Tackling the Protein Folding Problem by a Generalized-Ensemble Approach with Tsallis Statistics

Ulrich H.E. Hansmann[a,*] and Yuko Okamoto[b,†]

[a] *Department of Physics*
Michigan Technological University
Houghton, MI 49931-1291, U.S.A.
[b] *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 07 December, 1998

We review uses of Tsallis statistical mechanics in the protein folding problem. Monte Carlo simulated annealing algorithm and generalized-ensemble algorithm with both Monte Carlo and stochastic dynamics algorithms are discussed. Simulations by these algorithms are performed for a penta peptide, Met-enkephalin. In particular, for generalized-ensemble algorithms, it is shown that from only one simulation run one can find the global-minimum-energy conformation and obtain probability distributions in canonical ensemble for a wide temperature range, which allows the calculation of any thermodynamic quantity as a function of temperature.

I. Introduction

For many important physical systems like biological macromolecules it is very difficult to obtain the accurate canonical distribution at low temperatures by conventional simulation methods. This is because the energy function has a huge number of local minima separated by high energy barriers, and at low temperatures simulations will necessarily get trapped in the configurations corresponding to one of these local minima. In order to overcome this multiple-minima problem, many methods have been proposed. Simulated annealing [1] is probably the most widely used algorithm that can alleviate the difficulty. Generalized-ensemble algorithms, most well-known of which is the multicanonical approach [2, 3], are also powerful ones, and were first introduced to the protein-folding problem in Ref. [4]. Simulations in the multicanonical ensemble perform 1D random walk in energy space, which allows the system to overcome any energy barrier. Besides multicanonical algorithms, simulated tempering [5, 6] and *1/k*-sampling [7] have been shown to be equally effective generalized-ensemble methods in the protein folding problem [8]. The simulations are usually performed with Monte Carlo (MC) scheme, but recently molecular dynamics (MD) version of multicanonical algorithm was also developed [9]-[11].

The generalized-ensemble approach is based on non-Boltzmann probability weight factors, and in the above three methods the determination of the weight factors is non-trivial. We have shown that a particular choice of the Tsallis weight factor [12] can be used for generalized-ensemble simulations [13, 14]. The advantage of this ensemble is that it greatly simplifies the determination of the weight factor.

In this article, we review simulated annealing and generalized-ensemble algorithms based on Tsallis statis-

* e-mail: hansmann@mtu.edu
† e-mail: okamotoy@ims.ac.jp

tics. The performances of the algorithms are tested with the system of an oligopeptide, Met-enkephalin.

II. Methods

II.1 Energy Function of Protein Systems

The total potential energy function E_{tot} for the protein systems 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:

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

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 [15]. The computer code KONF90 [16] was used for the MC simulations and SMC [17] was used for the MD simulations. We neglected the solvent contributions for simplicity and set the dielectric constant ϵ equal to 2. The peptide-bond dihedral angles ω were fixed at the 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 MC sweep consists of updating all these angles once with Metropolis evaluation [18] for each update.

II.2 Simulated Annealing Algorithm with Tsallis Statistics

In the canonical ensemble at temperature T each state with potential energy E is weighted by the Boltzmann factor

$$w_B(E,T) = e^{-\beta E} \ , \tag{2}$$

where the inverse temperature is given by $\beta = 1/k_B T$ with Boltzmann constant k_B. This weight factor gives the usual bell-shaped canonical probability distribution of energy:

$$P_B(E,T) \propto n(E)\ w_B(E,T) \ , \tag{3}$$

where $n(E)$ is the density of states. For systems with many degrees of freedom, it is usually very difficult to generate a canonical distribution at low temperatures. This is because there are many local minima in the energy function, and simulations will get trapped in states of these local minima.

A now almost classical way to alleviate this difficulty is simulated annealing [1]. Its underlying idea of modeling the crystal grow process in nature is easy to understand and simple to implement. Any MC or MD technique can be converted into a simulated annealing algorithm in a straightforward manner. During a simulation temperature is lowered very slowly from a sufficiently high initial temperature T_I where the structure changes freely with MC or MD updates to a "freezing" temperature T_F where the system undergoes no significant changes with respect to the MC sweeps or MD steps. If the rate of temperature decrease is slow enough for the system to stay in thermodynamic equilibrium, then it is ensured that the system can avoid getting trapped in local minima and that the global minimum will be found.

The performance of simulated annealing depends crucially on the annealing schedule. It could be shown that convergence to the global minimum can be secured for a logarithmic annealing schedule [19], but this is of little use in applications of the method. Constraints in available computer time enforce the choice of faster annealing schedules where success is no longer guaranteed. As in the growth of real crystals, which can hardly be

achieved by a simple cooling process, elaborated and system-specific annealing schedules are often necessary to obtain the global minimum within the CPU time available. In our simulations the temperature was lowered exponentially in N_S (number of total MC sweeps) steps by setting the inverse temperature $\beta = 1/k_B T$ at k-th MC sweep to [16, 20]

$$\beta_k = \beta_I \gamma^{k-1} \ . \tag{4}$$

Here, β_I is the initial inverse temperature and γ is given in terms of initial and final temperatures by

$$\gamma = \left(\frac{\beta_F}{\beta_I}\right)^{\frac{1}{N_S-1}} = \left(\frac{T_I}{T_F}\right)^{\frac{1}{N_S-1}} \ . \tag{5}$$

For a fixed value of the total MC sweeps N_S, the initial and final temperatures (T_I and T_F) are free parameters and have to be tuned in such a way that the annealing process is optimized for the specific problem.

Simulated annealing was first successfully used to predict the global minimum-energy conformations of polypeptides and proteins in Refs. [21]-[23] and to refine protein structures from NMR and X-ray data in Refs. [24, 25]. Since then many promising results have been obtained. Latest applications include Refs. [26]-[29].

Attempts have been made to improve the performance of simulated annealing in practical applications, see for instance Refs. [30, 20, 8]. More recent attempts [31]-[34] are inspired by Tsallis generalized statistical mechanics [12].

In the Tsallis formalism [12], a generalized statistical mechanics is constructed by maximizing a generalized entropy

$$S = -k_B \frac{1 - \sum_i p_i^q}{q-1} \tag{6}$$

with the constraints

$$\sum_i p_i = 1 \ , \qquad \sum_i p_i^q E_i = \text{const} \ . \tag{7}$$

Here, q is a real number. A generalized probability weight

$$w(E) \propto [1 + (q-1)\beta E]^{-\frac{1}{q-1}} \tag{8}$$

follows, which tends to the Boltzmann factor of Eq. (2) for $q \to 1$, and therefore regular statistical mechanics is recovered in this limit. The important feature of Tsallis generalized statistical mechanics for optimization problems is that the probability distribution of energy does no longer decrease exponentially with energy but according to a power law, where the exponent is determined by the free parameter q (compare Eqs. (2) and (8)). The resulting probability distribution has a tail to higher energies for $q > 1$, which enhances the excursion into high-energy regions and escape from local-minimum states.

This observation inspired the construction of various generalized simulated annealing algorithms based on Tsallis weight factors [31]-[34]. As an example we present here the generalized simulated annealing technique proposed in Ref. [34]. The configurations are weighted with

$$w(E) = [1 + (q-1)\,\beta\,(E - E_{GS})]^{-\frac{q}{q-1}} \ , \tag{9}$$

where E_{GS} is the ground-state energy and the Tsallis parameter q has been set to be: $q = 1 + \frac{1}{n_F}$. Here, n_F is the number of degrees of freedom. Note that through the substraction of E_{GS} it is ensured that the weights will always be positive definite. However, in general E_{GS} is not known. We therefore approximate E_{GS} in the course of a simulated annealing simulation by $E_0 \equiv E_{min} - c$, where E_{min} is the lowest energy ever encountered in the simulation and c a small number. E_0 is reset every time a new value for E_{min} is found. Changing the value of E_0 is a disturbance of the Markov chain and while we expect the disturbance to be small, we clearly cannot use our algorithm to calculate thermodynamic averages. Moreover, because of the finite step size of the temperature annealing we cannot assume convergence against an equilibrium distribution. As in the regular simulated annealing algorithm, our method is thus valid only as a global optimization method.

II.3 Generalized-Ensemble Algorithm with Tsallis Statistics

Generalized-ensemble algorithms are the methods that perform random walks in energy space, allowing simulations to escape from any state of energy local minimum. To name a few, multicanonical algorithms [2, 3], simulated tempering [5, 6], and *1/k*-sampling [7] are such

algorithms. Here, we discuss one of the latest examples of simulation techniques in generalized ensemble [13, 14]. The probability weight factor of this method is given by

$$w(E) = \left(1 + \beta_0 \frac{E - E_{GS}}{m}\right)^{-m} , \qquad (10)$$

where $T_0 = 1/k_B\beta_0$ is a low temperature, E_{GS} is the global-minimum potential energy, and $m(> 0)$ is a free parameter the optimal value of which will be given below. This is the Tsallis weight of Eq. (8) at a fixed temperature T_0 with the following choice of q:

$$q = 1 + \frac{1}{m} . \qquad (11)$$

The above choice of the weight was motivated by the following reasoning [13]. We are interested in an ensemble where not only the low-energy region can be sampled efficiently but also the high-energy states can be visited with finite probability. In this way the simulation can overcome energy barriers and escape from local minima. The probability distribution of energy should resemble that of an ideal low-temperature canonical distribution, but with a tail to higher energies. The Tsallis weight of Eq. (10) at low temperature T_0 has the required properties when the parameter m is carefully chosen. Namely, for suitable $m > 0$ it is a good approximation of the Boltzmann weight $w_B(E, T_0) = \exp(-\beta_0(E - E_{GS}))$ for $\beta_0(E - E_{GS})/m \ll 1$, while at high energies it is no longer exponentially suppressed but only according to a power law with the exponent m.

In this work we consider a system with continuous degrees of freedom. At low temperatures the harmonic approximation holds, and the density of states is given by

$$n(E) \propto (E - E_{GS})^{\frac{n_F}{2}} , \qquad (12)$$

where n_F is the number of degrees of freedom of the system under consideration. Hence, by Eqs. (10) and (12) the probability distribution of energy for the present ensemble is given by

$$P(E) \propto n(E)w(E) \propto (E - E_{GS})^{\frac{n_F}{2} - m} , \qquad (13)$$

for $\beta_0 \frac{E - E_{GS}}{m} \gg 1$. This implies that we need $m > \frac{n_F}{2}$. For, otherwise, the sampling of high-energy configurations will be enhanced too much. On the other hand, in the limit $m \to \infty$ our weight tends for all energies to the Boltzmann weight and high-energy configurations will not be sampled.

In order for low-temperature simulations to be able to escape from energy local minima, the weight should start deviating from the (exponentially damped) Boltzmann weight at the energy near its mean value (because at low temperatures there are only small fluctuations of energy around its mean). In Eq. (10) we may thus set

$$\beta_0 \frac{< E >_T - E_{GS}}{m} = \frac{1}{2} . \qquad (14)$$

The mean value at low temperatures is given by the harmonic approximation:

$$< E >_T = E_{GS} + \frac{n_F}{2} k_B T_0 = E_{GS} + \frac{n_F}{2\beta_0} . \qquad (15)$$

Substituting this value into Eq. (14), we obtain the optimal value for the exponent m:

$$m_{opt} = n_F . \qquad (16)$$

Hence, the optimal weight factor is given by

$$w(E) = \left(1 + \beta_0 \frac{E - E_0}{n_F}\right)^{-n_F} , \qquad (17)$$

where E_0 is the best estimate of the global-minimum energy E_{GS}.

We remark that the calculation of the weight factor is much easier than in other generalized-ensemble techniques, since it requires one to find only an estimator for the ground-state energy E_{GS}.

As in the case of other generalized ensembles, we can use the reweighting techniques [35] to construct canonical distributions at various temperatures T. This is because the simulation by the present algorithm samples a large range of energies. The thermodynamic average of any physical quantity $\mathcal{A}$ can be calculated over a wide temperature range by

$$< \mathcal{A} >_T = \frac{\int dx \; \mathcal{A}(x) \; w^{-1}(E(x)) \; e^{-\beta E(x)}}{\int dx \; w^{-1}(E(x)) \; e^{-\beta E(x)}} , \qquad (18)$$

where $w(E)$ is the weight in Eq. (17) and x stands for configurations.

Once the weight factor is given, we can implement the Metropolis MC algorithm [18] in a straightforward manner.

We now describe MD algorithm in the new ensemble defined by the weight of Eq. (17). We remark that similar stochastic dynamics algorithms were also developed in the context of Tsallis statistical mechanics in Refs. [36, 37].

The classical MD algorithm is based on a Hamiltonian

$$H(q,\pi) = \frac{1}{2}\sum_{i=1}^{N}\pi_i^2 + E(q_1,\cdots,q_N) \ , \qquad (19)$$

where π_i are the conjugate momenta corresponding to the coordinates q_i. Hamilton's equations of motion are then given by

$$\begin{cases} \dot{q}_i &= \dfrac{\partial H}{\partial \pi_i} = \pi_i \ , \\ \dot{\pi}_i &= -\dfrac{\partial H}{\partial q_i} = -\dfrac{\partial E}{\partial q_i} \ = \ f_i \ , \end{cases} \qquad (20)$$

and they are used to generate representative ensembles of configurations. For numerical work the time is discretized with a step Δt and the equations are integrated according to the *leapfrog* (or other time reversible integration) scheme:

$$\begin{cases} q_i(t+\Delta t) &= q_i(t) + \Delta t \ \pi_i\left(t+\dfrac{\Delta t}{2}\right) \ , \\ \pi_i\left(t+\dfrac{3}{2}\Delta t\right) &= \pi_i\left(t+\dfrac{\Delta t}{2}\right) + \Delta t \ f_i(t+\Delta t) \ . \end{cases} \qquad (21)$$

The initial momenta $\{\pi_i(\frac{\Delta t}{2})\}$ for the iteration are prepared by

$$\pi_i\left(\frac{\Delta t}{2}\right) = \pi_i(0) + \frac{\Delta t}{2} f_i(0) \ , \qquad (22)$$

with appropriately chosen $q_i(0)$ and $\pi_i(0)$ ($\pi_i(0)$ is from a Gaussian distribution).

In order to generalize this widely used technique to simulations in our generalized ensemble, we rewrite the weight factor in Eq. (17) as

$$w(E) = \exp\left\{-\beta_0\left[\frac{n_F}{\beta_0}\ln\left(1+\beta_0\frac{E-E_{GS}}{n_F}\right)\right]\right\} , \qquad (23)$$

We then define an effective potential energy by [36, 37]

$$E_{eff}(E) = \frac{n_F}{\beta_0}\ln\left(1+\beta_0\frac{E-E_{GS}}{n_F}\right) \ . \qquad (24)$$

We can see that MD simulations in the new ensemble can be performed by replacing E by E_{eff} (of Eq. (24)) in Eq. (20). A new set of Hamilton's equations of motion are now given by

$$\begin{cases} \dot{q}_i &= \pi_i \ , \\ \dot{\pi}_i &= -\dfrac{\partial E_{eff}}{\partial q_i} = -\dfrac{\partial E_{eff}}{\partial E}\dfrac{\partial E}{\partial q_i} = \dfrac{1}{1+\dfrac{\beta_0}{n_F}(E-E_{GS})}\ f_i \ . \end{cases} \qquad (25)$$

This is the set of equations we adopt for MD simulations in our new ensemble [14]. For numerical work the time is again discretized with a step Δt and the equations are integrated according to the *leapfrog* scheme.

Langevin algorithm [38] and hybrid Monte Carlo algorithm [39] in the new ensemble can likewise be introduced. It was shown that the performances of these three stochastic dynamics algorithms and that of MC version are similar (for details, see Ref. [14]).

III. Results

III.1 Simulated Annealing Algorithm

The effectiveness of the algorithms presented in the previous section is tested for the system of an oligopeptide, Met-enkephalin. This peptide has the amino-acid sequence Tyr-Gly-Gly-Phe-Met.

We have compared the performance of Tsallis simulated annealing algorithm with that of regular sim-

ulated annealing method [34]. As in an earlier work on Met-enkephalin [20] we made 20 runs of 50,000 MC sweeps for various annealing schedules. Each run started from completely random conformations. One of the quantities we monitored to evaluate the performance was the average $< E_{min} >$ (taken over all 20 runs) of the lowest energies E_{min} obtained in each single run. The other quantity was the number n_{GS} of ground-state conformations found in the 20 independent runs. In Ref. [40] it was shown that with the energy function KONF90, conformations of energy less than -11.0 kcal/mol have essentially the same three-dimensional structure. Hence, we consider any conformation with $E \leq -11.0$ kcal/mol as the ground-state conformation.

In Table 1 we show the results for our implementation of Tsallis weight in simulated annealing algorithms using the acceptance probability of Eq. (9). E_0 is reset every time to $E_0 = E_{min} - 1$ kcal/mol when a new conformation with lower energy E_{min} than any previous conformation is found. We found for both canonical and generalized simulated annealing simulations an optimal performance for the initial temperature $T_I = 500$ K and final temperature $T_F = 50$ K. With this annealing schedule the ground-state conformation was found 8 out of 20 runs for regular simulated annealing and 12 out of 20 runs for generalized simulated annealing. This is a modest improvement of the new algorithm over the canonical simulated annealing. The improvement can also be seen in the estimate for $< E_{min} >$ which is 0.6 kcal/mol lower for the new algorithm and has a smaller standard deviation than regular simulated annealing.

Table 1. Number of times that reached the ground state (n_{GS}) and average of the lowest energy ($< E_{min} >$) (in kcal/mol) obtained by the 20 runs of various regular and Tsallis simulated annealing simulations.

T_I (K)	T_F (K)	Regular Simulated Annealing		Tsallis Simulated Annealing	
		n_G	$< E_{min} >$	n_G	$< E_{min} >$
1000	50	6	-10.0 (1.3)	7	-10.7 (0.9)
1000	1	8	-10.0 (2.2)	7	-10.7 (1.3)
500	50	8	-10.5 (1.3)	12	-11.1 (0.9))
500	1	2	-9.3 (1.3)	11	-10.9 (1.3))
300	50	5	-10.1 (1.3)	13	-11.0 (0.9)
300	1	3	-9.6 (1.4)	11	-11.0 (1.1)

Moreover, we notice that the new simulated annealing algorithm allows one to start simulations at lower temperatures. While regular simulated annealing works best with initial temperatures over 500 K, the performance of the new algorithm depends only little on the initial temperature and rather favors $T_I \leq 500$ K. This follows from the form of the Tsallis distributions which have a tail to high energies for $q > 1$. Equilibrization at lower temperatures is therefore enhanced.

III.2 Generalized-Ensemble Algorithm

In this subsection we present the results of our generalized-ensemble simulations based on Tsallis statistics [13, 41]. It is known from our previous work that the global-minimum value of KONF90 energy for Met-enkephalin is $E_{GS} = -12.2$ kcal/mol [20]. The peptide has essentially a unique three-dimensional structure at temperatures $T \leq 50$ K, and the average energy is about -11 kcal/mol at $T = 50$ K [40, 4]. Hence, in the present work we always set $T = 50$ K (or, $\beta = 10.1$ $[\frac{1}{\mathrm{kcal/mol}}]$) in our new probability weight factor. All simulations were started from completely random initial configurations (Hot Start).

To demonstrate that thermalization is greatly enhanced in our ensemble, we first compare the "time series" of energy as a function of MC sweep. In Fig. 1 we

show the results from a regular canonical MC simulation at temperature $T = 50$ K (dotted curve) and those from a generalized-ensemble simulation of the new algorithm (solid curve). Here, the weight we used for the latter simulation is given by Eq. (17) with $n_F = 19$ and $E_0 = E_{GS} = -12.2$ kcal/mol. For the canonical run the curve stays around the value $E = -7$ kcal/mol with small thermal fluctuations, reflecting the low-temperature nature. The run has apparently been trapped in a local minimum, since the mean energy at this temperature is $< E >_T = -11.1$ kcal/mol as found by a multicanonical simulation in Ref. [20]. On the other hand, the simulation based on the new weight covers a much wider energy range than the canonical run. It is a random walk in energy space, which keeps the simulation from getting trapped in a local minimum. It indeed visits the ground-state region several times in 1,000,000 MC sweeps. These properties are common features of generalized-ensemble methods.

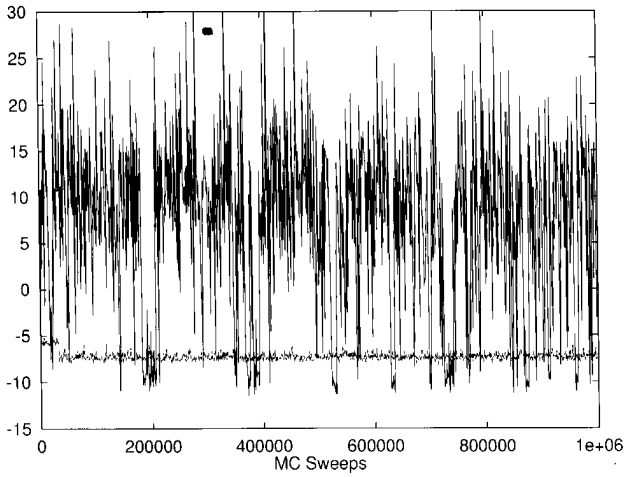

Figure 1. Time series of the total energy E_{tot} (kcal/mol) from a regular canonical simulation at temperature $T = 50$ K (dotted curve) and that from a simulation of the present method with the parameters: $E_0 = -12.2$ kcal/mol, $n_F = 19$, and $T = 50$ K (solid curve).

We now examine the dependence of the simulations on the values of the exponent m in our weight (see Eqs. (10) and (17)) and demonstrate that $m = n_F$ is indeed the optimal choice. Setting $E_0 = E_{GS} = -12.2$ kcal/mol, we performed 10 independent simulation runs of 50,000 MC sweeps with various choices of m. In Table 2 we list the lowest energies obtained during each of the 10 runs for five choices of m values: 9.5 ($= \frac{n_F}{2}$), 14, 19 ($= n_F$), 50, and 100. The results from regular canonical simulations at $T = 50$ K with 50,000 MC sweeps are also listed in the Table for comparison. If m is chosen to be too small (e.g., $m = 9.5$), then the weight follows a power law in which the suppression for higher energy region is insufficient (see Eq. (13)). As a result, the simulations tend to stay at high energies and fail to sample low-energy configurations. On the other hand, for too large a value of m (e.g., $m = 100$), the weight is too close to the canonical weight, and therefore the simulations will get trapped in local minima. It is clear from the Table that $m = n_F$ is the optimal choice. In this case the simulations found the ground-state configurations 80 % of the time (8 runs out of 10 runs). This should be compared with 90 % and 40 % for multicanonical annealing and simulated annealing algorithms, respectively, in simulations with the same number of MC sweeps [20].

The weight factor of the present algorithm just depends on the knowledge of the global-minimum energy E_{GS} (see Eq. (17)). If its value is known, which is the case for some systems, the weight is completely determined. However, if E_{GS} is not known, we have to obtain its best estimate E_0. In Table 3 we list the lowest energies obtained during each of 10 independent simulation runs of 200,000 MC sweeps with $m = n_F = 19$. Four choices were considered for the E_0 value: -12.2, -13.2, -14.2, and -15.2 kcal/mol. We remark that E_0 has to underestimate E_{GS} to ensure that $E - E_0$ cannot become negative. Our data show that an accuracy of $1 \sim 2$ kcal/mol in the estimate of the global-minimum energy is required for Met-enkephalin.

Since the simulation by the present algorithm samples a large range of energies (see Fig. 1), we can use the reweighting techniques [35] to construct canonical distributions and calculate thermodynamic quantities as a function of temperature over a wide temperature range.

All thermodynamic quantities were then calculated from a single production run of 1,000,000 MC sweeps which followed 10,000 sweeps for thermalization. At the end of every fourth sweep we stored the energies of the conformation, the corresponding volume, and the overlap of the conformation with the (known) ground state for further analyses. Here, we approximate the volume of the peptide by its solvent excluded volume (in Å^3)

which is calculated by a variant [42] of the double cubic lattice method [43]. Our definition of the overlap, which measures how much a given conformation differs from the ground state, is given by

$$O(t) = 1 - \frac{1}{90\ n_F} \sum_{i=1}^{n_F} |\alpha_i^{(t)} - \alpha_i^{(GS)}| \ , \qquad (26)$$

where $\alpha_i^{(t)}$ and $\alpha_i^{(GS)}$ (in degrees) stand for the n_F dihedral angles of the conformation at t-th MC sweep and the ground-state conformation, respectively. Symmetries for the side-chain angles were taken into account and the difference $\alpha_i^{(t)} - \alpha_i^{(GS)}$ was always projected into the interval $[-180°, 180°]$. Our definition guarantees that we have

$$0 \leq < O >_T \leq 1 \ , \qquad (27)$$

with the limiting values

$$\begin{cases} < O(t) >_T \ \rightarrow 1 \ , & T \rightarrow 0 \ , \\ < O(t) >_T \ \rightarrow 0 \ , & T \rightarrow \infty \ . \end{cases} \qquad (28)$$

We expect the folding of proteins and peptides to occur in a multi-stage process. A common scenario for folding may be that first the polypeptide chain collapses from a random coil to a compact state. This coil-to-globular transition is characterized by the collapse transition temperature T_θ. In the second stage, a set of compact structures are explored. The final stage involves a transition from one of the many local minima in the set of compact structures to the native (ground-state) conformation. This final transition is characterized by the folding temperature T_f ($\leq T_\theta$).

The first process is connected with a collapse of the extended coil structure into an ensemble of compact structures. This transition should be connected with a pronounced change in the average potential energy as a function of temperature. At the transition temperature we therefore expect a peak in the specific heat. Both quantities are shown in Fig. 2. We clearly observe a steep decrease in total potential energy around 300 K and the corresponding peak in the specific heat defined by

$$C \equiv \frac{1}{N\ k_B} \frac{d\,(< E_{tot} >_T)}{dT} = \beta^2\ \frac{< E_{tot}^2 >_T - < E_{tot} >_T^2}{N} \ , \qquad (29)$$

where N (= 5) is the number of amino-acid residues in the peptide. In Fig. 3 we display the average values of each of the component terms of the potential energy (defined in Eq. (2)) as a function of temperature. As one can see in the Figure, the change in average potential energy is mainly caused by the Lennard-Jones term and therefore is connected to a decrease of the volume occupied by the peptide. This can be seen in Fig. 4, where we display the average volume as a function of temperature. As expected, the volume decreases rapidly in the same temperature range as the potential energy.

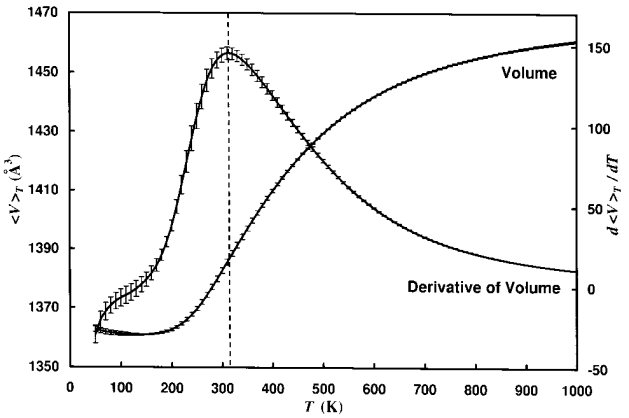

Figure 2. Average total potential energy $< E_{tot} >_T$ and specific heat C as a function of temperature. The dotted vertical line is added to aid the eyes in locating the peak of specific heat. The results were obtained from a generalized-ensemble simulation of 1,000,000 MC sweeps.

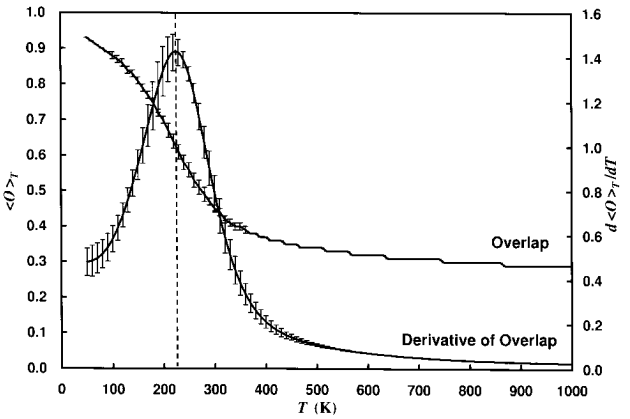

Figure 3. Average potential energies as a function of temperature. The results were obtained from a generalized-ensemble simulation of 1,000,000 MC sweeps.

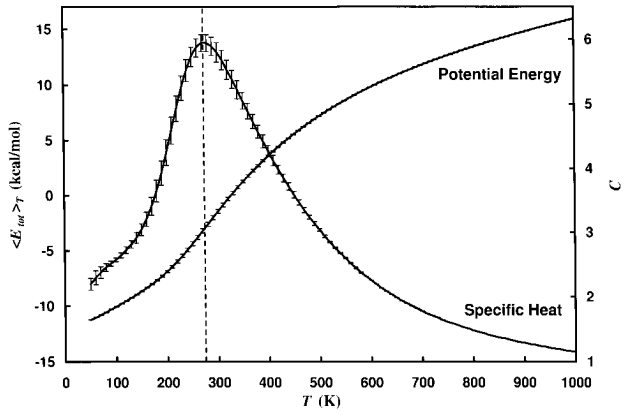

Figure 4. Average volume $< V >_T$ and its derivative $d < V >_T /dT$ as a function of temperature. The dotted vertical line is added to aid the eyes in locating the peak of the derivative of volume. The results were obtained from a generalized-ensemble simulation of 1,000,000 MC sweeps.

If both energy and volume decrease are correlated, then the transition temperature T_θ can be located both from the position where the specific heat has its maximum and from the position of the maximum of

$$\frac{d < V >_T}{dT} \equiv \beta^2 \left(< V E_{tot} >_T - < V >_T < E_{tot} >_T \right) , \quad (30)$$

which is also displayed in Fig. 4. The second quantity measures the steepness of the decrease in volume in the same way as the specific heat measures the steepness of decrease of potential energy. To quantify its value we divided our time series in 4 bins corresponding to 250,000 sweeps each, determined the position of the maximum for both quantities in each bin and averaged over the bins. In this way we found a transition temperature $T_\theta = 280 \pm 20$ K from the location of the peak in specific heat and $T_\theta = 310 \pm 20$ K from the maximum in $d < V >_T /dT$. Both temperatures are indeed consistent with each other within the error bars.

Table 2. Lowest energy (in kcal/mol) obtained by the present method with several different choices of the exponent m. The case for $m = \infty$ stands for a regular canonical run at $T = 50$ K. $< E_{min} >$ is the average of the lowest energy obtained by the 10 runs (with the standard deviations in parentheses), and n_{GS} is the number of runs in which a conformation with $E \leq -11.0$ kcal/mol (the average energy at $T = 50$ K) was obtained.

E_0	$E_{GS} = -12.2$	-12.2	-12.2	-12.2	-12.2	
m	$\frac{n_F}{2} = 9.5$	14	$n_F = 19$	50	100	∞
Run						
1	0.8	-5.2	-11.8	-6.9	-6.8	-4.2
2	-1.4	-2.6	-11.5	-7.1	-7.7	-5.2
3	0.1	-6.8	-11.5	-6.9	-4.9	-11.8
4	0.5	-5.5	-11.7	-8.2	-9.9	-7.1
5	-1.0	-3.4	-11.6	-7.4	-12.0	-3.3
6	1.1	-6.4	-11.6	-10.1	-8.8	0.9
7	-1.3	-5.1	-8.5	-8.7	-8.7	-5.3
8	0.4	-3.3	-9.7	-10.8	-9.5	-6.3
9	1.2	-8.1	-11.6	-12.0	-6.8	-6.4
10	1.2	-3.3	-11.9	-10.8	-9.5	-4.7
$< E_{min} >$	0.2 (1.0)	-5.0 (1.8)	-11.1 (1.1)	-8.9 (1.9)	-8.5 (2.0)	-5.3 (3.2)
n_{GS}	0/10	0/10	8/10	1/10	1/10	1/10

Table 3. Lowest energy (in kcal/mol) obtained by the present method with several different choices of the free parameter E_0. $< E_{min} >$ and n_{GS} are the same as in Table 2.

E_0	$E_{GS} = -12.2$	-13.2	-14.2	-15.2
m	$n_F = 19$	19	19	19
Run				
1	-11.8	-11.1	-10.5	-9.0
2	-11.9	-10.8	-8.3	-10.3
3	-11.9	-11.3	-11.6	-9.7
4	-11.9	-10.2	-10.9	-10.8
5	-11.8	-11.2	-6.9	-9.2
6	-11.3	-11.5	-10.8	-9.6
7	-11.9	-11.3	-8.3	-10.3
8	-11.8	-11.4	-5.9	-6.8
9	-12.0	-11.5	-10.6	-8.6
10	-11.7	-10.0	-10.3	-8.9
$< E_{min} >$	-11.8 (0.2)	-11.0 (0.5)	-9.4 (1.9)	-9.3 (1.1)
n_{GS}	10/10	7/10	1/10	0/10

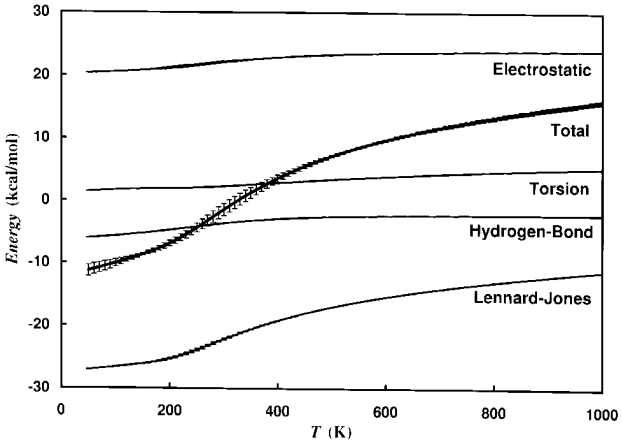

Figure 5. Average overlap $< O >_T$ and its derivative $d < O >_T /dT$ as a function of temperature. The dotted vertical line is added to aid the eyes in locating the peak of the derivative of overlap. The results were obtained from a generalized-ensemble simulation of 1,000,000 MC sweeps.

The second transition which should occur at a lower temperature T_f is that from a set of compact structures to the "native conformation" that is considered to be the ground state of the peptide. Since these compact conformations are expected to be all of similar volume and energy, we do not expect to see this transition by pronounced changes in $< E_{tot} >_T$ or to find another peak in the specific heat. Instead this transition should be characterized by a rapid change in the average overlap $< O >_T$ with the ground-state conformation (see Eq. (26)) and a corresponding maximum in

$$\frac{d < O >_T}{dT} \equiv \beta^2 \left(< OE_{tot} >_T - < O >_T < E_{tot} >_T\right) . \tag{31}$$

Both quantities are displayed in Fig. 5, and we indeed find the expected behavior. The change in the order parameter is clearly visible and occurs at a temperature lower than the first transition temperature T_θ. We again try to determine its value by searching for the peak in $d < O >_T /dT$ in each of the 4 bins and averaging over the obtained values. In this way we find a transition temperature of $T_f = 230 \pm 30$ K. We remark that the average overlap $< O >_T$ approaches its limiting value zero only very slowly as the temperature increases. This is because $< O >_T = 0$ corresponds to a completely random distribution of dihedral angles which is energetically highly unfavorable because of the

steric hindrance of both main and side chains.

We remark that the above algorithm was also succesfully applied to a more direct evaluation of the free-energy landscape of small peptides [44], which allowed us to visualize the folding funnel of the molecule.

IV. Conclusions

In this article we have reviewed the uses of Tsallis statistical mechanics in the protein folding problem. Monte Carlo simulated annealing algorithm and generalized-ensemble algorithm with both Monte Carlo and stochastic dynamics algorithms were introduced. A penta peptide, Met-enkephalin was used to study the performances of these algorithms. While other generalized-ensemble algorithms suffer from the difficulty that the determination of the optimal weight factor is non-trivial and tedious, it was shown that its determination in the Tsallis generalized-ensemble algorithm is much simpler than other versions.

Acknowledgements:
Our simulations were performed on the computers of the Computer Center at the Institute for Molecular Science, Okazaki, Japan. This work is supported, in part, by a Grant-in-Aid for Scientific Research from the Japanese Ministry of Education, Science, Sports and Culture, by a grant from the Research for the Future Program of Japan Society for the Promotion of Science (JSPS-RFTF98P01101) and by a Research Excellence Fund (E27448) of the State of Michigan.

References

[1] S. Kirkpatrick, C.D. Gelatt, Jr., and M.P. Vecchi, *Science* **220**, 671 (1983).

[2] B.A. Berg and T. Neuhaus, *Phys. Lett.* **B267**, 249 (1991); *Phys. Rev. Lett.* **68**, 9 (1992).

[3] B.A. Berg, *Int. J. Mod. Phys.* **C3**, 1083 (1992).

[4] U.H.E. Hansmann and Y. Okamoto, *J. Comp. Chem.* **14**, 1333 (1993).

[5] A.P. Lyubartsev, A.A.Martinovski, S.V. Shevkunov, and P.N. Vorontsov-Velyaminov, *J. Chem. Phys.* **96**, 1776 (1992).

[6] E. Marinari and G. Parisi, *Europhys. Lett.* **19**, 451 (1992).

[7] B. Hesselbo and R.B. Stinchcombe, *Phys. Rev. Lett.* **74**, 2151 (1995).

[8] U.H.E. Hansmann and Y. Okamoto, *J. Comp.Chem.* **18**, 920 (1997).

[9] U.H.E. Hansmann, Y. Okamoto and F. Eisenmenger, *Chem. Phys. Lett.* **259**, 321 (1996).

[10] N. Nakajima, H. Nakamura and A. Kidera, *J. Phys. Chem.* **101**, 817 (1997).

[11] C. Bartels and M. Karplus, *J. Phys. Chem. B* **102**, 865 (1998).

[12] C. Tsallis, *J. Stat. Phys.* **52**, 479 (1988).

[13] U.H.E. Hansmann and Y. Okamoto, *Phys. Rev. E* **56**, 2228 (1997).

[14] U.H.E. Hansmann, F. Eisenmenger, and Y. Okamoto, *Chem. Phys. Lett.* **297**, 374 (1998).

[15] F.A. Momany, R.F. McGuire, A.W. Burgess, and H.A. Scheraga, *J. Phys. Chem.* **79**, 2361 (1975); G. Némethy, M.S. Pottle, and H.A. Scheraga, *J. Phys. Chem.* **87**, 1883 (1983); M.J. Sippl, G. Némethy, and H.A. Scheraga, *J. Phys. Chem.* **88**, 6231 (1984).

[16] H. Kawai, Y. Okamoto, M. Fukugita, T. Nakazawa, and T. Kikuchi, *Chem. Lett.* **1991**, 213 (1991); Y. Okamoto, M. Fukugita, T. Nakazawa, and H. Kawai, *Protein Eng.* **4**, 639 (1991).

[17] The program SMC was written by F. Eisenmenger.

[18] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, *J. Chem. Phys.* **21**, 1087 (1953).

[19] S. Geman and D. Geman, *IEEE Trans. Patt Anal. Mach. Intel* **6**, 721 (1984).

[20] U.H.E. Hansmann and Y. Okamoto, *J. Phys. Soc. Jpn.* **63**, 3945 (1994); *Physica A* **212**, 415 (1994).

[21] S.R. Wilson, W. Cui, J.W. Moskowitz, and K.E. Schmidt, *Tetrahedron Lett.* **29**, 4373 (1988).

[22] H. Kawai, T. Kikuchi, and Y. Okamoto, *Protein Eng.* **3**, 85 (1989).

[23] C. Wilson and S. Doniach, *Proteins* **6**, 193 (1989).

[24] A.T. Brünger, *J. Mol. Biol.* **203**, 803 (1988).

[25] M. Nilges, G.M. Clore, and A.M. Gronenborn, *FEBS Lett.* **229**, 317 (1988).

[26] M. Pellegrini, N. Grønbech-Jensen, and S. Doniach, *Physica A* **239**, 244 (1997).

[27] M. Kinoshita, Y. Okamoto, F. Hirata, *J. Am. Chem. Soc.* **120**, 1855 (1998).

[28] L. Carlacci, *J. Comp-Aided Mol. Des.* **12**, 195 (1998).

[29] Y. Okamoto, M. Masuya, M. Nabeshima and T. Nakazawa, *Chem. Phys. Lett.* **299**, 17 (1999).

[30] H. Szu and R. Hartley, *Phys. Lett.* **A122**, 157 (1987).

[31] D.A. Stariolo and C. Tsallis, in *Annual Reviews of Computational Physics II*, edited by D. Stauffer (World Scientific, Singapore, 1995), p. 343.

[32] T.J.P. Penna, *Phys. Rev. E* **51**, R1 (1995).

[33] I. Andricioaei and J.E.Straub, *Phys. Rev. E* **53**, R3055 (1996).

[34] U.H.E. Hansmann, *Physica A* **242**, 250 (1997).

[35] A.M. Ferrenberg and R.H. Swendsen, *Phys. Rev. Lett.* **61**, 2635 (1988); *ibid.* **63**, 1658(E) (1989).

[36] D.A. Stariolo, *Phys. Lett.* **A185**, 262 (1994).

[37] I. Andricioaei and J.E. Straub, *J. Chem. Phys.* **107**, 9117 (1997).

[38] G. Parisi and Y.-S. Wu, *Sci. Sin.* **24**, 483 (1981).

[39] S. Duane, A.D. Kennedy, B.J. Pendleton, and D. Roweth, *Phys. Lett.* **B195**, 216 (1987). references given in the erratum.

[40] Y. Okamoto, T. Kikuchi, and H. Kawai, *Chem. Lett.* **1992**, 1275 (1992).

[41] U.H.E. Hansmann, M. Masuya, and Y. Okamoto, *Proc. Natl. Acad. Sci. USA* **94**, 10652 (1997).

[42] M. Masuya, in preparation.

[43] F. Eisenhaber, P. Lijnzaad, P. Argos, C. Sander, and M. Scharf, *J. Comp. Chem.* **16**, 273 (1995).

[44] U.H.E. Hansmann, Y. Okamoto, and J.N. Onuchic, "The folding funnel landscape for the peptide Met-enkephalin," *Proteins: Structure, Function, and Genetics* (1999), in press.