Protein folding simulations in a deformed energy landscape

U.H.E. Hansmann[a]

Department of Physics Michigan Technological University Houghton, MI 49931-1295, USA

Abstract. A modified version of stochastic tunneling, a recently introduced global optimization technique, is introduced as a new generalized-ensemble technique and tested for a benchmark peptide, Met-enkephalin. It is demonstrated that the new technique allows to evaluate folding properties and especially the glass temperature T_{g} of this peptide.

PACS. 05.10.-a Computational methods in statistical physics and nonlinear dynamics – 87.16.Ac Theory and modeling; computer simulation

Numerical simulations of biological molecules can be extremely difficult when the molecule is described by "realistic" energy functions where interactions between all atoms are taken into account. For a large class of molecules, in particular for peptides or proteins, the various competing interactions lead to frustration and a rough energy landscape. At low temperatures canonical simulations will get trapped in one of the multitude of local minima separated by high energy barriers and physical quantities cannot be calculated accurately. One way to overcome this difficulty in protein simulations is by utilizing so-called *generalized ensembles* [1], which are based on a non-Boltzmann probability distribution. Multicanonical sampling [2] and simulated tempering [3] are prominent examples of such an approach. Application of these techniques to the protein folding problem was first addressed in reference [4] and their usefulness for simulation of biological molecules and other complex systems [4–8] has become increasingly recognized.

However, generalized-ensemble methods are not without problems. In contrast to canonical simulations the weight factors are not *a priori* known. Hence, for a computer experiment one needs estimators of the weights, and the problem of finding good estimators is often limiting the use of generalized-ensemble techniques. Here we describe and test a new generalized ensemble where determination of the weights is by construction of the ensemble simple and straightforward. Our method is based on a recently introduced global optimization technique, stochastic tunneling [9].

Canonical simulations of proteins at low temperature are hampered by the roughness of the potential energy surface: local minima are separated by high energy barriers. To enhance sampling we propose to weight conformations not with the Boltzmann factor $w_{\mathrm{B}}(E) = \exp(-E/k_{\mathrm{B}}T)$,

but with a weight

$$w_{\mathrm{f}}(E) = \exp(f(E)/k_{\mathrm{B}}T). \qquad (1)$$

Here, T is a low temperature, k_{B} the Boltzmann constant, and $f(E)$ is a non-linear transformation of the potential energy onto the interval $[0, -1]$ chosen such that the relative location of all minima is preserved. The physical idea behind such an approach is to allow the system at a given low temperature T to "*tunnel*" through energy barriers of arbitrary height. A transformation with the above characteristics can be realized by

$$f_1(E) = -\mathrm{e}^{-(E-E_0)/n_{\mathrm{F}}}. \qquad (2)$$

Here, E_0 is an estimate for the ground state and n_{F} the number of degrees of freedom of the system. Equation (2) is a special choice of the transformation recently introduced under the name "stochastic tunneling" [9] to the corresponding problem of global minimization in complex potential energy landscapes. One can easily find further examples for transformations with the above stated properties, for instance,

$$f_2(E) = -(1 + (E - E_0)/n_{\mathrm{F}})^{-1}. \qquad (3)$$

We will restrict our investigation to these two transformations without claiming that they are an optimal choice.

A simulation in the above ensemble, defined by the weight of equation (1) with a suitable chosen non-linear transformation $f(E)$, will sample a broad range of energies. Hence, application of re-weighting technique [10] allows to calculate the expectation value of any physical quantity $\mathcal{O}$ over a large range of temperatures T by

$$\langle \mathcal{O} \rangle_T = \frac{\int \mathrm{d}E \mathcal{O}(E) P_{\mathrm{f}}(E) w_{\mathrm{f}}^{-1}(E) \mathrm{e}^{-E/k_{\mathrm{B}}T}}{\int \mathrm{d}E P_{\mathrm{f}}(E) w_{\mathrm{f}}^{-1}(E) \mathrm{e}^{-E/k_{\mathrm{B}}T}}. \qquad (4)$$

[a] e-mail: `hansmann@mtu.edu`

In this point our method is similar to other generalized-ensemble techniques such as the multicanonical sampling [2], however, our method differs from them in that the weights are explicitly given by equation (1). One only needs to find an estimator for the ground-state energy E_0 in the transforming functions $f_1(E)$ or $f_2(E)$ (see Eqs. (2, 3)) which in earlier work [11,12] was found to be much easier than the determination of weights for multicanonical algorithm [2] or simulated tempering [3].

The new simulation technique was tested for Met-enkephalin, one of the simplest peptides, which has become a often used model to examine new algorithms. Met-enkephalin has the amino-acid sequence Tyr-Gly-Gly-Phe-Met. The potential energy function E_{tot} that was used is given by the sum of the electrostatic term E_{es}, 12-6 Lennard-Jones term E_{vdW}, and hydrogen-bond term E_{hb} for all pairs of atoms in the peptide together with the torsion term E_{tors} for all torsion angles:

$$E_{\text{tot}} = E_{\text{es}} + E_{\text{vdW}} + E_{\text{hb}} + E_{\text{tors}}, \tag{5}$$

$$E_{\text{es}} = \sum_{(i,j)} \frac{332 q_i q_j}{\epsilon r_{ij}}, \tag{6}$$

$$E_{\text{vdW}} = \sum_{(i,j)} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right), \tag{7}$$

$$E_{\text{hb}} = \sum_{(i,j)} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right), \tag{8}$$

$$E_{\text{tors}} = \sum_l U_l \left(1 \pm \cos(n_l \chi_l) \right), \tag{9}$$

where r_{ij} is the distance between the atoms i and j, and χ_l is the lth torsion angle. The parameters $(q_i, A_{ij}, B_{ij}, C_{ij}, D_{ij}, U_l$ and $n_l)$ for the energy function were adopted from ECEPP/2 [13].

The computer code SMC[1] was used. The simulations were started from completely random initial conformations (Hot Start) and one Monte-Carlo sweep updates every torsion angle of the peptide once. The peptide bond angles ω were fixed to their common value 180°, which left 19 torsion angles (ϕ, ψ, and χ) as independent degrees of freedom (i.e., $n_F = 19$). The interaction of the peptide with the solvent was neglected in the simulations and the dielectric constant ϵ set equal to 2. In short preliminary runs it was found that $T = 8$ K was the optimal temperatures for simulations relying on the transformation $f_1(E)$ (Eq. (2)), and $T = 6$ K for simulations relying on the second chosen transformation $f_2(E)$ (Eq. (3)). The free parameter E_0 was set in equation (2) or (3) to $E_0 = -10.72$ kcal/mol, the ground state energy as known from previous work. In addition, simulations were also performed where E_0 was dynamically updated in the course of the simulation and set to the lowest ever encountered energy. In these runs the (known) ground state was found in less than 5000 MC sweeps. Hence, determination of the weights is easier than in other generalized-ensemble

techniques since in earlier work [4] it was found that at least 40 000 sweeps were needed to calculate multicanonical weights. We remark that a Monte-Carlo sweep in both algorithm takes approximately the same amount of CPU time.

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 sweep we stored the energies of the conformation and the radius of gyration

$$R = \frac{1}{N_{\text{atoms}}^2} \sum_{i,j}^{N_{\text{atoms}}} (\mathbf{r}_i - \mathbf{r}_j)^2 \tag{10}$$

for further analyses.

In order to demonstrate the dynamical behavior of the algorithm the "time series" and histograms of potential energy are shown for both choices of the transforming functions $f_1(E)$ (Fig. 1) and $f_2(E)$ (Fig. 2). Both choices of the non-linear transformation with which the energy landscape was deformed in the simulations lead to qualitatively the same picture. In Figures 1a and 2a, respectively, one can see that the whole energy range between $E < -10$ kcal/mol (the ground state region) and $E \approx 20$ kcal/mol (high-energy, coil states) is sampled. However, unlike in the multicanonical algorithm the energies are not sampled uniformly and low-energy states appear with higher frequency than high energy states. However, as one can see from the logarithmic scale of Figures 1b and 2b, where the histograms are displayed for these simulations, high-energy states are only suppressed by three orders of magnitude and their probability is still large enough to allow crossing of energy barriers. Hence large parts of the configuration space are sampled by our method and it is justified to calculate from these simulations thermodynamic quantities by means of re-weighting, see equation (4).

Here, the average radius of gyration $\langle R \rangle$, which is is a measure for the compactness of protein configurations and defined in equation (10), was calculate for various temperatures. In Figure 3 the results for the new ensemble, using the defining non-linear transformations $f_1(E)$ or $f_2(E)$, are compared with the ones of a multicanonical run with equal number of Monte-Carlo sweeps. As one can see, the values of $\langle R \rangle(T)$ agree for all three simulations over the whole temperature range. Hence, it is obvious that simulations in the new ensemble are indeed well able to calculate thermodynamic averages over a wide temperature range.

After having established the new techniques as a possible alternative to other generalized-ensemble techniques (such as multicanonical sampling or simulated tempering) its usefulness shall be further demonstrated by calculating the free energy of Met-enkephalin as a function of radius of gyration R:

$$G(R) = -k_{\text{B}} T \log P(R) \tag{11}$$

where

$$P(R) = P_{\text{f}}(R) w_{\text{f}}^{-1}(E(R)) e^{-E(R)/k_{\text{B}}T}. \tag{12}$$

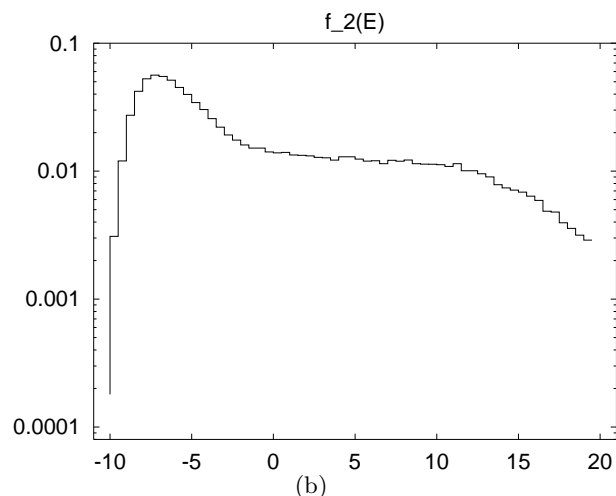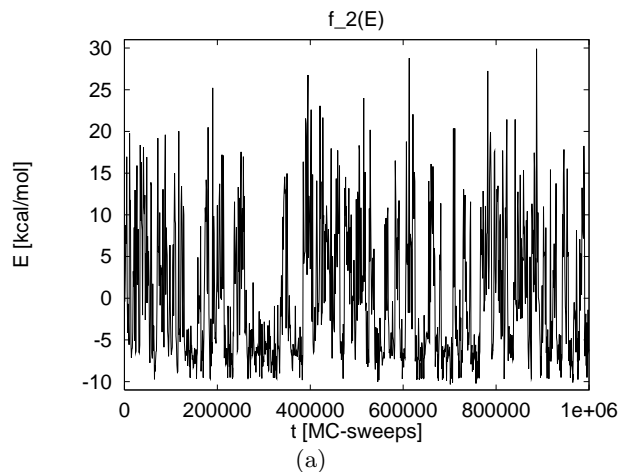[1] The program SMC was written by Dr. Frank Eisenmenger (eisenmenger@rz.hu-berlin.de).
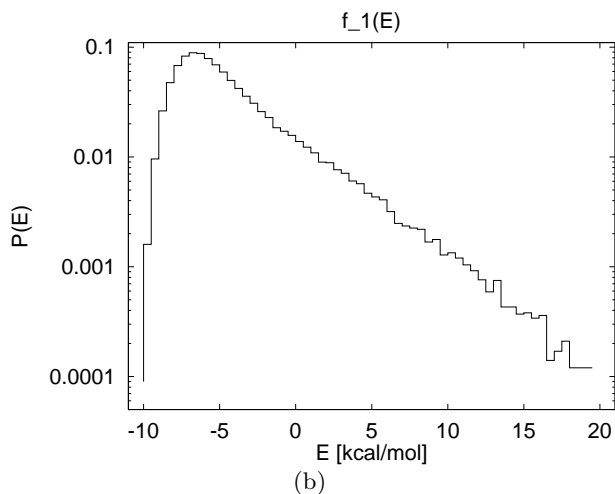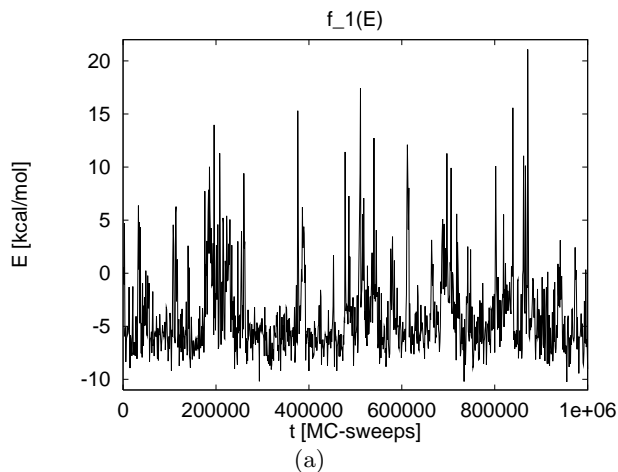
Fig. 1. "Time series" (a) of potential energy E of Met-enkephalin for a simulation in a generalized ensemble defined by the transformation $f_1(E)$ of equation (2) and the corresponding histogram (b) of potential energy.
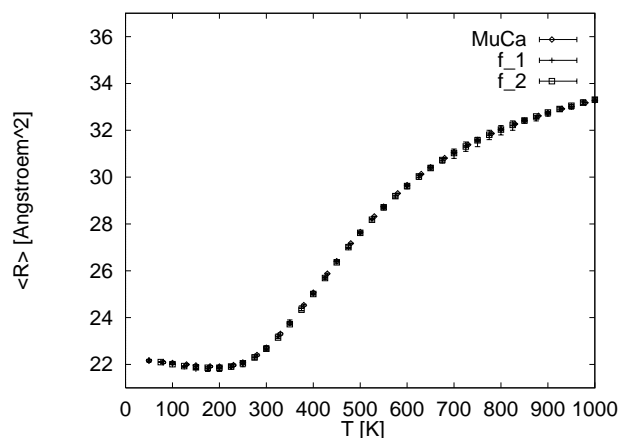

Fig. 2. "Time series" (a) of potential energy E of Met-enkephalin for a simulation in a generalized ensemble defined by the transformation $f_2(E)$ of equation (3) (a) and the corresponding histogram (b) of potential energy.

Here, a normalization is chosen where the minimal value of $G_{\min}(R) = 0$. The chosen temperature was $T = 230$ K, which was found in earlier work [7] as the folding temperature T_f of Met-enkephalin. The results, which rely on the transformation $f_1(E)$ of the energy landscape given by equation (2), are displayed in Figure 4. At this temperature one observes clearly a "funnel" towards low values of R which correspond to compact structures. Such a funnel-like landscape was already observed in reference [8] for Met-enkephalin, utilizing a different set of order parameters, and is predicted by the landscape theory of folding [14].

The essence of the funnel landscape idea is competition between the tendency towards the folded state and trapping due to ruggedness of the landscape. One way to measure this competition is by the ratio [15]:
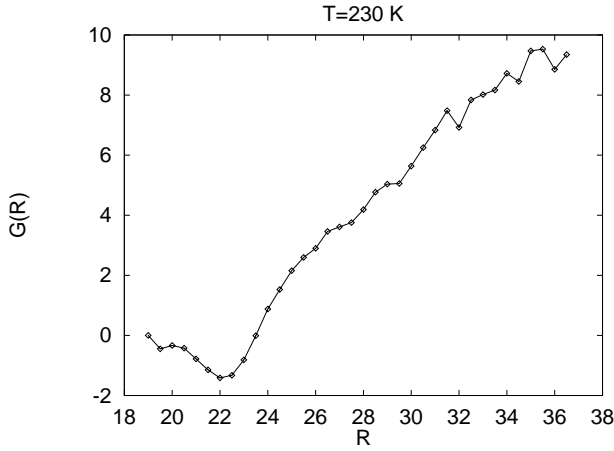
$$Q = \frac{\overline{E - E_0}}{\sqrt{\overline{E^2} - \bar{E}^2}}, \tag{13}$$

Fig. 3. Average radius of gyration $\langle R \rangle$ (in Å^2) as a function of temperature (in K). The results of a multicanonical simulation of 1 000 000 MC sweeps were compare with simulations of equal statistics in the new ensemble utilizing either the nolinear transformation $f_1(E)$ or $f_2(E)$.

Fig. 4. Free energy $G(R)$ as a function of the radius of gyration R for $T = 230$ K. The results rely on a generalized-ensemble simulation based on the transformation $f_1(E)$ of the energy landscape s defined in equation (2).

where the bar denotes averaging over compact configurations. The landscape theory asserts that good folding protein sequences are characterized by large values of Q [15]. Using the results of our simulations and defining a compact structure as one where $R(i) \leq 23$ Å, we find $\overline{E - E_0} = 13.96(3)$ kcal/mol, $\overline{E^2} - \bar{E}^2 = 0.49(2)$, from which we estimate for the above ratio $Q = 20.0(5)$. This value indicates that Met-enkephalin is good folder and is consistent with earlier work [7] where we evaluated an alternative characterization of folding properties. Thirumalai and collaborators [16] have conjectured that the kinetic accessibility of the native conformation can be classified by the parameter

$$\sigma = \frac{T_\theta - T_f}{T_\theta}, \tag{14}$$

i.e., the smaller σ is, the more easily a protein can fold. Here T_f is the folding temperature and T_θ the collapse temperature. With values for $T_\theta = 295$ K and $T_f = 230$ K, as measured in reference [7], one has for Met-enkephalin $\sigma \approx 0.2$, indicating again that the peptide has good folding properties.

Yet another characterization of folding properties relies on knowledge of the glass temperature T_g and is closely related to equation (13). As the number of available states gets reduced with the decrease of temperature, the possibility of local trapping increases substantially. Glassy behavior appears when the residence time in some local traps becomes of the order of the folding event. Folding dynamics is now non-exponential since different traps have different escape times [17]. For temperatures above the glass transition temperature T_g, the folding dynamics is exponential and a configurational diffusion coefficient average the effects of the short lived traps [18]. It is expected that for a good folder the glass transition temperature, T_g, where glass behavior sets in, has to be significantly lower than the folding temperature T_f, *i.e.* a good folder can be

characterized by the relation [19]

$$\frac{T_f}{T_g} > 1. \tag{15}$$

I present here for the first time a numerical estimate of this glass transition temperature for the peptide Met-enkephalin. The calculation of the estimate is based on the approximation [19]

$$T_g = \sqrt{\frac{\overline{E^2} - \bar{E}^2}{2k_B S_0}}, \tag{16}$$

where the bar indicates again averaging over compact structures and S_0 is the entropy of these states chosen such that the entropy of the ground state becomes zero. The results of the simulation in the new ensemble defined by the transformation $f_1(E)$, leads to a value of $s_0 = 2.3(7)$. Together with the above quoted value for $\overline{E^2} - \bar{E}^2 = 0.49(2)$ (in (kcal/mol)^2) one therefore finds as an estimate for the glass transition temperature

$$T_g = 160(30) \text{ K}. \tag{17}$$

Since it was found in earlier work [7] that $T_f = 230(30)$, it is obvious that the ratio $T_f/T_g > 1$ and again one finds that Met-enkephalin has good folding properties. Hence, we see that there is a strong correlation between all three folding criteria.

Let me summarize my results. I have proposed to utilize a recently introduced global optimization technique, stochastic tunneling, in such a way that it allows calculation of thermodynamic quantities. The new generalized-ensemble technique was tested for a benchmark peptide, Met-enkephalin. It was demonstrated that the new technique allows to evaluate the folding properties of this peptide and an estimate for the glass transition temperature T_g in that system was presented. Currently I am evaluating the efficiency of the new method in simulations of larger molecules.

References

1. U.H.E. Hansmann, Y. Okamoto, in *Annual Reviews in Computational Physics VI*, edited by D. Stauffer (Singapore, World Scientific, 1999), pp. 129-157.
2. B.A. Berg, T. Neuhaus, Phys. Lett. B **267**, 249 (1991); Phys. Rev. Lett. **68**, 9 (1992).
3. A.P. Lyubartsev, A.A. Martinovski, S.V. Shevkunov, P.N. Vorontsov-Velyaminov, J. Chem. Phys. **96**, 1776 (1992); E. Marinari, G. Parisi, Europhys. Lett. **19**, 451 (1992).

4. U.H.E. Hansmann, Y. Okamoto, J. Comp. Chem. **14**, 1333 (1993).
5. Y. Okamoto, U.H.E. Hansmann, J. Phys. Chem. **99**, 2236 (1995).
6. F. Eisenmenger, U.H.E. Hansmann, J. Phys. Chem. B **101**, 3304 (1997).
7. U.H.E. Hansmann, M. Masuya, Y. Okamoto, Proc. Natl. Acad. Sci. USA **94**, 10652 (1997).
8. U.H.E. Hansmann, Y. Okamoto, J.N. Onuchic, Proteins **34**, 472 (1999).
9. W. Wenzel, K. Hamacher, Phys. Rev. Lett. **82**, 3003 (1999).
10. A.M. Ferrenberg, R.H. Swendsen, Phys. Rev. Lett. **61**, 2635 (1988); Phys. Rev. Lett. **63**, 1658(E) (1989), and references given in the erratum.
11. U.H.E. Hansmann, Physica A **242**, 250 (1997).
12. U.H.E. Hansmann, Y. Okamoto, Phys. Rev. E **56**, 2228 (1997).
13. M.J. Sippl, G. Némethy, H.A. Scheraga, J. Phys. Chem. **88**, 6231 (1984), and references therein.
14. K.A. Dill, H.S. Chan, Nature Struct. Biol. **4**, 10 (1997).
15. R.A. Goldstein, Z.A. Luthey-Schulten, P.G. Wolynes, Proc. Natl. Acad. Sci. USA **89**, 4918 (1992).
16. D.K. Klimov, D. Thirumalai, Phys. Rev. Lett. **76**, 4070 (1996).
17. N.D. Socci, J.N. Onuchic, P.G. Wolynes, Proteins **32**, 136 (1998).
18. N.D. Socci, J.N. Onuchic, P.G. Wolynes, J. Chem. Phys. **104**, 5860 (1996).
19. J.D. Bryngelson, J.N. Onuchic, N.D. Socci, P.G. Wolynes, Proteins **21**, 167 (1995).