Simulating Folding of Helical Proteins with Coarse Grained Models

Shoji TAKADA

Department of Chemistry, Kobe University, Kobe 657-8501, Japan

(Received October 11, 1999)

We describe how potential parameters in a coarse graind model of proteins can be optimized with use of available protein three dimensional database. With this optimized potentials, we simulated a three helix bundle protein and found that all trajectories reach at the native structure within 1 microsecond. Interestingly, a quasi-mirror image is successfully discriminated from the native topology.

§1. Introduction

Protein folding has been intensively studied for about 40 years since Anfinsen's famous experiments.[1] The problem includes two aspects; physical understanding of folding mechanisms[2),3)] and predicting the three dimensional structure. For both aspects, it is crucial to construct a model that is realistic enough to discriminate the native structure from many non-native ones *and* that is simple enough to be able to sample wide range of conformational spaces with currently available computers. Models that include all atoms and solvent molecules, such as CHARM or AMBER, are still somewhat too demanding for this purpose. On the other hand, so called minimal models that include one bead per amino acid seems to be too crude for, at least, prediction.

Recently, models that are in between above mentioned two limits have been proposed and studied. In this paper, we describe our recent work to this direction. Our model includes 4 united atoms per amino acid (3 for glycine), by which backbone dynamics is modeled quite realistically, while side chain atoms are grouped into a bead and solvent effects are taken into account only indirectly. Functional form of interactions is devised to be consistent with physico-chemical knowledge; especially solvent effects are carefully taken into account via an idea of context dependent dielectric constant.

With this model, we performed simulation of a 54 residue long *pro*tein-like peptide made of three kinds of amino acids (PRO54), where three types of amino acids include hydrophobic, polar, and flexible ones.[4] The sequence of PRO54 is designed to have three helix bundle structure imitating a laboratory designed four helix bundle of DeGrado's group. For PRO54, restricting the ranges of parameters not too different from experimentally anticipated values, we tried to tune parameters empirically so that the peptide can reach at three helix bundle form starting from any random coil structure. After months of trial, we ended up with a set of parameters that indeed enables the peptide to fold within a microsecond.

Although its promising result, we found some significantly different properties for the simulated PRO54 comparing with natural proteins.[4] Among them are 1) native

like state of PRO54 has significant residual fluctuation, in which three helix bundle form is kept, but their relative alignment changes, 2) folding-unfolding transition is much less cooperative for PRO54, and 3) the most seriously, two quasi-mirror images of three helix bundle forms have almost same stability for PRO54. The first two characters were actually observed in laboratory designed peptide of DeGrado's, too. The third one may need more explanation: For three helix bundle topology, there can be two different ways of alignment of three helices. For both, all amino acids in the core are hydrophobic ones and the surface amino acids in helices are polar ones. Thus it is natural not to be able to have energy gap between the two topology. Thinking these together, we concluded that major reason for above mentioned differences is due to three letter codes, instead of due to inappropriate modeling.

Now we go forward in trying to simulate a natural protein, namely that made of 20 types of amino acids (actually 17 amino acids exist in a protein studied in this paper, though). Apparently, the model includes many more parameters and ad hoc determination of them is hopeless. Thus, we need some systematic ways to determine them. We use an idea developed by Wolynes and his co-workers;[5] namely optimize parameters so that relative stability of the native structure against misfold ones normalized by the standard deviation of energy fluctuation is maximal. This will be discussed in detail in the next section. With the optimized potential parameters, folding simulation is performed for a three helix bundle protein, albumin binding domain with 47 residues (pdb code; 1prb). Some preliminary result is reported in §3. Conclusion is given in the last section.

§2. Optimization of energy parameters

Here, we start with a short summary of the model used; an amino acid is modeled as three backbone united atoms, NH, CH, and CO, and a bead for the side chain. The latter is located near the center of mass of non-hydrogen atoms. Molecular dynamics simulation is performed by the position Langevin equation, where the Stokes law is utilized to decide the friction coefficients of atoms. All chemical bond lengths (real for backbone, and virtual between CH and a side chain) and bond-bond angles are fixed by the LINCS algorithm,[6] which is significantly better than the so-called SHAKE, the well-known algorithm. The systematic force in the Langevin equation is calculated by the derivative of the potential function, which consists of various interactions,

$$V = V_\omega + V_\phi + V_\psi + V_{\mathrm{Rama}} + V_{\mathrm{vdW}} + V_{\mathrm{HB}} + V_{\mathrm{HP}} + V_{\mathrm{EL}}. \qquad (2\cdot1)$$

Meaning of each term is the following; (in order) the hindered rotation around ω dihedral angle (1st), that around ϕ (2nd), that around ψ (3rd), the side chain entropy effect representing the secondary structure propensity (4th), the van der Waals potential (5th), the hydrogen bonding interaction (6th), the hydrophobic interaction (7th), and the electrostatic interactions (8th). The explicit expression will be described elsewhere.

The potential function includes many energetic parameters ϵ in linear form in

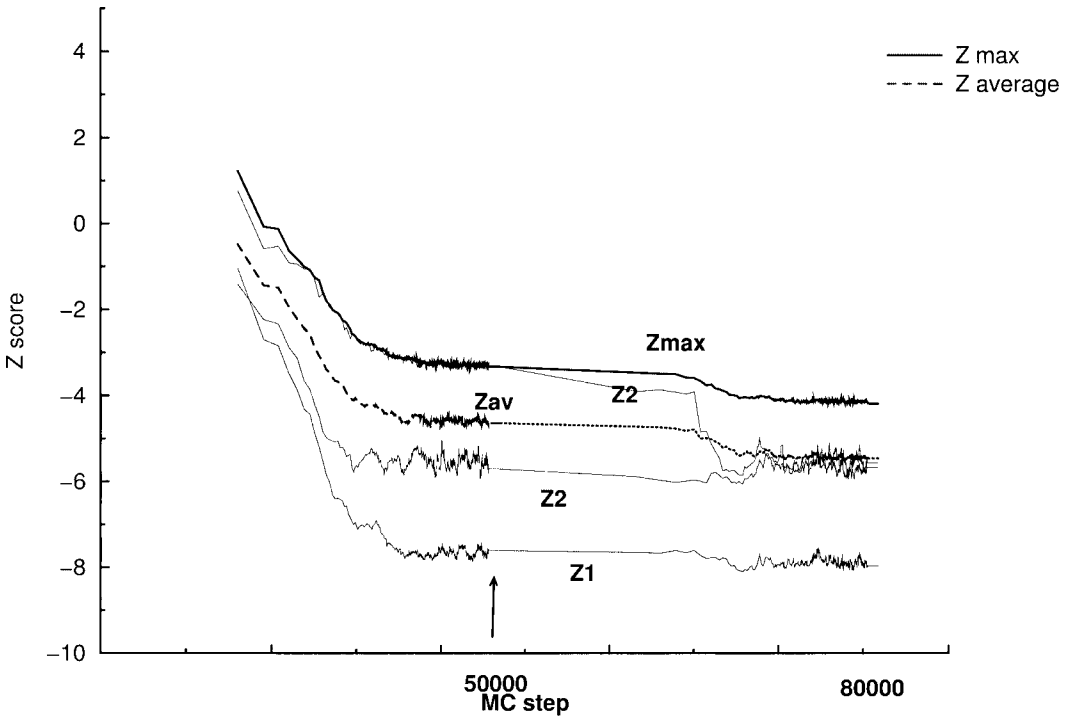

Fig. 1. Z score optimization procedure as a function of Monte Carlo steps. The top curve is for $Z_{\max}$, the dashed curve is for average of 39 Z scores, and other three are Z scores of the first three proteins out of 39's. For the first 50000 step, only the hydrophobic interaction parameters are optimized that is followed by the optimization of the rest of parameters with fixed hydrophobic ones.

the potential energy term;

$$V(r,\epsilon) = \sum_i \epsilon_i u_i(r), \tag{2·2}$$

where ϵ_i is a parameter and $u_i(r)$ is a function of protein conformation collectively denoted as r. Now we introduce the so-called Z score,

$$Z(\epsilon) = \frac{V(r_{\rm nat},\epsilon) - \langle V(r,\epsilon)\rangle_D}{\Delta V(\epsilon)}, \tag{2·3}$$

the (potential) energy of the native structure $V(r_{\rm nat})$ relative to average energy $\langle V(r)\rangle_D$ of denatured ensemble divided by the standard deviation of energy fluctuation ΔV. (Note that the opposite sign to this definition is sometimes used.) It was theoretically analyzed that for the protein to fold quickly avoiding severe trap in misfolded states the protein has to have reasonably small Z score, i.e., negative and large in absolute value in the current definition. [5)]

Our strategy to optimize parameters is as follows; We first choose some training set of proteins for which native three dimensional structures are known from experiments. We use 39 proteins in this paper. Our goal is to find *a* set of potential parameters that can be used for simulation of *any* proteins. Therefore we decided

to use the maximum value of Z score, $Z_{\max}$, as an index representing quality of the energy function. We performed simulated annealing runs in parameter space ϵ with use of $Z_{\max}$ as a scoring function; Namely, assuming some initial set of parameters ϵ, we compute Z score of training proteins and get the maximal value of them, $Z_{\max}$. We then make small change in ϵ and recompute $Z_{\max}$. Metoropolis criteria is used for $Z_{\max}$ to decide whether the change is accepted or rejected. The procedure is repeated with decreasing temperature until getting an annealed parameters.

Figure 1 represents a "trajectory" of an annealing run, where in addition to $Z_{\max}$, average of Z scores for 39 proteins, and Z scores of the first three proteins are plotted as a function of Monte Carlo step.

With use of the optimized parameter set, we computed Z scores of several small proteins that are not in the training set. The Z values are as follows; -3.48 for 1bdd (the pdb code), -2.64 for 1r69, -0.75 for 1coa, -1.20 for 2gb1, -1.39 for 1srl, and -0.82 for 1nmg where the first two are all α proteins, while the others include β sheet. Namely, for all-α proteins, the current energy function seems to be useful even if they are not involved in the training set. Unfortunately, this is not the case for proteins with β sheet.

§3. Simulating protein folding with 20 letter code: Albumin binding domain

Since the current energy function is supposed to be good for helical proteins, we performed folding simulation of a three helix bundle protein, 47 residue of albumin binding domain (6 residues are cleaved out from the sequence in the pdb file 1prb). We tune up one parameter that is responsible for strength of overall collapse. After tuning up this one parameter, we found that almost all trajectories (13/14 at this moment) can reach at the native like structure within $1\mu s$, starting from random conformations. The simulated native structure, after quenching, has about 3 A root mean square deviation (RMSD) from the experimental structure. Figure 2 shows snapshots of a typical folding trajectory; after several tens of nanoseconds, collapse and helix formation simultaneously occured. After forming about-right topology, it takes somewhat long time to reach at the native structure. In contrast to the simulation result for PRO54, the protein-like peptide with three types of amino acids, we found that a quasi-mirror image is seldom reached through folding runs. Energetic analysis suggested that quasi-mirror image has total energy about 7 kcal/mol higher than the right topology. This difference arises from the vdW interaction, the HB interaction, and the HP interaction.

§4. Conclusions

An automated optimization of potential parameters is proposed and is tested. We found that training of parameters with 39 protein database leads to better modeling not only for proteins in the database but also many other proteins. In particular all-α proteins have better score with the current energy function. Langevin dynamics simulation for a three helix bundle protein is performed finding that folding run from

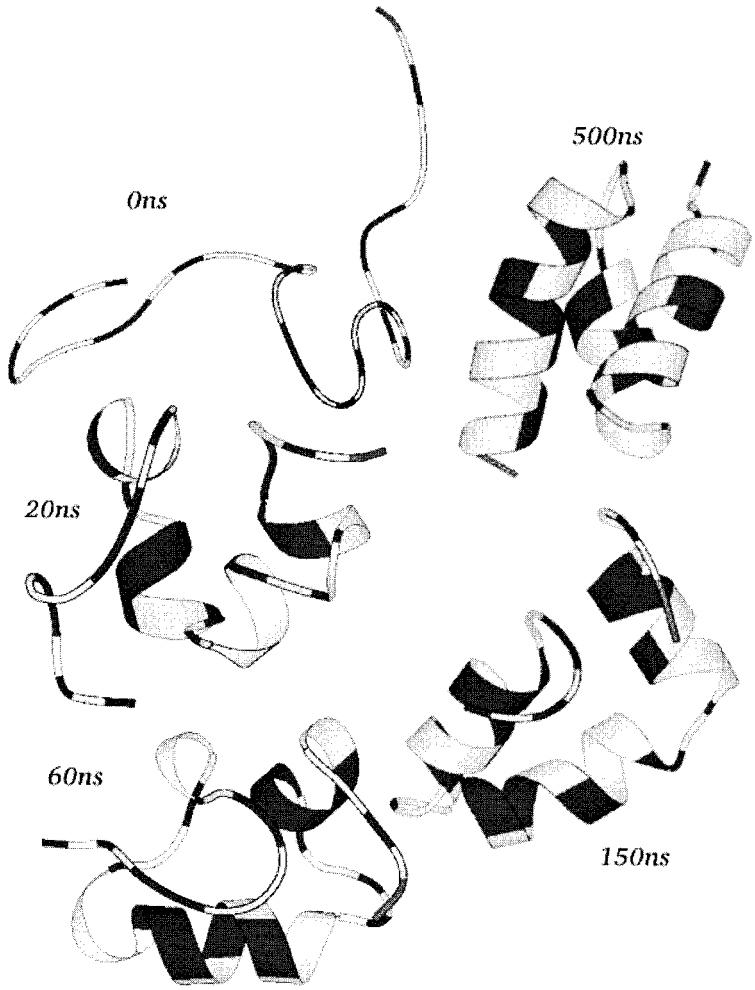

Fig. 2. Snapshots of a typical folding trajectory for albumin binding domain (drawn with Molscript [7]).

any random conformations always reaches at a native-like structure within about 3 A RMSD. A quasi-mirror image where helix alignment is opposite to the native structure is discriminated about 7 kcal/mol and almost all trajectories fall into the right conformation.

Acknowledgments

I would like to appreciate Peter G. Wolynes and Zaida Luthey-Schulten for useful discussions. This work has been supported by JSPS Research for the Future Program "Photo Science" and by the Grant-in-Aid on Priority Areas "Molecular Physical Chemistry".

References

1) A. R. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (WH Freeman and Co NY 1999).
2) J. D. Bryngelson, J. N. Onuchic, N. D. Socci and P. G. Wolynes, PROTEINS: Struct, Funct, Genetics. **21** (1995), 167.
3) J. J. Portman, S. Takada and P. G. Wolynes, Phys. Rev. Lett. **81** (1998), 5237.
4) S. Takada, Z. Luthey-Schulten and P. G. Wolynes, J. Chem. Phys. **110** (1999), 11616.
5) R. Goldstein, Z. Luthey-Schulten and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **89** (1992), 4918.
6) B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, J. Comp. Chem. **18** (1997), 1463.
7) P. J. Kraulis, Molscript, J. Appl. Crystallogr. **24** (1991), 946.