Inter-C^{α} Atomic Potentials Derived from the Statistics of Average Interresidue Distances in Proteins: Application to Bovine Pancreatic Trypsin Inhibitor

TAKESHI KIKUCHI

International Research Laboratories, Ciba-Geigy (Japan) Ltd., 10-66 Miyuki-cho, Takarazuka 665, Japan

Received 8 August 1994; accepted 9 May 1995

ABSTRACT

New effective potentials acting between pairs of residues in proteins are proposed based on statistics of average distances and standard deviations between C^{α} atoms of residues in protein tertiary structures. Gaussian functions are adopted as analytical forms of the potentials. A protein structure is modeled as a chain molecule with a fixed bond length connecting particles approximating the effects of amino acid residues. The potentials derived in this study are used for conformational sampling of trypsin inhibitor from bovine pancreas. Sampling is done with the Monte Carlo simulated annealing method. Sampled conformations can be classified into a few groups or structural classes, and one of these classes contains structures relatively close (with 7.8–8.7 Å root mean square [rms] deviation) to the X-ray structure. The native structure exhibits relatively low energy. These results denote a rather smooth landscape of the present potential energy surfaces. One class of classified structures contains nativelike structures, which suggests that the native structure can be predicted by further refinement of structures in this class. We discuss other properties and the effectiveness of the present potentials for description of protein structures.

Introduction

The prediction of tertiary structures of proteins from their sequences is one of the most significant unsolved problems in molecular biophysics. The methodologies of protein structure prediction proposed up to the present may be grouped roughly into two categories: *ab initio* predictions and knowledge-based predictions. The *ab initio* predictions attempt to search a protein structure by a minimization of total interatomic potential energy. This kind of method is based on the

CCC 0192-8651 / 96 / 020226-12

principle that the global minimum energy structure of a protein is the native structure.[1,2] The main problem of the *ab initio* predictions is finding the global minimum out of the huge number of minima on the potential energy hypersurface of a protein. *Ab initio* predictions usually require vast amounts of computer time for completion. Thus, their applications have been limited to small peptides with treatable size,[3-6] and no globular protein has been solved so far.

On the other hand, knowledge-based predictions are carried out using information about conformational propensities of amino acid residues in protein sequences, information which is derived from statistical analyses of protein structures determined by X-ray crystallographic or nuclear magnetic resonance (NMR) studies. The methodologies of knowledge-based predictions have been developed and applied originally to predict secondary structures in proteins.[7,8] Interactions formed only by four or five consecutive residues (i.e., short-range interactions) along a sequence are taken into consideration in standard secondary structure prediction methods. Hence these techniques contain an essential limitation in their predictability.[9]

A method to predict gross features of protein tertiary structures which are formed by long-range interactions, such as domain structures, has also been proposed.[10] This method is based on average values of interresidue distances computed from known protein structures.[10] Generally, such statistically estimated distance propensities can be transformed into effective potentials or potentials of mean force. For example, if we take a position of a C^{α} or a C^{β} atom as a representative or material point of a residue, pseudo interaction potentials can be derived from probability of occurrence of contacts between these material points. Such potentials are usually used on a lattice[11-13] or defined with inter-C^{β} atomic distances as a set of discrete values.[14-16] Covell[12b] demonstrated that it is possible to obtain protein structures relatively close to native structures (7.5–8.5 Å of rms values) by a lattice Monte Carlo method with the potentials defined by Miyazawa and Jernigan.[11] Skolnick et al.[13b-d,17] succeeded in obtaining with their refined lattice model folded protein structures which deviated from native structures by 2.5–4.5 Å for designed and natural helical bundles with short turns. Although these rms values obtained by Skolnick et al.[13b-d,17] are small, it is still uncertain whether their technique is applicable to various protein classes in general. Structures from a simulation with a lattice model might contain unrealistic packing densities.[18]

Knowledge-based potentials have also been applied to detect the native structure in the inverse folding problem,[19,20] and their usefulness has been demonstrated. (Readers are also referred to ref. 21 for the inverse folding problem or 3D–1D methods.) Furthermore, Rooman et al.[22] classified protein backbone structures on Ramachandran plots and derived potentials of mean force for these structures from the statistics of protein structures. With these potentials, they tried to predict tertiary structures of protein segments in which local interactions between residues are dominant and obtained structures close to native conformations.

As seen in these studies, positions of C^{α} or C^{β} atoms seem to be appropriate as representative points to describe protein conformations. This means that statistics of C^{α} or C^{β} atomic positions in proteins contain information of effective interatomic interactions in residues and backbones. However, the effective potentials in proteins have not yet been sufficiently studied. Such potentials should be able to reproduce the properties of both structure and dynamics of actual proteins. For example, the global minimum of an effective potential energy surface should give a structure close to the native conformation. Some of the sampled structures using the potentials proposed by Wilson and Doniach are very similar to the native structure,[15] but it is not clear whether such structures are actually near the minimum on the potential surface. Another significant property of real proteins is the approximate two-state folding–unfolding transition. To examine rigorously whether a potential function can describe this property, we have to carry out sufficiently large simulations to calculate thermodynamic properties of a protein.

In the present work, new knowledge-based potentials are derived from average distances between C^{α} atoms and their standard deviations in known protein structures. We try to express a potential in the form of a relatively simple analytical function. A potential surface expressed by an analytical function (not a discrete set of coordinates) is convenient for analysis of the gross features of the potential surface, and prediction of a protein structure by search for the global minimum is also more tractable. With the new potentials, we perform sampling of structures of a protein by means of the Monte Carlo simulated annealing technique and analyze the structures

obtained. This article focuses on the analysis of the gross features of our new potential energy surfaces. That is, we attempt to classify low-energy structures obtained by the new potentials and examine convergence of those structures into a few groups. (We will discuss the problem of dynamical properties of proteins on the new potentials elsewhere.) We apply the new potentials to the folding of BPTI and discuss the properties and effectiveness of the potentials.

Theoretical Background

The internal potential energy of a protein is expressed by a summation of interatomic interactions as follows:

$$E = \sum_{ij} u_{ij}(r_{ij}) = \sum_{\{k,l \text{ on } A\}} u_{kl}(r_{kl}) + \sum_{\substack{\{k \text{ on } A\} \\ \{\gamma \text{ on } B\} \\ (A \neq B)}} u_{k\gamma}(r_{k\gamma}) + \sum_{\substack{\{\alpha^A, \alpha^B\} \\ (=C^\alpha \text{ atoms}) \\ (A \neq B)}} u_{\alpha A \alpha B}(R_{AB}) \quad (1)$$

The term $u_{ij}(r_{ij})$ in the first line of eq. (1) means the interaction energy between atoms i and j separated by the distance r_{ij}. The total energy of a protein is formally decomposed into the three terms of eq. (1). The first term is the summation of interatomic interactions within each amino acid residue. Amino acids are labeled by A, B, and so on. The second term includes the interactions between two atoms (except C^α atoms) that belong to different amino acids. The inter-C^α atomic interactions are summed in the third term, where R_{AB} denotes the distance between C^α atoms in amino acid residues A and B. In eq. (1), α^A means the C^α atom in the residue A.

The partition function of this system can be formally written as

$$Z = \iiint \exp\left(-\beta\left(\sum_A u_{kl} + \sum_{AB} u_{k\gamma} + \sum_{AB} u_{\alpha A \alpha B}\right)\right) \times d\{r_{kl}\} d\{r_{k\gamma}\} d\{R_{AB}\} \quad (2)$$

Here, $d\{r_{ij}\} = dr_{12} dr_{13} \ldots$, and $\beta = 1/kT$.

If the motion of sidechains in a protein can be decoupled from that of the backbone chain (i.e., sidechains move fast enough compared with the backbone chain), we can perform the integrations related to $d\{r_{ij}\}$ and $d\{r_{k\gamma}\}$ separately from those related to $d\{R_{AB}\}$ in eq. (2). That is, the partition function, eq. (2), can be rewritten as Z_{app} as follows:

$$Z_{app} = \int F \cdot \exp\left(-\beta\left(\sum_{AB} u_{\alpha A \alpha B}(R_{AB})\right)\right) d\{R_{AB}\} \quad (3)$$

Here, $F = \iint \exp(-\beta(\sum_A u_{kl} + \sum_{AB} u_{k\gamma}))$ $d\{r_{kl}\} d\{r_{k\gamma}\}$.

Considering a protein as a polymer molecule, the typical relaxation time of a backbone structural change of a protein, τ, can be approximated as[23]

$$\tau \backsim N^\delta \quad (4)$$

Here, N is the number of monomers (i.e., residues, in the protein). (In addition, $\delta = 2$ in the Rouse model[24] and $\delta = 9/5$ in the Kirkwood approximation.[25]) On the other hand, the typical relaxation time of sidechain motion is on the order of 10^{-9}–10^{-6} s. (For example, the typical relaxation time of rotation of a phenyl ring in Tyr is on the order of at most 10^{-6} s.) Therefore, when N is large enough, sidechain dynamics can be expressed as motion with a frozen backbone chain conformation. Conversely, structural change of the backbone is regarded as motion in the mean field produced by the sidechains.

Introducing a probability distribution $\rho(R_{AB})$, an average distance between C^α atoms of residues A and B is expressed as

$$\langle R_{AB} \rangle = \int R_{AB} \rho(R_{AB}) \, dR_{AB} \quad (5)$$

where

$$\rho(R_{AB}) = \frac{\exp(-\beta u_{\alpha A \alpha B}(R_{AB})) \times \int F \exp\left(-\beta\left(\sum_{\substack{CD \\ (\neq AB)}} u_{\alpha C \alpha D}(R_{CD})\right)\right) d\{R_{CD}\}}{Z_{app}} \quad (6)$$

Therefore,

$$\ln(\rho(R_{AB})) = -\beta \varepsilon_{AB}(R_{AB}) - \ln Z_{app} \quad (7)$$

where

$$\varepsilon_{AB}(R_{AB}) = u_{\alpha A_\alpha B}(R_{AB}) - (1/\beta)\ln J$$

and

$$J = \int F \exp\left(-\beta\left(\sum u_{\alpha C_\alpha D}(R_{CD})\right)\right) d\{R_{CD}\}$$

Thus, an interresidue potential $\varepsilon_{AB}(R_{AB})$ is obtained formally from the probability distribution $\rho(R_{AB})$. [As long as we discuss the difference of total energy of conformers, the term $\ln Z_{app}$ in eq. (7) does not appear explicitly in the calculations.] It is expected that ε_{AB} depends strongly on types of residues A and B. Therefore, instead of an explicit form of $\rho(R_{AB})$, we use a statistical distribution, ρ_{ab}, obtained from X-ray structures of proteins as an empirical model in this article. That is, eq. (7) is replaced by

$$(1/\beta)\ln \rho_{ab}(R_{ab}) = -\varepsilon_{ab}(R_{ab}) - (1/\beta)\ln Z_{app} \quad (8)$$

In eq. (8), a and b denote the residue types. From this equation, the effective potential, ε_{ab}, is calculated from the distribution function ρ_{ab} derived from X-ray structures.

In this study, a continuous analytic function is introduced for ρ_{ab}. The problem is how the analytic form of ρ_{ab} is determined. There is no unique answer yet. The simplest and most reasonable form is a Gaussian function whose average value and standard deviation coincide with the values computed from the statistics of X-ray structures. This simplest case is investigated in the present work. Of course, this function corresponds to the approximation of ε_{ab} in eq. (8) by a harmonic potential. Furthermore, we also examine the superposition of two Gaussian functions as an approximation of ρ_{ab} based on the fact that an arbitrary function might be approximated by superposition of several Gaussian functions.* Gaussian potentials have also been used as constraints between residues in the homology modeling of proteins by Sali and Blundell.[26]

To transform ρ_{ab} to ε_{ab} according to eq. (8), we need the value of β as a parameter. In this study, we use 1.667 for this value, which corresponds to 0.6 kcal/mol of kT. As Sippl[14a] pointed out, the choice of temperature here is not critical. β is regarded as a multiplicative factor.

* For example, in electronic structure theory of molecules, a Slater-type orbital can be approximated by a superposition of several Gaussian functions. See, for example, D. Feller and E. R. Davidson, *Review in Computational Chemistry*, Vol. 1, K. B. Lipkowitz and D. B. Boyd, Eds., VCH Publishers, New York, 1990, pp. 1–43.

Statistics and Distribution Functions

The published data of average distances between C^{α} atoms in proteins[10a] are used. In ref. 10, the average distances and their standard deviations have been computed with 42 nonhomologous proteins. A range M is defined by the formula $|i - j| = k$, the separation of two residues i and j on a sequence. According to ref. 10a, $M = 1$ when $1 \le k \le 8$, $M = 2$ when $9 \le k \le 20$, $M = 3$ when $21 \le k \le 30$, and so on. For each range, an average distance $\langle R_{ab}^M \rangle$ and its standard deviation σ_{ab}^M have been calculated, where a and b denote amino acid types and M a range. Therefore, the distribution function depends on a, b, and M—namely, $\rho_{ab}^M(R_{ab})$. R_{ab}^M is a distance between a residue pair a and b in a range M.

Let us assume that $\rho_{ab}^M(R_{ab})$ is a Gaussian function with the average value, $\langle R_{ab}^M \rangle$, and the standard deviation, σ_{ab}^M, as follows:

$$\rho_{ab}^M(R_{ab}) = \left(1/\sqrt{2\pi}\,\sigma_{ab}^M\right) \times \exp\left[-(R_{ab}^M - \langle R_{ab}^M \rangle)^2/2(\sigma_{ab}^M)^2\right] \quad (9)$$

We refer to the distribution function expressed by eq. (9) as a single Gaussian distribution function.

Next we examine a distribution function in the following form:

$$\rho_{ab}^M(R_{ab}) = \sum_{i=1}^{l} \left(C_i/\sqrt{2\pi}\,\sigma_{iab}^M\right) \times \exp\left(-(R_{ab}^M - \Delta_{iab}^M)^2/2(\sigma_{iab}^M)^2\right) \quad (10)$$

In eq. (10), C_i, σ_{iab}^M, and Δ_{iab}^M are arbitrary parameters to be determined to give observed values of $\langle R_{ab}^M \rangle$, and σ_{ab}^M. Here, $0 \le C_i \le 1.0$. We take the case of $l = 2$ in this article. Then

$$\langle R_{ab}^M \rangle = C_1\Delta_{1ab}^M + C_2\Delta_{2ab}^M \quad (11)$$

$$(\sigma_{ab}^M)^2 = C_1\left[(\sigma_{1ab}^M)^2 + (\Delta_{1ab}^M)^2\right] + C_2\left[(\sigma_{2ab}^M)^2 + (\Delta_{2ab}^M)^2\right] - \langle R_{ab}^M \rangle^2 \quad (12)$$

Here, $C_2 = 1 - C_1$.

We define Δ^M_{1ab}, Δ^M_{2ab}, σ^M_{1ab}, and Δ^M_{2ab} as follows:

$$\Delta^M_{1ab} = \langle R^M_{ab} \rangle - C_2\delta,\ \Delta^M_{2ab} = \langle R^M_{ab} \rangle + C_1\delta,$$
$$\delta = \alpha\sigma^M_{ab}/\sqrt{C_1C_2}$$

where α is a positive adjustable parameter. Let us take the σ^M_{iab} as follows:

$$C_1(\sigma^M_{1ab})^2 = D_1\left[(\sigma^M_{ab})^2 + \langle R^M_{ab}\rangle^2 - C_1(\Delta^M_{1ab})^2 - C_2(\Delta^M_{2ab})\right]$$
$$C_2(\sigma^M_{2ab})^2 = D_2\left[(\sigma^M_{ab})^2 + \langle R^M_{ab}\rangle^2 - C_1(\Delta^M_{1ab})^2 - C_2(\Delta^M_{2ab})\right] \quad (13)$$

Here, $0 < D_1 < 1.0$ and $D_2 = 1 - D_1$. Thus, the parameters to be adjusted are C_1, α, and D_1. We always take the values of these parameters such that $\Delta^M_{1ab} \le \Delta^M_{2ab}$ and $\sigma^M_{1ab} \le \sigma^M_{2ab}$. We call the potential function defined by eq. (10) a double Gaussian potential.

A Model of a Protein

The following model of a protein is adopted in this article. We consider a chain molecule consisting of N particles. Each pair of consecutive particles is connected by a virtual bond with the length of 3.8 Å. Each of these particles represents all effects of an amino acid in the protein. Any values between 0 and π radians can be taken as bond angles and between 0 and 2π as dihedral angles. Potentials defined by eq. (9) or (10) act among these particles, but interaction energy is adjusted to become a certain large value, E_{rep}, in the region of the interparticle distance of $R \le 3.8$ Å. An artificial potential acts for each residue of disulfide pairs in the following form:

$$E_{ss} = k_{ss}(r - r_{ss})^2 \quad (14)$$

where r is the distance between two particles to form a disulfide bond. We set $E_{rep} = 50$ kcal/mol, $k_{ss} = 50$ kcal/mol Å^2, and $r_{ss} = 5.5$ Å in this article.

Sampling Procedure

To investigate characteristics of the potentials, sampling of conformations of a protein was carried out by means of the Monte Carlo simulated annealing method. The Metropolis algorithm[27] was employed in this study. Starting with a structure with random bond angles and dihedral angles, each bond angle and dihedral angle of the structure was changed randomly in the interval of $0-\pi$ and $0-2\pi$, respectively, during the course of the simulation. The Monte Carlo procedure included changes of bond angles and dihedral angles of all residues. (In the present work, we did not take avoidance of steric crashes into account in the Monte Carlo movements.) This procedure was iterated 5000 times while decreasing the relative temperature, kT, from 100 to 0.1. This temperature range was chosen empirically. In the relative temperature range 100–5, an acceptance ratio of 60–40% could be obtained. We made a number of runs of this procedure, and 50 structures with energy values within 500 kcal/mol from the lowest were selected as sampled structures. Then we analyzed the final results. As a test, the present potentials were applied to trypsin inhibitor from bovine pancreas (BPTI).

Results

THE CASE OF A SINGLE GAUSSIAN DISTRIBUTION FUNCTION

In Table I, we show the average values of energy, rms deviation from the X-ray structure, and radius of gyration of 50 structures of BPTI obtained by the Monte Carlo simulation described in the previous section. The energy of the native structure was approximated in this work as the value of the slightly relaxed conformation produced from the X-ray structure by 10 Monte Carlo

TABLE I.
Properties of the 50 Sampled Structures by the Single Gaussian Function and the Relaxed Native Structure.

	Energy (kcal / mol)	Radius of Gyration (Å)	Rms Deviation (Å) from the X-ray Structure
Sampled structures (average)	5714.8	12.2	10.73
Relaxed native structure	3094.4	10.7	0.75

iterations at $T = 10$. We observed no important change in tertiary structure after this relaxation. The rms deviation of the relaxed structure from the X-ray structure was 0.75 Å. Radii of gyration of the X-ray and the relaxed structures are 10.6 and 10.7 Å, respectively. This relaxed structure is referred to as the relaxed native structure. Figures 1(a) and 1(b) illustrate the X-ray and relaxed native structures of BPTI. The values of the energy, the radius of gyration, and the rms deviation of the relaxed native structure of BPTI are also presented in Table I. We notice from this table that the energy of the relaxed structure is remarkably low compared with the average energy of the 50 structures. The average value of the radius of gyration of 50 structures is 12.2 Å, which is 15% larger than that of the X-ray structure. As expected, the native structure is rather compact.

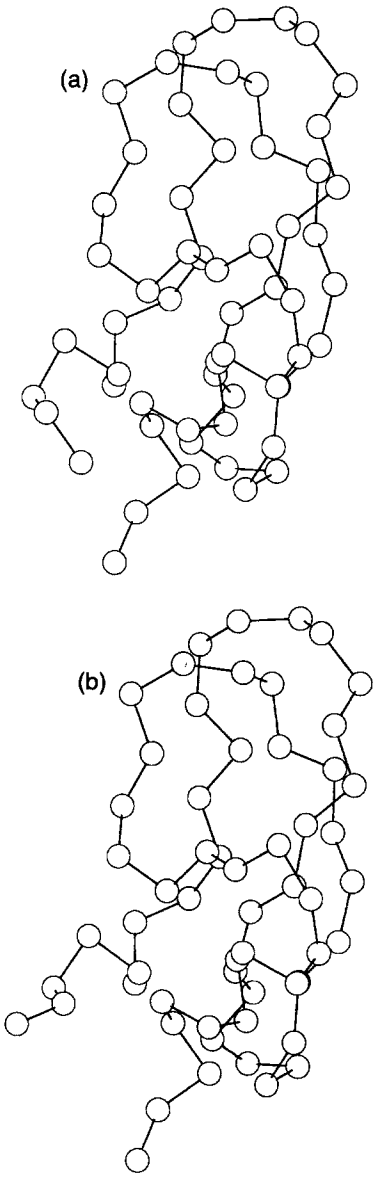

FIGURE 1. (a) C^{α} trace of the X-ray structure of BPTI. (b) C^{α} trace of the relaxed native structure of BPTI.

Kawai et al.[4a] proposed a technique to classify conformations of peptides and proteins generated by the Monte Carlo simulated annealing method. According to this technique, two conformations are classified to be in the same set if this pair of conformations deviates from each other by less than a cutoff rms distance. If we take too small a cutoff value, only a few of the generated conformations are put in the same set, and there are many conformations that do not belong to any set. On the other hand, if we take too large a cutoff, many conformations belong to more than one set simultaneously. Therefore, we have to determine the cutoff rms distance carefully so that generated conformations, as much as possible, belong to one and only one of classes (i.e., no conformation belongs to two classes concurrently[4a]). We applied this method to our problem.

As a result, among the 50 sampled conformations, we could choose six structures in which any two structures deviate each other if we take 8.0 Å as the cutoff rms value. We refer to these six structures as core structures. Core structures in a class are selected so that as many other sampled structures fall into a class as possible. We could find 13 other structures similar to at least one of these six structures within the cutoff rms deviation of 8.0 Å. Besides these 19 structures, we can choose six other core structures each resembling the others by less than the same cutoff rms distance; 15 more structures can be picked up to be similar to at least one of those six structures within the same cutoff value. Thus, a majority ($19 + 21 = 40$) of the 50 structures can be classified roughly into two classes. We call the former class I and the latter class II. There is no common structure between classes I and II. Therefore, the rms cutoff distance of 8.0 Å is appropriate for the present classification.

Table IIa represents the rms distances between the core structures in each class. We show the rms deviation between conformations from different classes in Table IIb. Each conformation is labeled by a number I-1, I-2 and so on, where, for example, I-1 means the conformation 1 in class I. With this table, we can confirm that the core conformations in each class resemble one another within the cutoff distance of 8.0 Å. The corresponding aver-

TABLE IIa.

Rms Distances (Å) between Core Conformers in Classes Sampled by the Single Gaussian Function.

Class I

	I-1	I-2	I-3	I-4	I-5	I-6
I-1	—	7.30	7.20	6.08	6.39	7.08
I-2		—	5.43	7.28	5.50	6.96
I-3			—	7.83	6.23	6.98
I-4				—	6.59	7.85
I-5					—	7.35
I-6						—

Average: 6.80

Class II

	II-1	II-2	II-3	II-4	II-5	II-6
II-2	—	6.80	7.33	6.01	7.13	7.75
II-2		—	7.74	7.20	7.94	7.88
II-3			—	6.59	7.74	7.53
II-4				—	6.84	6.99
II-5					—	5.92
II-6						—

Average: 7.16

Each core conformer is labeled by Roman and Arabic numerals (e.g., I-1 denotes the conformer 1 in class I).

TABLE IIb.

Rms Distances (Å) between Core Conformers from Classes I and II Sampled by the Single Gaussian Function.

	I-1	I-2	I-3	I-4	I-5	I-6
II-1	13.50	13.52	13.58	12.61	13.01	13.32
II-2	13.82	14.91	13.28	13.31	13.92	13.03
II-3	14.02	13.97	12.68	13.17	14.15	13.85
II-4	14.17	13.82	13.02	13.38	13.78	13.53
II-5	12.65	11.78	12.88	12.51	11.91	12.48
II-6	12.37	12.97	13.75	12.40	12.74	12.92

Average: 13.21

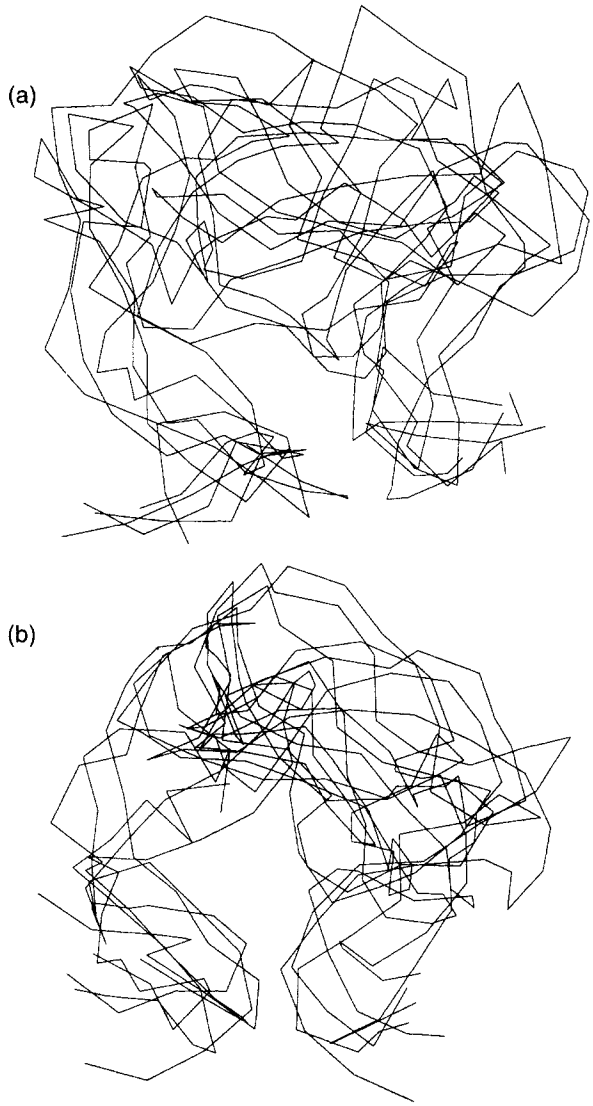

FIGURE 2. (a) Superposition of C^{α} traces of the core structures in class I sampled by the single Gaussian function. (b) Superposition of C^{α} traces of the core structures in class II sampled by the single Gaussian function.

age rms values are 6.80 Å and 7.16 Å, respectively. On the other hand, there is less similarity between any of two structures from different classes, in which the average rms value is 13.21 Å.

Figures 2(a) and 2(b) indicate the superposition of the six core structures of class I and the six core structures of class II, respectively. The properties of these core structures are summarized in Table III. There is no remarkable difference between classes I and II in the average values of energy and radii of gyration. However, the rms values from the X-ray structure denote the obvious difference between classes. The average rms value of the core structures in class I from the X-ray structure is clearly smaller than that of class II (9.06 and 11.68 Å, respectively). The values of the rms deviation of three of the core members in class I from the X-ray structure fall in the range of 8.2–8.7 Å (Table III). One more conformation in class I (not a core structure) also shows 8.61 Å of rms deviation. These structures are closest to the X-ray conformation. These facts attest to the relative resemblance of conformations in class I to the X-ray structure.

TABLE III.
Properties of the Core Structures in Classes I and II Derived by the Single Gaussian Function.

	Energy (kcal / mol)	Radius of Gyration (Å)	Rms Deviation from the X-ray Structure (Å)
Class I			
I-1	5820.4	12.1	8.17
I-2	5711.0	12.5	9.62
I-3	5807.5	12.0	9.59
I-4	5766.8	12.1	8.44
I-5	5637.5	11.9	9.88
I-6	5801.0	11.9	8.63
Average	5757.4	12.1	9.06
Class II			
II-1	5882.5	12.0	12.08
II-2	5868.6	12.2	12.57
II-3	5601.9	13.0	11.72
II-4	5463.5	12.7	11.82
II-5	5605.4	12.4	10.89
II-6	5652.7	12.0	11.0
Average	5679.1	12.4	11.68

Thus, it is demonstrated that the sampled structures with the present potential function are mainly classified into two classes. That is, the sampled conformations converge into two conformational families based on rms deviations. In particular, from the structural similarity of the members in class I to the native structure, we expect that some of the structures in class I can be energetically relaxed to conformations close to the relaxed native structure taking the relatively low energy of the relaxed native structure into account. However, both the average values of energy and radius of gyration of the sampled structures in class I are larger than the native values. This difficulty of the proximity to the native values by the Monte Carlo simulated annealing method suggests that the present potential surface still contains high complexity.

THE CASE OF A DOUBLE GAUSSIAN DISTRIBUTION FUNCTION

We chose the values of the parameters in eqs. (12) and (13) as $C_1 = 0.7$, $D_1 = 0.3$, and $\alpha = 0.8$. These values have been adjusted to obtain values of rms deviations from X-ray structure and radii of gyration as small as possible. The potential function attains asymmetric shape compared with the symmetric form of a single Gaussian case. As an example, Figure 3 shows the profile of the double Gaussian potential defined by eqs. (8) and (10) for the amino acid pair Ala–Ala in the range 3 ($21 \leq k \leq 30$). Twenty-five structures were sampled with the double Gaussian function using the same procedure as with the single Gaussian case. Table IV summarizes the average values of energy, radii of gyration, and rms deviation from the X-ray structure. As with the single Gaussian case, this

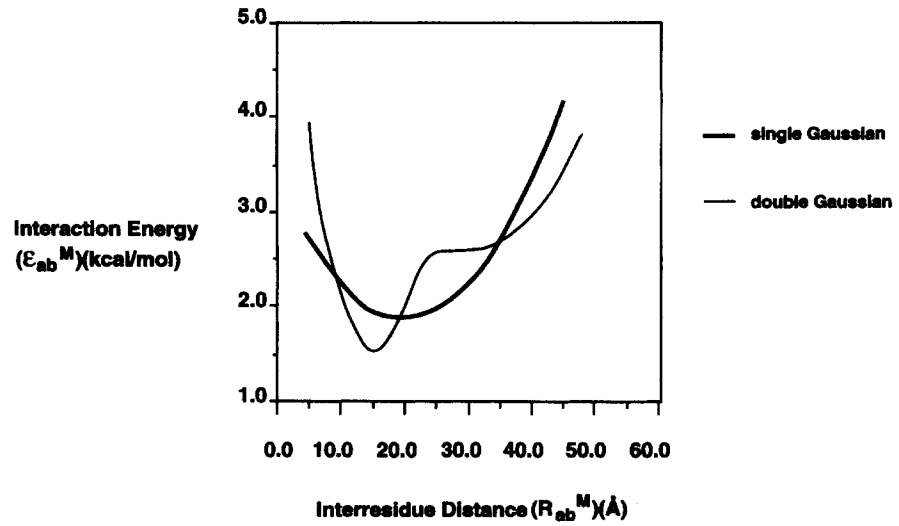

FIGURE 3. The profiles of the single and double Gaussian potentials defined by eq. (8) for the amino acid pair Ala–Ala in the range 3 ($M = 3$, i.e., $21 \leq k \leq 30$). The average distance and standard deviation are 19.71 and 9.08, respectively.

TABLE IV.
Properties of the 25 Sampled Structures by the Double Gaussian Function and the Relaxed Native Structure.

	Energy (kcal / mol)	Radius of Gyration (Å)	Rms Deviation (Å) from the X-ray Structure
Sampled structures (average)	5786.5	11.7	10.65
Relaxed native structure	3226.4	10.6	0.15

double Gaussian function leads to a low energy of the relaxed native structures compared to the average energy of the sampled structures.

Again, we tried to classify the 25 structures into a few classes using the 8.0-Å cutoff distance. As a result, the sampled structures could be classified into two classes. When we take two conformers as core structures of class I, the class contains seven conformations as a whole. These two structures are also similar to the X-ray structure by 7.8–8.3 Å rms deviation (Table VI). This class contains another conformation with 8.34-Å rms deviation from the X-ray structure. Those three structures show the smallest rms distances from the native structure among the 25 generated structures in the present simulations. As class II, four structures are recognized as core members. We can find five other structures that resemble at least one of four core structures within the cutoff rms distance.

The values of the rms deviation between core conformations within each class are shown in Table Va. The conformational differences among the core structures belonging to different classes are presented in Table Vb. Again, there is no common structure between conformations in class I and those in Class II. The conformational differences between classes I and II are clear (i.e., the average rms deviation between them is 10.38 Å).

TABLE Va.
Rms Distances (Å) between Core Conformers in Classes Sampled by the Double Gaussian Function.

Class I II-1 vs II-2: Class II	7.99			
	I-1	I-2	I-3	I-4
I-1	—	7.27	7.09	7.53
I-2		—	7.06	5.78
I-3			—	7.28
I-4				—
Average: 7.00				

TABLE Vb.
Rms Distances (Å) between Core Conformers from Classes I and II Sampled by the Doubled Gaussian Function.

	I-1	I-2	I-3	I-4
II-1	10.69	9.48	10.08	9.30
II-2	11.90	11.43	9.74	10.36
Average: 10.38				

The average values of the properties of the core structures are summarized in Table VI. The average rms deviation of the core structures in class I from the X-ray structure is 8.06 Å, and this is clearly smaller than that for class II. Thus, the proximity of the sampled structures to the native-like structures in class I is somewhat improved in the double Gaussian function compared to the 9.06 Å in Table III. The average value of radii of gyration, 11.7, is slightly smaller than the corresponding value (12.2) of the single Gaussian potential, as shown in Tables I and IV. Moreover, as seen in Table VI, the average radius of gyration of class I (11.05) is clearly smaller than that of class II (11.7). The 11.05 value is close to the native value, 10.6 (Table VI).

It is interesting to compare the structures obtained with the single Gaussian function to those

TABLE VI.
Properties of the Core Structures in Classes I and II Derived by the Double Gaussian Function.

	Energy (kcal / mol)	Radius of Gyration (Å)	Rms Deviation from the X-ray Structure (Å)
Class I			
I-1	5910.8	11.0	7.88
I-2	5830.5	11.1	8.23
Average	5865.7	11.05	8.06
Class II			
II-1	5977.3	12.0	11.15
II-2	5787.0	11.9	11.04
II-3	5852.2	11.0	10.92
II-4	5751.9	11.8	10.59
Average	5842.1	11.7	10.93

TABLE VII.

Average Rms Distances (Å) between Structures in Classes Defined by the Single and Double Gaussian Functions, Respectively.

	SGI	SGII
DGI	7.79	12.05
DGII	11.24	8.59

Classes are labeled by SGI, DGI, and so on. For example, SGI and DGI denote the classes I defined by the single and double Gaussian functions, respectively.

derived from the double Gaussian potential surface. We refer to the classes I and II derived by the single Gaussian function as SGI and SGII, respectively, and the classes I and II defined by the double Gaussian function as DGI and DGII, respectively. Table VII shows average rms distances between structures. The rms distances between SGI and DGI and between SGII and DGII are 7.99 Å and 8.59 Å, respectively, suggesting that SGI and SGII correspond to DGI and DGII, respectively. These results imply that sampled conformations on both single and double Gaussian potential functions converge into the same two conformational classes.

Discussion

As described in the previous section, the sampled structures of BPTI by the Monte Carlo simulated annealing method can be classified into a few classes with both the single and double Gaussian functions. This implies that the gross shape of the landscape of each potential function can be expressed by only a few wells. That is, each of the potential functions can be approximated by a smooth surface, especially at high temperatures. The degree of approximation is reflected by the criterion of the 8.0-Å rms distance used for the classification of the sampled structures. This value is expected to be improved by optimization of the cooling schedule in the Monte Carlo simulated annealing procedure. As seen in Table VII, the structures of classes I and II in the case of the single Gaussian function appear to correspond to classes I and II on the double Gaussian potential surface. As expected, the corresponding potential wells on both the surfaces appear to be fairly close. It is interesting that one of these classes of each potential function contains conformations close (with 7.8 Å–8.7 Å rms) to the X-ray structure, and the relaxed native structure has extremely low energy compared with the sampled structures. Thus, it is expected that, if conformations of this class are energetically refined further, the native-like structures will be attainable.

These properties of the present potential energy surface seem to be related to the convergence of a protein to its native structure. However, the convergence of the sampled conformations into a few classes is somewhat crude, especially with the single Gaussian potential function. For example, the average value of the rms deviation of the structures in class I from the X-ray structure is 9.06 Å (Table III). This value is relatively large, although it is smaller than the corresponding value of class II. This rms indicates that the shape of the potential surface is still rather rugged. The situation has been improved to some degree in the double Gaussian function. We obtained a smaller average rms deviation of the core structures as 8.06 Å in the class I sampled by the double Gaussian function (Table VI). This value is close to the result of a lattice Monte Carlo simulation carried out by Covell,[12b] Skolnick and Kolinski,[13a] and Godzik et al.[20] Of course, the value in the present analysis cannot be compared directly with those values because their values have been obtained from the finally optimized conformations, whereas our value is taken from the average of the values of the core structures. The value in this study will be refined further.

As pointed out, the single Gaussian potential function corresponds to the harmonic approximation of the potentials. On the other hand, the present double Gaussian version includes an asymmetric form of the function, as shown in Figure 3. We think that this mimics correction of the anharmonicity of the real potentials. Therefore, nonlinearity of the potential function seems to smooth the landscape and makes the convergence of the sampled structures better. This effect is also reflected in the smaller values of the radii of gyration of the sampled structures on the double Gaussian potential surface. In the single Gaussian case, the average value of radius of gyration of the structures in class I is 12.2 Å, which is rather large compared with the native value, 10.6 Å. This difference suggests that a chain has difficulty being compact and is trapped in an expanded conformation because of the fairly rugged single Gaussian potential surface. With the double Gaussian function, the average radius of gyration value is smaller (11.7 vs. 12.2), suggesting that a conformation on

this potential energy surface tends to be more compact than on the single Gaussian surface because of the relative smoothness of the double Gaussian surface. It is interesting that the average radius of gyration of class I is closer to the native value in the double Gaussian simulations.

In the present work, one of our main objectives was to understand the gross features of the landscape produced by the potential functions. Therefore, we stopped each Monte Carlo simulation at 5000 iterations. It is expected that further simulations for a structure in each class would lead to a more refined structure. However, we checked that a simulation with 20,000 iterations gives essentially the same results as 5000 simulations. Therefore, to obtain a more refined structure in a class, we should perform much more than 20,000 iterations, perhaps on the order of 10^5. We are currently working to make larger simulations for the structures in these classes.

In an application of the present method to protein structure prediction, the resolution of the classification of sampled conformations should be improved to at least around 4.0 Å.[12b,20,28] To attain this level of the resolution, we would need to (1) increase the number of the simulation steps, (2) take more detailed statistics of average distances, or (3) improve the distribution function by increasing the number of Gaussian functions superposed. For example, the present treatment does not show clear secondary structure formation in the structures [Figs. 2(a) and 2(b)]. This is because we used the statistics averaged over the range $1 \leq k \leq 8$. Therefore, the present potentials are insensitive to the formation of secondary structures. This property will be improved by refinement of the statistics of the range $1 \leq k \leq 8$.

We found that the native structure corresponds to a very low energy (i.e., the secondary structures decrease the energy of the protein even with the present potentials). On the other hand, Saitoh et al.[29] have pointed out that the consideration of secondary structures in building of a protein tertiary structure does not remarkably improve rms deviation of a built-up structure from the native structure.

In the present study, we take C^{α} atoms as the representative particles describing a protein's tertiary structure. Further improvement might be achieved when we treat more detailed statistics (e.g., considering average distances between C^{β} atoms and between centers of mass of residues simultaneously). These improvements can be carried out by use of multivariable Gaussian functions. It would also be interesting to increase the number of Gaussian functions superposed. However, the number of parameters also considerably increases. We should seek criteria to determine the optimum values of the parameters. Then improvement of the situation might take place more than we saw in the passage from the single to double Gaussian function. Even at the present stage, our method will help reduce conformational space of a protein by the classification of generated structures into a few classes of the backbone. After that, further refinement of a whole protein structure can be performed by detailed Monte Carlo or molecular dynamics calculation within the restricted conformational space.

As shown in this article, knowledge-based potentials of mean force are efficient in modeling protein structures to some extent. This denotes that motion of a backbone of a protein can be decoupled from that of sidechains in eq. (3) to a certain degree. During folding of a protein with a high degree of freedom of a backbone, this approximation is fairly valid. Deviation of motion of a backbone and sidechains from this approximation may become large when the degree of freedom is almost lost at the final stage of protein folding. This stage might include transition from a molten globule to the native state.[30] However, the fact that knowledge-based potentials of mean force are generally effective to describe a protein tertiary structure suggests the validity of this approximation in a major part of the process of protein folding.

Of course, as a final step of a tertiary structure prediction with knowledge-based potentials, refinement of the structures should include the effects of sidechain conformations. Deviation of a final structure predicted by such potentials from the native structure denotes the roughness of this approximation. On the other hand, small peptide, for which we cannot decouple the motion of the main chain from that of sidechains, does not form a unique structure but exists in a conformational ensemble. We can empirically infer that the minimum number of residues required to fold a protein into its unique structure is 40–50.

In the present study, a continuous coordinate system is adopted to express potentials of a protein. Therefore, the present model is free from possible fictitious effects of lattice models of proteins. The characteristic of our method is an approximation of the potential function of proteins by an analytical function. Hence the calculation is considerably simplified compared to lattice mod-

els and to models using discrete potential sets. It is relatively easy to apply the present technique to a dynamical problem of a protein. We are currently working on analysis of dynamics of a protein on the present potential energy surfaces. We will try to improve our potentials by increasing of the number of Gaussian functions superposed. Then we may expect that the accuracy of the present method will be improved.

References

1. (a) C. B. Anfinsen and H. A. Scheraga, *Adv. Prot. Chem.*, **29**, 205 (1975); (b) G. Némethy and H. A. Scheraga, *Quart Rev. Biophys.*, **10**, 239 (1977).
2. H. A. Scheraga, *Review in Computational Chemistry*, Vol. 3, K. B. Lipkowitz and D. B. Boyd, Eds., VCH Publishers, New York, 1992, p. 73.
3. (a) M. Vásquez and H. A. Scheraga, *J. Biomol. Struc. Dyn.*, **5**, 757 (1988); (b) L. Piela and H. A. Scheraga, *Biopolymers*, **26**, S33 (1987); (c) D. R. Ripoll, L. Piela, M. Vásquez, and H. A. Scheraga, *Proteins*, **10**, 188 (1991); (d) Z. Li and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.*, **84**, 6611 (1987); (e) L. Piela, J. Kostrowicki, and H. A. Scheraga, *J. Phys. Chem.*, **93**, 3339 (1989).
4. (a) H. Kawai, T. Kikuchi, and Y. Okamoto, *Protein Eng.*, **3**, 85 (1989); (b) Y. Okamoto, T. Kikuchi, and H. Kawai, *Chem. Lett.*, **1992**, 1275; (c) H. Kawai, Y. Okamoto, M. Fukugita, T. Nakazawa, and T. Kikuchi, *Chem. Lett.*, **1991**, 213.
5. B. von Freyberg and W. Braun, *J. Comp. Chem.*, **12**, 1065 (1991).
6. S. R. Wilson and W. Cui, *Tetrahedron Lett.*, **29**, 4373 (1988).
7. P. Y. Chou and G. D. Fasman, *Adv. Enzymol.*, **47**, 45 (1978).
8. J. Garnier, D. J. Osguthorpe, and B. Robson, *J. Mol. Biol.*, **120**, 97 (1978).
9. K. Nishikawa, *Biochem. Biophys. Acta*, **748**, 285 (1983).
10. (a) T. Kikuchi, G. Némethy, and H. A. Scheraga, *J. Protein Chem.*, **7**, 427 (1988); (b) T. Kikuchi, G. Némethy, and H. A. Scheraga, *J. Protein Chem.*, **7**, 473 (1988); (c) T. Kikuchi, G. Némethy, and H. A. Scheraga, *J. Protein Chem.*, **7**, 491 (1988).
11. S. Miyazawa and R. L. Jernigan, *Macromolecules*, **18**, 534 (1985).
12. (a) D. G. Covell and R. L. Jernigan, *Biochemistry*, **29**, 3287 (1990); (b) D. G. Covell, *Proteins*, **14**, 409 (1992); (c) D. G. Covell, *J. Mol. Biol.*, **235**, 1032 (1994).
13. (a) J. Skolnick and A. Kolinski, *Science*, **250**, 1121 (1990); (b) A. Kolinski, A. Godzik, and J. Skolnick, *J. Chem. Phys.*, **98**, 7420 (1993); (c) A. Kolinski and J. Skolnick, *Proteins*, **18**, 338 (1994); (d) A. Kolinski and J. Skolnick, *Proteins*, **18**, 353 (1994).
14. (a) M. J. Sippl, *J. Mol. Biol.*, **213**, 859 (1990); (b) G. Casari and M. J. Sippl, *J. Mol. Biol.*, **224**, 725 (1992).
15. C. Wilson and S. Doniach, *Proteins*, **6**, 193 (1989).
16. S. H. Bryant and C. E. Lawrence, *Proteins*, **16**, 92 (1993).
17. J. Skolnick, A. Kolinski, C. L. Brooks III, A. Godzik, and A. Rey, *Curr. Biol.*, **3**, 414 (1993).
18. L. M. Gregoret and F. E. Cohen, *J. Mol. Biol.*, **219**, 109 (1991).
19. D. T. Jones, W. R. Taylor, and J. M. Thornton, *Nature*, **358**, 86 (1992).
20. A. Godzik, A. Kolinski, and J. Skolnick, *J. Mol. Biol.*, **227**, 227 (1992).
21. R. Lüthy, J. U. Bowie, and D. Eisenberg, *Nature*, **356**, 83 (1992).
22. (a) M. J. Rooman, J.-P. A. Kocher, and S. J. Wodak, *J. Mol. Biol.*, **221**, 961 (1991); (b) M. J. Rooman, J.-P A. Kocher and S. J. Wodak, *Biochemistry*, **31**, 10226 (1992); (c) M. J. Rooman and S. J. Wodak, *Biochemistry*, **31**, 10239 (1992).
23. P.-G. de Gennes, *Scaling Concepts in Polymer Physics*, Cornell University Press, Ithaca and London, 1979, p. 180.
24. P. E. Rouse, *J. Chem. Phys.*, **21**, 1273 (1953).
25. J. Kirkwood and J. Riseman, *J. Chem. Phys.*, **16**, 565 (1948).
26. A. Sali and T. L. Blundell, *J. Mol. Biol.*, **234**, 779 (1993).
27. N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, *J. Chem. Phys.*, **21**, 1087 (1953).
28. P. Correa, *Proteins*, **7**, 366 (1990).
29. S. Saitoh, T. Nakai, and K. Nishikawa, *Proteins*, **15**, 191 (1993).
30. K. Kuwajima, *Proteins*, **6**, 87 (1989).