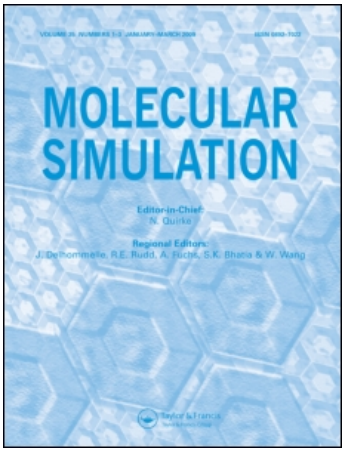

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

First-Principles Protein Folding Simulations

Yuko Okamoto[a]
[a] Department of Theoretical Studies, Institute for Molecular Science and Department of Functional Molecular Science, The Graduate University for Advanced Studies, Okazaki, Aichi, Japan

To cite this Article Okamoto, Yuko(2000) 'First-Principles Protein Folding Simulations', Molecular Simulation, 24: 4, 351 — 368
To link to this Article: DOI: 10.1080/08927020008022381
URL: http://dx.doi.org/10.1080/08927020008022381

Molecular Simulation, 2000, Vol. 24, pp. 351 - 368
Reprints available directly from the publisher
Photocopying permitted by license only

FIRST-PRINCIPLES PROTEIN FOLDING SIMULATIONS

YUKO OKAMOTO*

Department of Theoretical Studies, Institute for Molecular Science and Department of Functional Molecular Science, The Graduate University for Advanced Studies, Okazaki, Aichi 444-8585, Japan

(Received May 1999; accepted May 1999)

It is widely believed that the prediction of the three-dimensional structures of proteins from the first principles is impossible. This view is based on the fact that the number of possible structures for each protein is astronomically large. The question is then why a protein folds into its native structure with the proper biological functions in the time scale of milliseconds to minutes, and this is called *Levinthal's paradox*. In this article I will discuss our strategy for attacking the protein folding problem. Our approach consists of two elements: the inclusion of accurate solvent effects and the development of powerful simulation algorithms that can avoid getting trapped in states of energy local minima. For the former, we discuss several models varying in nature from crude (distance-dependent dielectric function) to rigorous (reference interaction site model). For the latter, we show the effectiveness of Monte Carlo simulated annealing and generalized-ensemble algorithms.

Keywords: Protein folding; simulated annealing; generalised-ensemble algorithms

1. INTRODUCTION

Proteins are the most complicated molecules that exist in nature. Proteins under their native physiological conditions spontaneously fold into unique three-dimensional structures (tertiary structures) in the time scale of milliseconds to minutes. Although the tertiary structures of proteins appear to be dependent on various environmental factors within the cell, it was shown by experiments *in vitro* that unfolded proteins can refold back into their native conformations once the denaturants are removed, implying that the

*e-mail: okamotoy@ims.ac.jp

three-dimensional structure of a protein is determined solely by its amino-acid sequence information [1]. This is called the *Anfinsen's dogma*. However, the prediction of the native protein tertiary structure from the first principles has yet to be accomplished. The difficulty comes from the fact that the number of possible conformations for each protein is astronomically large [2]. Simulations by conventional methods such as Monte Carlo or molecular dynamics algorithms in canonical ensemble will necessarily get trapped in one of many local-minimum states in the energy function.

In this article, we discuss our strategy for overcoming the above difficulty. In Section 2 we explain the so-called *Levinthal's paradox*. In Section 3 we discuss the traveling salesman problem that belongs to the same class of mathematical difficulty as in the protein folding problem. In Section 4 we give the energy functions of the protein systems. In Section 5 we describe the two powerful methods, Monte Carlo simulated annealing [3] and multicanonical algorithms [4], which are effective in overcoming the common difficulty in the protein folding problem and traveling salesman problem. In Section 6 the results of our applications of these algorithms to the protein folding problem are given. Finally, Section 7 is devoted to conclusions.

2. *LEVINTHAL'S PARADOX*

The *Anfinsen's dogma* [1] led to a hope that once the correct Hamiltonian of the system is given, one can predict the native protein tertiary structure from the first principles by computer simulations. However, this has not been accomplished to date. In fact, it is generally believed that the *ab initio* prediction of protein tertiary structure is impossible. The most important basis for this widely spread view is probably the *Levinthal's paradox* [2]. Simply stated, *Levinthal's paradox* can be summarized as follows. The three-dimensional structure of the main chain of a protein is determined by the dihedral angles ϕ and ψ ($\omega = 180°$ is assumed). If only local interactions are considered, these dihedral angles have a few preferred values that correspond to the local minima of the torsion energy around each rotation bond. For instance, when we consider four atoms along a rotation bond, three values of angles in gauche ($\pm 60°$) and trans ($180°$) correspond to rather stable local minima of the torsion energy. Hence, we have to consider only about 10 conformations per each amino acid. But this means that we have to examine at least as many as 10^N conformations for a protein with N amino acids. This number increases very rapidly with N. For example, with $N = 40$ there are 10^{40} possible conformations. Considering an average oscillation frequencies of protein molecules, one can assume that

a protein can sample of the order of 10^{14} structures per second [2]. Hence, it would take this protein about 10^{26} seconds $\approx 10^{18}$ years to examine all the possible conformations. This is of course much longer than the present age of the Universe. Then why can proteins fold into their native conformations on the time scale of milliseconds to minutes? This is the *Levinthal's paradox*. Likewise, the problem can also be stated in terms of computational difficulty. Assuming the computation power of the fastest supercomputer to be 1 TFLOPS (*i.e.*, it can perform 10^{12} floating point operations per second), it would take this computer roughly $\approx 10^{20}$ years to sample all the possible conformations. This is certainly impossible.

3. "*LEVINTHAL'S PARADOX*" IN THE TRAVELING SALESMAN PROBLEM

The difficulty in the prediction of protein tertiary structures belongs to a common class of difficulty encountered in systems with frustrations (*e.g.*, traveling salesman problem, spinglass, optimum electrical circuit wiring problem, *etc.*). The problem is mathematically classified as NP complete (nonpolynomical complete). Here, NP complete means that for a system with size n, the time, $T(n)$, it takes to solve the problem grows faster than any power of n. (That is, $T(n)$ grows not like n^a but *e.g.*, a^n.) This kind of problem is impossible to solve when n becomes large. In fact, for a protein with N amino acids the rough estimate of the necessary computation time was $T(N) \sim 10^N$.

We now discuss a completely different problem that belongs to the same class, NP complete, as in the protein folding problem. This is the traveling salesman problem. The problem is stated as follows. "A salesman travels N cities by starting from a certain city and visiting each city once, and comes back to the starting city. Find the shortest path." Here, the number of possible paths is $(N-1)!$ (hence, $T(N) \sim (N-1)!$). Thus, the traveling salesman problem is also NP complete (if we compare only the N dependence of $T(N)$, it is even harder than the protein folding problem). For instance, when $N = 42$, $T(N) \sim 41! \sim 3 \times 10^{49}$. It would take a 1 TFLOPS supercomputer $\approx 10^{30}$ years to sample all the possible paths. This is again absolutely impossible.

In order to solve the traveling salesman problem numerically, let us assume, for simplicity, that cities are regularly distributed on a two-dimensional lattice. Then one can tell whether an obtained path is the shortest or not at first sight, while the number of possible paths is still the same as above. Namely, the path made of only line segments in east-west

or north-south direction is the shortest, whereas a path with diagonally connecting line segments is not. In fact, if one tries conventional minimization routines on the problem, one finds out that if the initial path is selected to be random, one never reaches the global-minimum path. This is because conventional minimization routines find only the local minimum near the initial state. However, if one applies *Monte Carlo simulated annealing* [3] or *multicanonical algorithms* [4], which are algorithms that can avoid getting trapped in local minima, one can find a shortest path even for the case $N = 420$ by several hours of computations on a workstation (see Fig. 1) (A. Mitsutake and Y. Okamoto, unpublished). I will call this "*Levinthal's paradox*" in the traveling salesman problem.

Why did the impossible become possible? The answer can be stated as follows. In general, the number, $n(E)$, of paths with length E grows rapidly with E. This is a common characteristic in systems with many degrees of freedom. For example, if we consider E as energy, the number of states, $n(E)$, with energy E for the system of monatomic ideal gas with N molecules grows as a power of E (*i.e.*, $n(E) \propto E^{3N/2}$). Hence, in the traveling

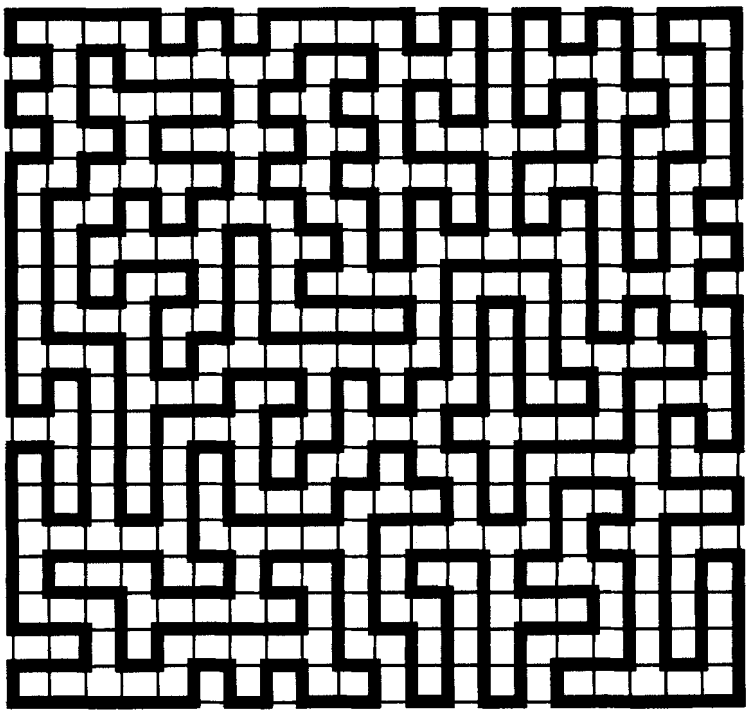

FIGURE 1 A solution to the 420-city traveling salesman problem obtained by a simulation.

salesman problem, there are much more long paths (as shown in Fig. 2a) than short paths (as shown in Fig. 2b). Then if we use the algorithms that can avoid getting trapped in local minima, we can find the shortest path (the global-minimum-energy state) by examining much fewer paths (states) without ever sampling most of long paths (high energy states) [5].

In Figure 3a, a generic density of states $n(E)$ is depicted. This rapidly increasing characteristic of the density of states can also be represented in

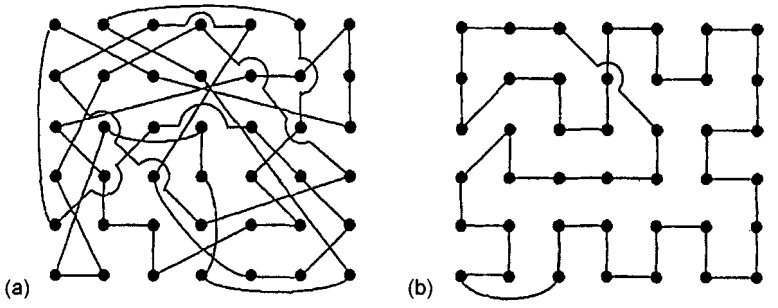

FIGURE 2 Examples of paths in the 42-city traveling salesman problem. (a) A long path, (b) A short path.

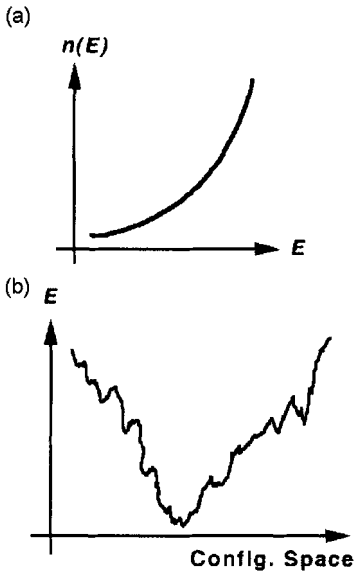

FIGURE 3 (a) A generic density of states of a system with many degrees of freedom. It is a rapidly increasing function of energy, (b) A funnel picture based on the density of states of (a).

a slightly different way as in Figure 3b, where the *abscissa* represents the configuration space. The point is that in Levinthal's argument, one of important parameters, namely *temperature*, was not taken into consideration. Although there exist astronomically large number of local-minimum states in total, the number of relevant local-minimum states at *room temperature* is much smaller. Many people now call this picture a funnel [6].

4. ENERGY FUNCTIONS OF PROTEIN SYSTEMS

The energy function for the protein systems is given by the sum of two terms: the conformational energy E_P for the protein molecule itself and the solvation free energy E_S for the interaction of protein with the surrounding solvent. The conformational energy function E_P (in kcal/mol) for the protein molecule that we used is one of the standard ones. Namely, it is given by the sum of the electrostatic term E_C, 12–6 Lennard-Jones term E_{LJ}, and hydrogen-bond term E_{HB} for all pairs of atoms in the molecule together with the torsion term E_{tor} for all torsion angles:

$$\begin{aligned} E_P &= E_C + E_{LJ} + E_{HB} + E_{tor}, \\ E_C &= \sum_{(i,j)} \frac{332\ q_i q_j}{\varepsilon r_{ij}}, \\ E_{LJ} &= \sum_{(i,j)} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right), \\ E_{HB} &= \sum_{(i,j)} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right), \\ E_{tor} &= \sum_{i} U_i (1 \pm \cos(n_i \chi^i)). \end{aligned} \tag{1}$$

Here, r_{ij} is the distance (in Å) between atoms i and j, ε is the dielectric constant, and χ^i is the torsion angle for the chemical bond i. Each atom is expressed by a point at its center of mass, and the partial charge q_i (in units of electronic charges) is assumed to be concentrated at that point. The factor 332 in E_C is a constant to express energy in units of kcal/mol. These parameters in the energy function as well as the molecular geometry were adopted from ECEPP/2 [7]. The computer code KONF90 [8] was used for all the Monte Carlo simulations. For gas phase simulations, we set the dielectric constant ε equal to 2. The peptide-bond dihedral angles ω were fixed at their usual experimental value 180° for simplicity. So, the remaining

dihedral angles ϕ and ψ in the main chain and χ in the side chains constitute the variables to be updated in the simulations. One Monte Carlo (MC) sweep consists of updating all these angles once with Metropolis evaluation [9] for each update.

One of the simplest ways to represent solvent effects is by the sigmoidal, distance-dependent dielectric function [10]. The explicit form of the function we used is given by [11]

$$\varepsilon(r) = D - \frac{D-2}{2}[(sr)^2 + 2sr + 2]e^{-sr}, \tag{2}$$

which is a slight modification of the one used in Ref. [12]. Here, we use $s = 0.3$ and $D = 78$. It approaches 2 (the value inside a protein) in the limit the distance r going to zero and 78 (the value for bulk water) in the limit r going to infinity. The distance-dependent dielectric function is simple and also computationally only slightly more demanding than the gas-phase case.

Another commonly used term that represents solvent contributions more accurately than the distance-dependent dielectric function is the term proportional to the solvent-accessible surface area of protein molecule. The solvation free energy E_S in this approximation is given by

$$E_S = \sum_i \sigma_i A_i, \tag{3}$$

where A_i is the solvent-accessible surface area of i-th functional group, and σ_i is the proportionality constant. There are several versions of the set of the proportionality constants and functional groups. Five parameter sets were compared for the systems of peptides and a small protein, and we found that the parameter sets of Refs. [13, 14] are valid ones [15].

The most widely-used and rigorous method of inclusion of solvent effects is probably the one that deals with the explicit solvent molecules with all-atom representations. Many molecular dynamics simulations of protein systems now directly include these explicit solvent molecules (for a review, see, for instance, Ref. [16]). Another rigorous method is based on the statistical mechanical theory of liquid and solution and is called the reference interaction site model (RISM) [17]. A robust and fast algorithm for solving RISM equations was recently developed [18], which made folding simulations of peptides a feasible possibility [19]. Although this method is computationally much more time-consuming than the first two methods (terms with distance-dependent dielectric function and those proportional to surface area), it gives the most accurate representation of the solvation free energy.

5. SIMULATION METHODS

Once the appropriate energy function of the protein system is given, we have to employ a simulation method that does not get trapped in states of energy local minima. We have been advocating the uses of Monte Carlo simulated annealing [3] and multicanonical algorithm [4] (for reviews, see Refs. [20, 21]).

5.1. Monte Carlo Simulated Annealing

In the regular canonical ensemble with a given inverse temperature $\beta \equiv 1/k_B T$, the probability weight of each state with energy E is given by the Boltzmann factor:

$$W_B(E) = \exp(-\beta E). \tag{4}$$

The probability distribution in energy is then given by

$$P_B(T, E) \propto n(E) W_B(E), \tag{5}$$

where $n(E)$ is the density of states. Since the density of states $n(E)$ is a rapidly increasing function of E and the Boltzmann factor $W_B(E)$ decreases exponentially with E, the probability distribution $P_B(T, E)$ has a bell-like shape in general. When the temperature is high, β is small, and $W_B(E)$ decreases slowly with E. So, $P_B(T, E)$ has a wide bell-shape. On the other hand, at low temperature β is large, and $W_B(E)$ decreases rapidly with E. So, $P_B(T, E)$ has a narrow bell-shape (and in the limit $T \to 0\,\mathrm{K}$, $P_B(E) \propto \delta(E - E_{\min})$, where $E_{\min}$ is the global-minimum energy). This is shown in Figure 4a. If the system is an NP complete problem, however, it is very difficult to obtain canonical distributions at low temperatures with these methods by starting a simulation from a random initial state. This is because the simulation will certainly get trapped in energy local minima. Note also that if an initial state is selected randomly, it will necessarily have a high energy. This is because the number of states is generally a rapidly increasing function of energy (see Fig. 3a).

Simulated annealing [3] is based on the process of crystal making. Namely, by starting a simulation at a sufficiently high temperature (much above the melting temperature), one lowers the temperature gradually during the simulation until it reaches the global-minimum-energy state (crystal). As shown in Figure 4a, if the rate of temperature decrease is sufficiently slow so that thermal equilibrium may be maintained throughout

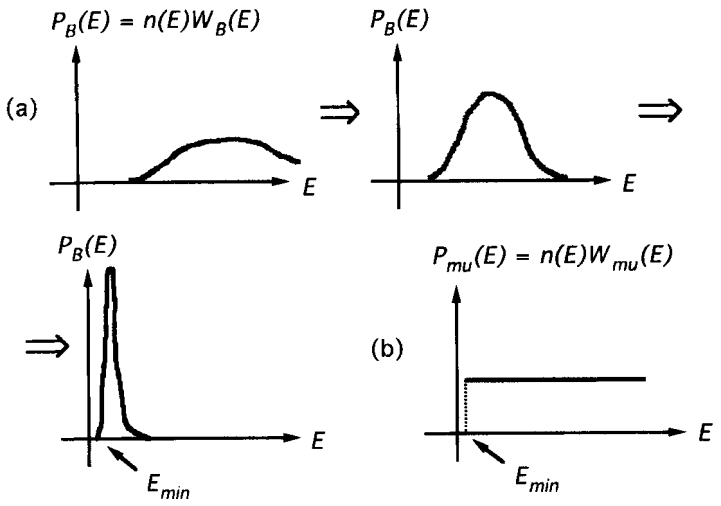

FIGURE 4 (a) Change of canonical probability distribution during a Monte Carlo simulated annealing run. The simulation is started at a high temperature, and the temperature is lowered very slowly, (b) The probability distribution of energy in multicanonical ensemble. It is proportional to a step function $\theta(E - E_{\min})$, where $E_{\min}$ is the global-minimum energy.

the simulation, only the state with the global energy minimum is obtained (when the final temperature is 0 K). However, if the temperature decrease is rapid (quenching), the simulation will get trapped in a state of energy local minimum in the vicinity of the initial state.

Simulated annealing was first successfully used to predict the global-minimum-energy conformations of polypeptides and proteins [22–24] and to refine protein structures from X-ray and NMR data [25,26] almost a decade ago. Since then this method has been extensively used in the protein folding and structure refinement problems (for reviews, see Refs. [20,27]). Our group has been testing the effectiveness of the method mainly in oligopeptide systems. The procedure of our approach is as follows. While the initial conformations in the protein simulations are usually taken from the structures inferred by the experiments, our initial conformations are *randomly generated.* Each Monte Carlo sweep updates every dihedral angle (in both the main chain and side chains) once. Our annealing schedule is as follows: The temperature is lowered exponentially from $T_I = 1000$ K to $T_F = 250$ K (the final temperature T_F was sometimes set equal to 100 K, 50 K, or 1 K) [8]. The temperature for the n-th MC sweep is given by [8]

$$T_n = T_I \gamma^{n-1}, \tag{6}$$

where γ is a constant which is determined by T_I, T_F, and the total number of MC sweeps of the run. Each run consisted of $10^4 \sim 10^6$ MC sweeps, and we usually made 10 to 20 runs from randomly-generated initial conformations.

5.2. Generalized-ensemble Algorithms

While a regular Monte Carlo method generates states according to the canonical distribution, generalized-ensemble algorithms [28] generate states so that a one-dimensional random walk in a pre-chosen physical quantity (for instance, the energy) is realized. Hence, any energy barrier can be overcome, and one can avoid getting trapped in states of energy local minima.

Multicanonical algorithm [4] is one of the most well-known such methods. In the "multicanonical ensemble" the probability distribution of energy is *defined* as follows:

$$P_{\mathrm{mu}}(E) \propto n(E) W_{\mathrm{mu}}(E) = \text{constant}. \tag{7}$$

This is illustrated in Figure 4b. The multicanonical weight factor then satisfies

$$W_{\mathrm{mu}}(E) \propto n^{-1}(E). \tag{8}$$

Since this weight factor is not *a priori* known, one has to determine it for each system by a few iterations of trial Monte Carlo simulations. See Refs. [29] and [30] for details of the method to determine the multicanonical weight factor $W_{\mathrm{mu}}(E)$. Once this weight factor is obtained, one performs a long production simulation run. The advantage of multicanonical algorithms lies in the fact that from this single production run, one can obtain not only the global-minimum-energy state but also the canonical distribution $P_B(T, E) = n(E)\mathrm{e}^{-\beta}$ for wide range of temperatures $T = 1/k_B\beta$. The latter is accomplished by the use of the reweighting techniques [31]. Namely, $P_B(T, E)$ can be expressed in terms of the predetermined weight $W_{\mathrm{mu}}(E)$ and the obtained distribution $P_{\mathrm{mu}}(E)$ as follows:

$$P_B(T, E) = \frac{P_{\mathrm{mu}}(E) W_{\mathrm{mu}}^{-1}(E) \mathrm{e}^{-\beta E}}{\int dE' P_{\mathrm{mu}}(E') W_{\mathrm{mu}}^{-1}(E') \mathrm{e}^{-\beta E'}}. \tag{9}$$

The expectation value of a physical quantity $\mathcal{A}$ at temperature T is then given by

$$\langle \mathcal{A} \rangle_T = \int dE\, \mathcal{A}(E) P_B(T, E). \tag{10}$$

The application of multicanonical algorithm and its variants to the prediction of protein tertiary structures was proposed several years ago [32, 33]. Since then there have been various applications of the method in the protein folding problem (for reviews, see Refs. [20, 21]). A formulation of multicanonical algorithm for the molecular dynamics method was also developed [34–36].

While multicanonical algorithm performs a free random walk in energy space, simulated tempering [37] performs a free random walk in temperature, and $1/k$ sampling [38] a free random walk in entropy. All these generalized-ensemble simulations in turn result in (weighted) random walk in energy space, allowing the escape from states of energy local minima. The performances of these three algorithms for the protein folding problem were recently compared [28].

The weight factors for the generalized-ensemble algorithms are not *a priori* known and usually have to be determined by interactive procedures. This weight determination is often non-trivial, and thus an easy-to-determine weight factors are in demand. We have proposed to use a simple weight factor [39] that is a generalization of the weight in the Tsallis statistics [40].

With all these generalized-ensemble algorithms, we first determined the weight factors by iterations of short preliminary runs. We then made one long production run of 200,000 to 1,000,000 MC sweeps from a random initial conformation for each system.

6. RESULTS

We now present the results of our simulations based on Monte Carlo simulated annealing and generalized ensembles. *All the simulations were started from randomly-generated conformations.*

Since simulations in generalized ensembles can sample much wider configurational space than conventional methods, the method is particularly effective for studying the free energy landscape of protein systems. Let us emphasize again that a single simulation run in a generalized ensemble can give various thermodynamics quantities as a function of temperature. Performing a Monte Carlo simulation of 1,000,000 MC sweeps for Met-enkephalin in gas phase ($\varepsilon = 2$) by the newly-developed generalized-ensemble algorithm [39], we have calculated the characteristic temperatures of folding of this peptide [41] and obtained a detailed picture of the free-energy landscape [42].

As an example for such calculations, we show the average volume and its derivative with respect to temperature in Figure 5.

The peak in the derivative of average volume with respect to temperature gives the collapse temperature T_θ, and from the figure we have $T_\theta = 280 \pm 20$ K [41]. The reader is referred to Refs. [41] and [42] for detailed analyses of the free-energy landscape of Met-enkephalin in gas phase.

We now present our results of the first-principles predictions of the tertiary structures of peptides and proteins.

The first example is Met-enkephalin again. This peptide consists of 5 amino acids with the amino-acid sequence: Tyr-Gly-Gly-Phe-Met. In Figure 6 we compare the 5 superposed structures inferred from NMR experiments (Fig. 2 of Ref. [43]) and the 8 superposed ones of the lowest-energy conformations from independent Monte Carlo simulated annealing runs in water (the solvent contributions were calculated by the RISM theory) [19]. The figures were created with RasMol [44]. We see a striking similarity between simulation results in water and those of NMR experiments.

The solvation free energy based on the RISM theory is very accurate, but it is also computationally very demanding. We are currently trying to solve this problem making the algorithm more efficient and robust [18]. Hereafter, we discuss how well other solvation theories can still describe

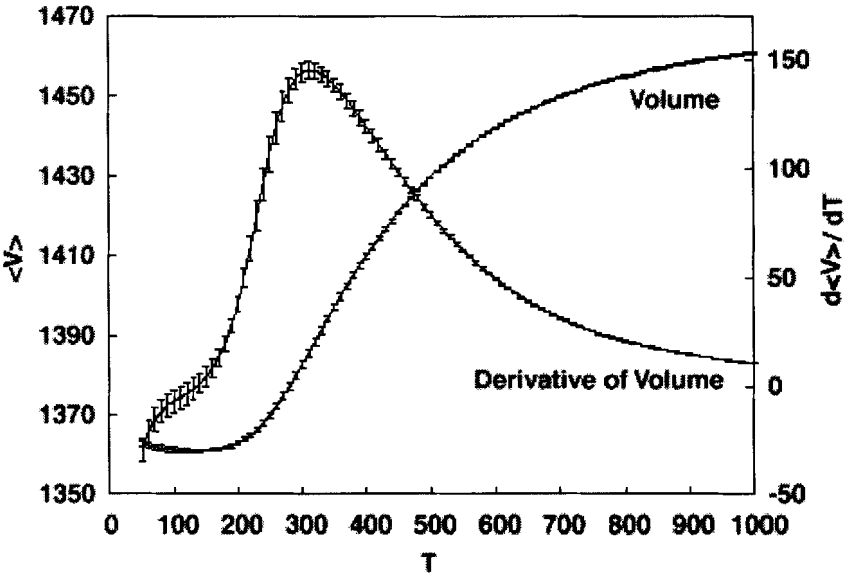

FIGURE 5 Average volume V(Å^3) (a) and its derivative with respect to temperature T(K) (b).

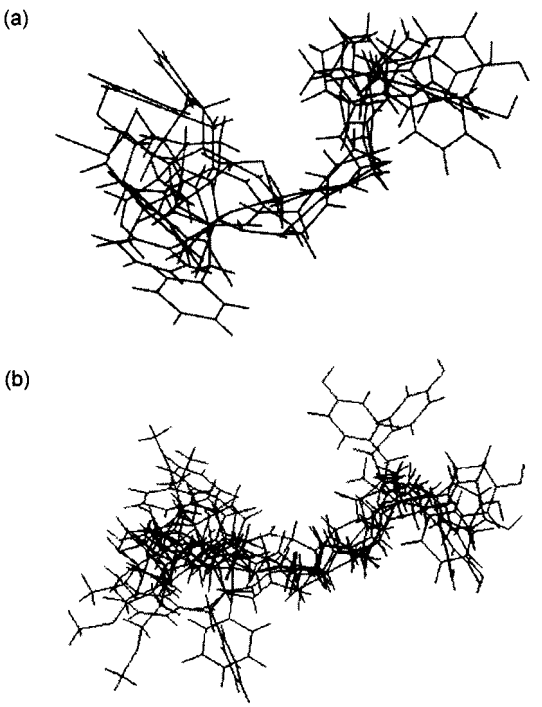

FIGURE 6 (a) Superposition of five conformations of Met-enkephalin deduced from the NMR experiment, (b) Superposition of the eight conformations obtained as the lowest-energy structures by Monte Carlo simulated annealing in water.

the effects of solvent in the prediction of three-dimensional structures of oligopeptides and small proteins.

Next example is the C-peptide, residues 1 – 13 of ribonuclease A. This peptide has the amino-acid sequence: Lys-Glu-Thr-Ala-Ala-Ala-Lys-Phe-Glu-Arg-Gln-His-Met. It is known from the X-ray diffraction data of the whole enzyme that the segment from Ala-4 to Gln-11 exhibits a nearly 3-turn α-helix [45]. It was also found by CD [46] and NMR [47] experiments that the isolated C-peptide also has significant α-helix formation in aqueous solution at temperatures near 273 K.

We have performed a multicanonical simulation of 1,000,000 MC sweeps for C-peptide with the inclusion of solvent effects by the distance-dependent dielectric function (see Eq. (2)) [48]. The lowest-energy conformation obtained has an α-helix from Ala-4 to Gln-11 and does have the characteristic salt bridge between Glu-2^- and Arg-10^+. This conformation and the corresponding X-ray structure are compared in Figure 7. The figures were created with Molscript [49] and Raster3D [50].

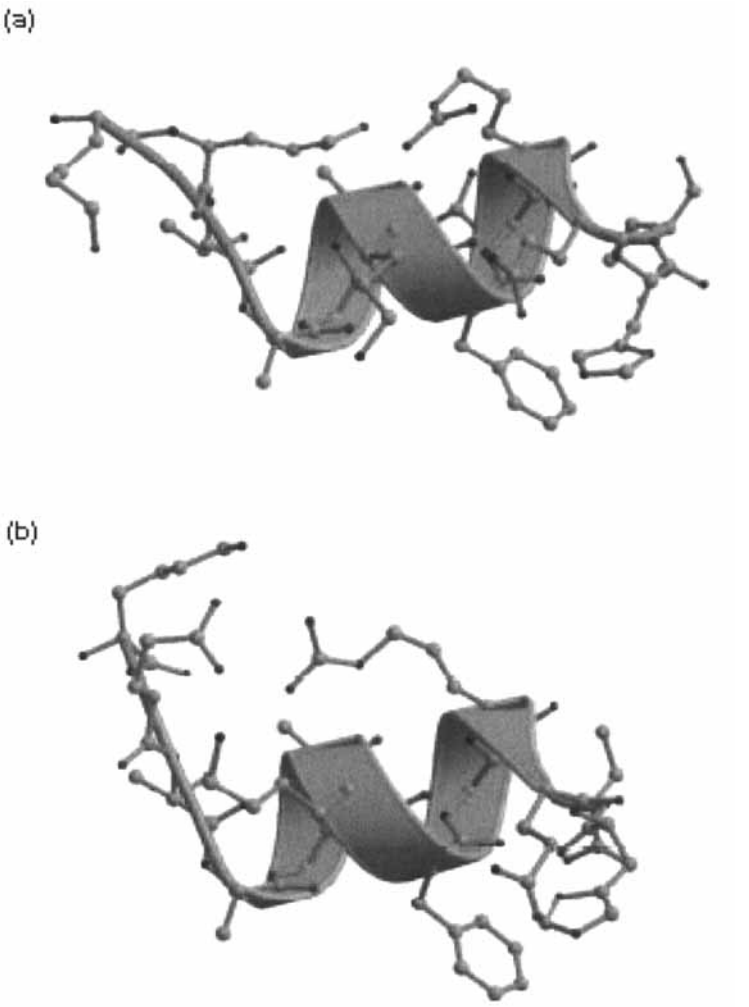

FIGURE 7 X-ray structure of C-peptide (a) and the lowest-energy conformation of C-peptide obtained from a multicanonical Monte Carlo run with the distance-dependent dielectric function (b).

The similarity between the two is apparent. The root-mean-square deviations between the two are 1.4 Å and 2.7 Å for non-hydrogen atoms in the backbone and in the whole molecule, respectively [48].

We have also studied β-sheet formations by Monte Carlo simulated annealing. The peptide that we studied is the fragment corresponding to residues 16–36 of bovine pancreatic trypsin inhibitor (BPTI) and has the

amino-acid sequence: Ala[16]-Arg-Ile-Ile-Arg-Tyr-Phe-Tyr-Asn-Ala-Lys- Ala-Gly-Leu-Cys-Gln-Thr-Phe-Val-Tyr-Gly[36]. An antiparallel β-sheet structure in residues 18–35 is observed in X-ray crystallographic data of the whole protein [51].

BPTI(16–36) was studied in aqueous solution that is represented by solvent-accessible surface area of Eq. (3) by Monte Carlo simulated annealing [52]. Twenty simulation runs of 100,000 MC sweeps were made. It was indeed found that the lowest-energy structure obtained has a β-sheet structure. This structure and that deduced from the X-ray experiments are compared in Figure 8. The figures were created with Molscript [49] and Raster3D [50].

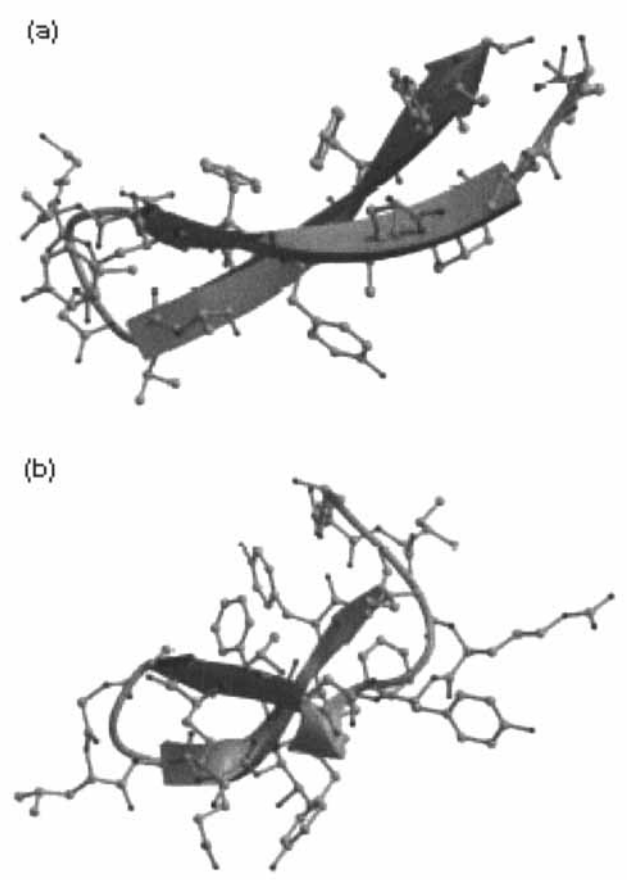

FIGURE 8 The structure of BPTI(16–36) deduced from X-ray experiments (a) and the lowest-energy conformation of BPTI(16–36) obtained from 20 Monte Carlo simulated annealing runs in aqueous solution represented by solvent-accessible surface area (b).

Although both conformations are β-sheet structures, there are important differences between the two: The positions and types of the turns are different. Since the X-ray structure is taken from the experiments on the whole BPTI molecule, it does not have to agree with that of the isolated BPTI(16–36) fragment. It was found that the simulated results in Figure 8b have remarkable agreement with those in the NMR experiments of the isolated fragment [52, 53].

7. CONCLUSIONS

In this article I have reviewed theoretical aspects of the protein folding problem and presented the results of my own work. Our strategy in attacking this problem consists of two steps: (1) inclusion of accurate solvent effects, (2) development of powerful simulation algorithms that can avoid getting trapped in states of energy local minima.

As for the simulation algorithms, we have shown the effectiveness of Monte Carlo simulated annealing and generalized-ensemble algorithms by directly folding α-helix and β-sheet structures from randomly-generated initial conformations. The latter methods are particularly useful in the sense that one can obtain from a single simulation run not only the conformation with the global-minimum energy but also any thermodynamic quantity as a function of temperature. We have presented various examples of such calculations.

As for the solvent effects, we considered several methods: a distance-dependent dielectric function, a term proportional to solvent-accessible surface area, and the reference interaction site model (RISM). These methods vary in nature from crude but computationally inexpensive (distance-dependent dielectric function) to accurate but computationally demanding (RISM theory). In the present article, we have shown that the inclusion of some solvent effects is very important for a successful prediction of the tertiary structures of small peptides and proteins.

Acknowledgments

The author is grateful to his collaborators for discussions and suggestions. Especially he would like to thank H. Kawai, T. Kikuchi, T. Nakazawa, and M. Fukugita (for the work with Monte Carlo simulated annealing), U. H. E. Hansmann and A. Mitsutake (for the work with generalized-ensemble algorithms), F. Hirata, M. Kinoshita, and M. Masuya (for the work with

various solvation free energy contributions), and J. Onuchic (for the work on free energy landscape). The simulations were performed on the computers at the Institute for Molecular Science. This work was supported, in part, by a grant from the Research for the Future Program of Japan Society for the Promotion of Science (JSPS-RFTF98P01101).

References

[1] Epstain, C. J., Goldberger, R. F. and Anfinsen, C. B. (1963). *Cold Spring Harbor Symp. Quant. Biol.*, **28**, 439.
[2] (a) Levinthal, C. (1968). *J. Chim. Phys.*, **65**, 44; (b) Wetlaufer, D. B. (1973). *Proc. Natl. Acad. Sci. USA*, **70**, 691.
[3] For example, Kirkpatrick, S., Gelatt, C. D. Jr. and Vecchi, M. P. (1983). *Science*, **220**, 671.
[4] Berg, B. A. and Neuhaus, T. (1991). *Phys. Lett.*, **B267**, 249; *Phys. Rev. Lett.*, **68**, 9 (1992).
[5] Okamoto, Y. (1995). *Assoc. Asia Pacif. Phys. Soc. Bull.*, **5**(Nos. 3 and 4), 4.
[6] (a) Leopold, P. E., Montal, M. and Onuchic, J. N. (1992). *Proc. Natl. Acad. Sci. USA*, **89**, 8721; (b) Bryngelson, J. D., Onuchic, J. N., Socci, N. D. and Wolynes, P. G. (1995). *PROTEINS: Struct. Funct. Genet.*, **21**, 167; (c) Dill, K. A. and Chan, H. S. (1997). *Nature Struc. Biol.*, **4**, 10.
[7] Momany, F. A., McGuire, R. F., Burgess, A. W. and Scheraga, H. A. (1975). *J. Phys. Chem.*, **79**, 2361; Némethy, G., Pottle, M. S. and Scheraga, H. A. (1983) *J. Phys. Chem.*, **87**, 1883; Sippl, M. J., Némethy G. and Scheraga H. A. (1984). *J. Phys. Chem.*, **88**, 6231.
[8] Kawai, H., Okamoto, Y., Fukugita, M., Nakazawa, T. and Kikuchi, T. (1991). *Chem. Lett.*, 1991, 213; Okamoto, Y., Fukugita, M., Nakazawa, T. and Kawai, H. (1991). *Protein Eng.*, **4**, 639.
[9] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). *J. Chem. Phys.*, **21**, 1087.
[10] (a) Hingerty, B. E., Ritchie, R. H., Ferrell, T. and Turner, J. E. (1985). *Biopolymers*, **24**, 427; (b) Ramstein, J. and Lavery, R. (1988). *Proc. Natl. Acad. Sci. USA*, **85**, 7231.
[11] Okamoto, Y. (1994). *Biopolymers*, **34**, 529.
[12] Daggett, V., Kollman, P. A. and Kuntz, I. D. (1991). *Biopolymers*, **31**, 285.
[13] Ooi, T., Oobatake, M., Némethy, G. and Scheraga, H. A. (1987). *Proc. Natl. Acad. Sci. USA*, **84**, 3086.
[14] Wesson, L. and Eisenberg, D. (1992). *Protein Sci.*, **1**, 227.
[15] Masuya, M. and Okamoto, Y., in preparation.
[16] Brooks III, C. L. (1998). *Curr. Opin. Struct. Biol.*, **8**, 222.
[17] (a) Chandler, D. and Andersen, H. C. (1972). *J. Chem. Phys.*, **57**, 1930; (b) Hirata, F. and Rossky, P. J. (1981). *Chem. Phys. Lett.*, **83**, 329.
[18] Kinoshita, M., Okamoto, Y. and Hirata, F. (1997). *J. Comp. Chem.*, **18**, 1320.
[19] Kinoshita, M., Okamoto, Y. and Hirata, F. (1998). *J. Am. Chem. Soc.*, **120**, 1855.
[20] Okamoto, Y. (1998). *Recent Research Devel. Pure Applied Chem.*, **2**, 1.
[21] Hansmann, U. H. E. and Okamoto, Y., In: *Annual Reviews of Computational Physics VI*, Stauffer, D. (Ed.) (Singapore, World Scientific, 1999), pp. 129–157.
[22] Wilson, S. R., Cui, W., Moskowitz, J. W. and Schmidt, K. E. (1988). *Tetrahedron Lett.*, **29**, 4373.
[23] Kawai, H., Kikuchi, T. and Okamoto, Y. (1989). *Protein Eng.*, **3**, 85.
[24] Wilson, C. and Doniach, S., *PROTEINS: Struct. Funct. Genet.*, **6**, 193.
[25] Brünger, A. T. (1988). *J. Mol. Biol.*, **203**, 803.
[26] Nilges, M., Clore, G. M. and Gronenborn, A. M. (1988). *FEBS Lett.*, **229**, 317.
[27] Wilson, S. R. and Cui, W., In: "*The Protein Folding Problem and Tertiary Structure Prediction*" Merz, K. M. Jr. and Le Grand, S. M. (Eds.) (Birkhäuser, 1994), pp. 43–70.
[28] Hansmann, U. H. E. and Okamoto, Y. (1997). *J. Comp. Chem.*, **18**, 920.

[29] Berg, B. A. (1992). *Int. J. Mod. Phys.*, **C3**, 1083.
[30] Hansmann, U. H. E. and Okamoto, Y. (1994). *J. Phys. Soc. Jpn.*, **63**, 3945; *Physica A*, **212**, 415 (1994).
[31] Ferrenberg, A. M. and Swendsen, R. H. (1988). *Phys. Rev. Lett.*, **61**, 2635; *ibid.*, **63**, 1658(E) (1989).
[32] Hansmann, U. H. E. and Okamoto, Y. (1993). *J. Comp. Chem.*, **14**, 1333.
[33] Hao, M. H. and Scheraga, H. A. (1994). *J. Phys. Chem.*, **98**, 4940.
[34] Hansmann, U. H. E., Okamoto, Y. and Eisenmenger, F. (1996). *Chem. Phys. Lett.*, **259**, 321.
[35] Nakajima, N., Nakamura, H. and Kidera, A. (1997). *J. Phys. Chem.*, **101**, 817.
[36] Bartels, C. and Karplus, M. (1998). *J. Phys. Chem. B*, **102**, 865.
[37] (a) Lyubartsev, A. P., Martinovski, A. A., Shevkunov, S. V. and Vorontsov-Velyaminov, P. N. (1992). *J. Chem. Phys.*, **96**, 1776; (b) Marinari, E. and Parisi, G., *Europhys. Lett.*, **19**, 451.
[38] Hesselbo, B. and Stinchcombe, R. B. (1995). *Phys. Rev. Lett.*, **74**, 2151.
[39] (a) Hansmann, U. H. E. and Okamoto, Y. (1997). *Phys. Rev. E*, **56**, 2228; (b) Hansmann, U. H. E. Eisenmenger, F. and Okamoto, Y. (1998). *Chem. Phys. Lett.*, **297**, 374.
[40] Tsallis, C. (1988). *J. Stat. Phys.*, **52**, 479.
[41] Hansmann, U. H. E., Masuya, M. and Okamoto, Y. (1997). *Proc. Natl. Acad. Sci. USA*, **94**, 10652.
[42] Hansmann, U. H. E., Okamoto, Y. and Onuchic, J. N. (1999). *PROTEINS: Struct. Funct. Genet.*, **34**, 472.
[43] Graham, W. H., Carter II, E. S. and Hicks, R. P. (1992). *Biopolymers*, **32**, 1755.
[44] Sayle, R. A. and Milner-White, E. J. (1995). *TIBS*, **20**, 374.
[45] (a) Wychoff, H. W., Tsernoglou, D., Hanson, A. W., Knox, J. R., Lee, B. and Richards, F. M. (1970). *J. Biol. Chem.*, **245**, 305; (b) Tilton, R. F. Jr., Dewan, J. C. and Petsko, G. A. (1992). *Biochemistry*, **31**, 2469.
[46] Shoemaker, K. R., Kim, P. S., Brems, D. N., Marqusee, S., York, E. J., Chaiken, I. M., Stewart, J. M. and Baldwin, R. L. (1985). *Proc. Natl. Acad. Sci. USA*, **82**, 2349.
[47] Osterhout, J. J., Baldwin, R. L., York, E. J., Stewart, J. M., Dyson, H. J. and Wright, P. E. (1989). *Biochemistry*, **28**, 7059.
[48] Hansmann, U. H. E. and Okamoto, Y. (1998). *J. Phys. Chem. B*, **102**, 653; *ibid.*, **103**, 1595 (1999).
[49] Kraulis, P. J. (1991). *J. Appl. Cryst.*, **24**, 946.
[50] (a) Bacon, D. and Anderson, W. F. (1988). *J. Mol. Graphics*, **6**, 219; (b) Merritt, E. A. and Murphy, M. E. P. (1994). *Acta Cryst.*, **D50**, 869.
[51] Deisenhofer, J. and Steigemann, W. (1985). *Acta Crystallogr.*, **B31**, 238.
[52] Okamoto, Y., Masuya, M., Nabeshima, M. and Nakazawa, T. (1999). *Chem. Phys. Lett.*, **299**, 17.
[53] Nakazawa, T., Okamoto, Y., Kobayashi, Y., Kyogoku, Y. and Aimoto, S., in preparation.