
- Reva, B.A., Finkelstein, A.V., Sanner, M.F., Olson, A.J., and Skolnick, J. 1997. Recognition of protein structure on coarse lattices with residue-residue energy functions. *Prot. Eng.* 10, 1123–1130.
- Rykunov, D.S., Reva, B.A., and Finkelstein, A.V. 1995. Accurate general method for lattice approximation of three-dimensional structure of a chain molecule. *Proteins* 22, 100–109.

Address reprint requests to:

Boris A. Reva
Department of Molecular Biology
The Scripps Research Institute
10550 North Torrey Pines Road
La Jolla, CA 92037

This article has been cited by:

1. Y. Ponty, R. Istrate, E. Porcelli, P. Clote. 2008. LocalMove: computing on-lattice fits for biopolymers. *Nucleic Acids Research* **36**:Web Server, W216-W222. [[CrossRef](#)]