A self-consistent field optimization approach to build energetically and geometrically correct lattice models of proteins

Boris A. Reva[1,2], Alexei V. Finkelstein[3] and Jeffrey Skolnick[1]

Abstract

Lattice modeling of proteins is commonly used in studying the protein folding problem. A finite 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", which drives the residues to optimize the chain energy, and a "geometrical field", which attracts the residues towards their native positions. By varying the contributions of these force fields in the combined pseudoenergy, one can also use the model building to test the accuracy of potentials: the better the potentials i.e., the more accurate the "interaction field", the smaller contribution of the "geometrical field" is required for building accurate lattice models.

Introduction

Lattice modeling of proteins is widely used in the numerical investigations of protein folding kinetics and thermodynamics[1-4]. A finite number of possible conformations of lattice models enormously facilitates exploration of the conformational space of a molecule.

The 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: in reproducing the protein chain geometry and in reproducing the protein energetics. This is not a trivial task because the model also has to satisfy the conditions of chain connectivity and self-avoiding[5-8]. A geometrically accurate lattice model that preserves chain connectivity can be built by dynamic programing[6-7]. Geometrically accurate self-avoiding models can be built by a self-consistent field (SCF)-based optimization of the error function which includes terms penalizing overapping of the chain residues[8].

In this work, we extend the approach of SCF-based optimization for building energetically and geometrically accurate models. The method is demonstrated on a coarse cubic lattice of spacing 3.8Å, which is difficult for building

[1]Department of Molecular Biology, The Scripps Research Institute, 10666 North Torrey Pines Road, Ca 92037, USA; (tel: 619 7848831, fax: 6197848895)

[2]on leave from Institute of Mathematical Problems of Biology Russian Academy of Sciences, 142292, Pushchino, Moscow Region, Russian Federation;

[3]Institute of Protein Research, Russian Academy of Sciences, 142292, Pushchino, Moscow Region, Russian Federation

RECOMB 98 New York NY USA
Copyright 1998 0-89791-976-9/98/3..$5.00

satisfactory models.

Methods

(a) A combined energy function for lattice models

Let us consider a chain of N monomers (residues $a_1,...,a_N$) with 3D coordinates $x_1,...,x_N$ (for simplicity we will assume further that x_i is given by C_α-atom coordinate of *i*th residue of a protein chain; a generalization for chains with side groups can be done according to Ref.7) and let, vectors R_i, $i = 1,...,N$, 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 term maintaining the chain connectivity, geometrical error and chain energy.

The chain connectivity condition is included by the terms:

$$U_i(R_i, R_{i+1}) = \begin{cases} 0, & if \ \ \left|\left|R_i - R_{i+1}\right| - d_i\right| \le \gamma \\ +\infty, & otherwise \end{cases} \qquad (1)$$

where $i = 1,...,N-1$. In this expression $d_i = \left|x_i - x_{i+1}\right|$ is the actual distance between residues i and $i+1$ (for a protein α-carbon chain without cys-Pro residues d=3.8 Å) 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[6-8], presented as:

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

where

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

gives the deviation 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 of the lattice model with respect to the native structure (RMSD) 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 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 show (see below) that the first shell is sufficient for building continuous lattice models.

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

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

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*;

$$\min_{R_1} \dots \min_{R_N} V\{a_1, R_1,..., a_N, R_N\} = V\left\{a_1, R_1^*,..., a_N, R_N^*\right\} \qquad (5)$$

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

In this study for energy calculations we use our recently derived lattice-adapted potentials[9] which take into account both long- and short- range interactions, as shown in Fig.1. With these energy functions, the lattice model energy is presented as:

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

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

(b) SCF-based optimization of the energy function

The energy function in the form given by Equation (4) cannot be minimized (as in Refs.6-7) using dynamic programming because the "long-range" terms of Equation (6) depend on coordinates of non-neighbor residues.

However, one can use a self-consistent field (SCF) theory[8,10,11] 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:

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

Here $\Psi_i(R) = f_i(R) + \Delta\Psi_i(R)$ is the potential acting on a residue i at a point R. The term $\Delta\Psi_i$ (which modifies the potential f_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 $\{W_j(R)\}$, *(j=1,...,N)*; $W_j(R)$ is a probability that residue j occupies lattice point R. The force field created by this distribution of residues in space is given by the potential:

$$
\begin{aligned}
\Delta\Psi_i(R) = &\sum_{\substack{j\neq i,\\ i-1,i+1}}^{N} \sum_r \varepsilon_{a_i a_j}(|R-r|) W_j(r) + \\
&\sum_{k=2}^{4} \Theta(i-k) \sum_r h^{(k)}_{a_{i-k} a_i}(|R-r|) W_{i-k}(r) + \\
&\sum_{k=2} \Theta(N-i-k+1) \cdot \\
&\sum_r h^{(k)}_{a_i a_{i+k}}(|R-r|) W_{i+k}(r) + \\
&\Theta(i-2) \sum_r b^{(2)}_{a_{i-1}}(|R-r|) W_{i-2}(r) + \\
&\Theta(N-i-1) \sum_r b^{(2)}_{a_{i+1}}(|R-r|) W_{i+2}(r) + \\
&\Theta(i-3) \sum_r b^{(3)}_{a_{i-2} a_{i-1}}(|R-r|) W_{i-3}(r) + \\
&\Theta(N-i-2) \sum_r b^{(3)}_{a_{i+1} a_{i+2}}(|R-r|) W_{i+3}(r)
\end{aligned}
\qquad (8)
$$

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

For the sake of computational efficiency we treat short-range and long-range interactions in Equation (9) equally. The form of the energy function (Eq.7) enables the use of 1D statistical mechanics of chain molecules to compute the probabilities $\{W_i^{\Psi}(R)\}$ provided the potentials $\{\Psi_i(R)\}$ are given. The corresponding algorithm can be found in Ref.8, Equations (7)-(13). As a result, one obtains $\{W_i^{\Psi}(R)\}$, the probability functions for distribution of residues in field $\{\Psi_i\}$, and free energy, F, of this distribution. This self-consistent solution can be found iteratively: one starts with some initial field $\Delta\Psi_i^{s=1}(R)$, $(i=1,..,N)$, i.e. with $\Delta\Psi_i^{s=1}(R) = 0$ (or with randomly generated $\Delta\Psi_i^{s=1}(R)$), and obtains the $\{W_i^{s=1}\}$ probabilities. In the next step of the iteration $(s \geq 2)$ one takes $\Delta\Psi_i^{s}(R) = \alpha\Delta\Psi_i^{W^{s-1}}(R) + (1-\alpha)\Delta\Psi_i^{s-1}(R)$ where $0<\alpha<1$ (usually $\alpha \sim 0.5-0.9$).

Then one uses the updated $\{\Delta\Psi_i^s\}$ values to obtain $\{\Psi_i^s\}$ potentials and then calculates $\left\{W_i^{\Psi^s}\right\}$, $\{\Delta\Psi_i^s\}$ values. According to general theory[10], before self-consistency is attained, F always decreases with the iteration number s when α value is chosen correctly, that is small enough. As a rule, we take α=0.8, but sometimes have to decrease it to α=0.3 when otherwise we observe an increase of F^s compared with F^{s-1}.

The self-consistent solution is obtained when the probabilities $\{W_i^{\Psi}\}$ no longer change upon iteration. This means that a self-consistent field is found and free energy minimum is achieved; 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.

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

A solution of SCF-equations at non-zero temperatures is a set of probabilities $\{W_i(R)\}$, $i=1,...N$. To get a unique model one needs to decrease a temperature to zero. Thus, one needs to use an annealing procedure: to start at some $T \neq 0$,

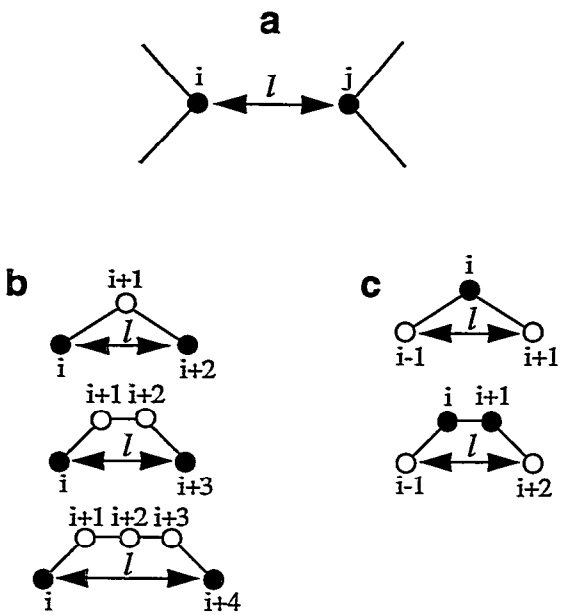

Fig. 1. Long-range and short-range interactions: residues for which potentials are derived are shown by filled circles. (**a**) long-range interactions: $|i-j| \geq 5$; the potential $\varepsilon_{a_i a_j}(l)$ depends on the distance l between remote chain residues and on chemical sorts of these residues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

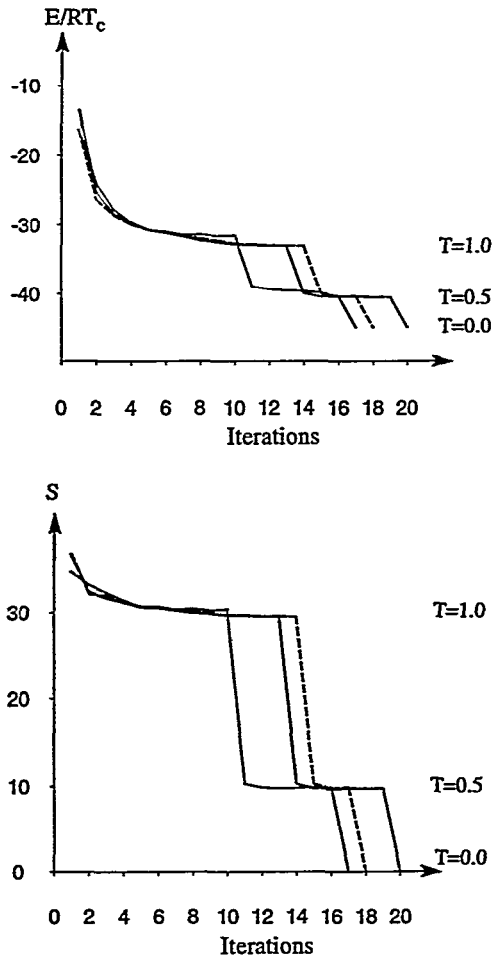

Fig. 2. Energy and entropy of a crambin molecule as a function of number of iterations in SCF-optimization. The calculation are done with the combined energy function (5) at A=0.9 on a lattice of 3.8Å. The protocols shown by filled, dashed and dotted lines correspond to three different randomly assigned 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

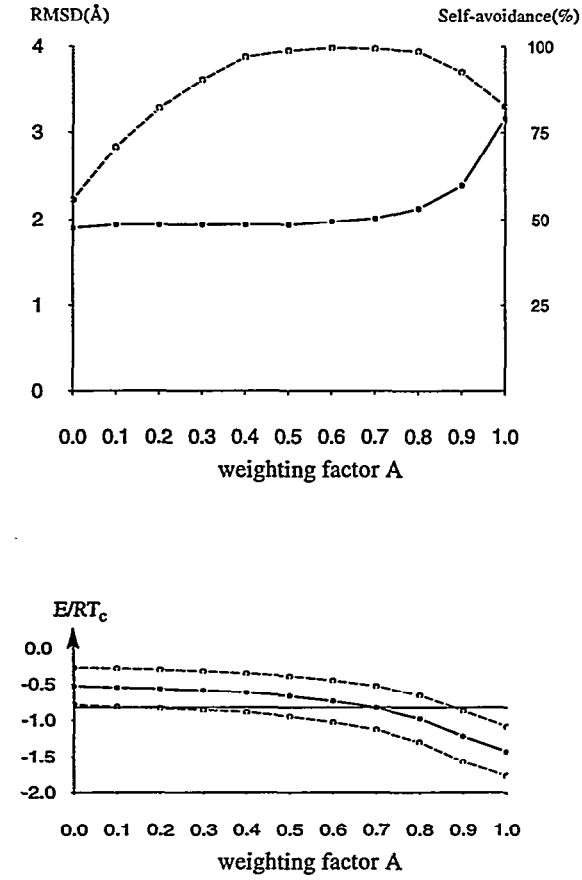

Fig. 3. (a) RMSD (filled circles) and percentage of self-avoiding models (open circles) as a function of the weighting factor A. The results are averaged over 10 proteins (PDB codes: 1crn (46); 1dtx (58); 1ptx (64); 2ctx (71); 1cks (78); 2bop (85); 7pcy (98); 3sic (107); 1bnd (109); 2rsl (120); chain lengths are given in parentheses); for each of the proteins 100 lattice models were built on a lattice of spacing 3.8Å 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.

to obtain the corresponding SCF solution, then decrease temperature, obtain 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[12]. This approach finds all the lowest energy pathways while taking into account a possibility of ground state degeneracy. Finally, dynamic programming 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 dependency results from using of a self-consistent field approximation[10-11].

Only the "interaction field" (see Equation (9)) is a subject of this approximation. When $A = 0$, (Equation 4), this field does not act; in this case one can take $T = 0$ from the very beginning (or that is absolutely the same, one use 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 temperatures because it 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 2 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 for a crambin molecule at $A = 1$ that 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 orientation ~4.

(b) Search for the optimal models

In Figure 3 we present averaged geometrical accuracy (RMSD), average residue energy and dispersion, and also a fraction of self-avoiding allowed to the chain to search for the lowest energy conformation. However, a deeper reason is that the employed lattice potentials are not models as functions of the weighting factor A of the pseudoenergy (see Eq.4). One can see that at $A \approx 0.7$ the SCF-based optimization algorithm builds a 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 which rather far from the true off-lattice chain pathway, i.e. RMSD is large. 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 $3.2\mathring{A} \sim \Delta$ 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. Fig.3 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) accurate enough to select the native structure, although they gave quite reasonable results in recognition of the native structure in threading[9].

Thus, the SCF-based optimization algorithm for lattice model building appears to provide a more severe test for lattice potentials.

Conclusion

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

Acknowledgments

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

References

[1] Dashevskii, V.G. Mol.Biol. (Moscow) 14, 105, (1980)

[2] Covell, D., Jernigan, R. *Biochemistry* **29**, 3287, (1990)

[3] Hind, D., Levitt, M. *J.Mol.Biol.* **243**, 668, (1994)

[4] Kolinski, A., Skolnick, J. *"Lattice Models of Protein Folding, Dynamics and Thermodynamics"*, R.G.Landes Co., Austin, (1996)

[5] Godzik, A., Kolinski, A., Skolnick, J. *J.Com.Chem.* **14**, 1194, (1992)

[6] Rykunov, D.S., Reva, B.A., Finkelstein, A.V. *Proteins* **22**, 100, (1995)

[7] Reva, B.A., Rykunov, D.S., Olson, A.J., Finkelstein, A.V. *J.Com.Biol.* **2**, 527, (1995)

[8] Reva, B.A., Finkelstein, A.V., Rykunov, D.S., Olson, A.J., *Proteins* **26**, 1, (1996)

[9] Reva, B.A., Finkelstein, A.V., Sanner, M.F., Olson A.J, Skolnick, *J. Prot.Engng.* in press, (1997)

[10] Finkelstein, A.V., Reva, B.A. *Prot.Engng.* **9**, 387, (1996)

[11] Kubo, R. *"Statisical Physics"* Amsterdam: North-Holand Publishing, (1965)