

# All-Atom Protein-Folding Simulations in Generalized-Ensembles

Nelson A. Alves,

*Departamento de Física e Matemática, FFCLRP, Universidade de São Paulo  
Av. Bandeirantes 3900. CEP 14040-901 Ribeirão Preto, SP, Brazil*

Yong Peng, and Ulrich H.E. Hansmann

*Department of Physics, Michigan Technological University, Houghton, MI 49931-1291, USA*

Received on 9 August, 2003

We review the generalized-ensemble approach to protein studies. Focusing on the problem of secondary structure formation, we show that these sophisticated techniques allow efficient simulations of all-atom protein models and may lead to a deeper understanding of the folding mechanism in proteins.

## 1 Introduction

Because a protein is only functional if it folds into its characteristic shape, it is important to understand how the structure and the function of proteins emerge from their sequence of amino acids (the monomers of the linear chain that builds up a protein). Such knowledge could not only lead to the *de novo* design of proteins that serve as novel drugs with customized properties, but also to a deeper understanding of various diseases that are caused by the misfolding of proteins.

Computer experiments offer one way to gain such knowledge but are extremely difficult for realistic protein models. This is because all-atom models of proteins lead to a rough energy landscape with a huge number of local minima separated by high energy barriers. Consequently, sampling of low-energy conformations becomes a hard computational task, and physical quantities cannot be calculated accurately from simple low-temperature molecular dynamics or Monte Carlo simulations.

The quest for overcoming this so-called multiple-minima problem is an active area of research (for a review, see Refs. [1, 2]). One of us suggested almost ten years ago that *generalized-ensemble* simulations may allow a better sampling of low-energy protein configurations [3]. Examples of this group of closely related techniques are multicanonical sampling [4], the broad histogram method [5], the Wang-Landau algorithm [6], techniques that rely on Tsallis weights [7, 8], or parallel tempering (also known as replica exchange method) [9], and over the last decade these techniques have been successfully applied to protein simulations (for a review, see Ref. [10]).

In the following we will present a short review of the generalized-ensemble approach and demonstrate its usefulness for protein simulations. We will focus in our examples on one particularly important aspect of the protein-folding problem, namely the role of secondary structure formation in the folding process.

## 2 Generalized-ensemble techniques

All generalized-ensemble techniques share the same key-idea: replace the canonical weights, that suppress the crossing of an energy barrier of height  $\Delta E$  by a factor  $\propto \exp(-\Delta E/k_B T)$  ( $k_B$  is the Boltzmann constant and  $T$  the temperature of the system), with such weights that allow the system to escape out of local minima. In most cases the weights are chosen in such a way that a Monte Carlo or molecular dynamics simulation will lead to a uniform distribution of a pre-chosen physical quantity. For instance, in multicanonical sampling [4] the weight  $w(E)$  leads to a distribution

$$P(E) \propto n(E)w(E) = \text{const}, \quad (1)$$

where  $n(E)$  is the spectral density. A free random walk in the energy space is performed that allows the simulation to escape from any local minimum. From this simulation one can calculate the thermodynamic average of any physical quantity  $A$  by re-weighting: [11]

$$\langle A \rangle_T = \frac{\int dx \mathcal{A}(x) w^{-1}(E(x)) e^{-E(x)/k_B T}}{\int dx w^{-1}(E(x)) e^{-E(x)/k_B T}}, \quad (2)$$

where  $x$  labels the configurations of the system. Note that the weight  $w(E)$  is not *a priori* known in generalized ensembles, and estimators have to be determined by an iterative procedure described in Refs. [4, 12].

Another way of enhancing the sampling of low-energy configurations in protein simulations is parallel tempering (also known as replica exchange or Multiple Markov Chain method) [9], a technique that was first introduced to protein folding in Ref. [13]. In its most common form, one considers  $N$  *non-interacting* copies of the molecule, each at a different temperature  $T_i$ . In addition to standard Monte Carlo or molecular dynamics moves that act only on one copy (i.e.

the molecule at a fixed temperature), an exchange of conformations between two copies  $i$  and  $j = i + 1$  is allowed with probability

$$w(\mathbf{C}^{old} \rightarrow \mathbf{C}^{new}) = \min(1, \exp(-\beta_i E(C_j) - \beta_j E(C_i) + \beta_i E(C_i) + \beta_j E(C_j))) . \quad (3)$$

The exchange of conformations will lead, especially at low temperatures, to a faster convergence of the Markov chain than is observed in regular canonical simulations with only local moves. This is because the resulting random walk in temperatures allows the configurations to move out of local minima and cross energy barriers. Note that parallel tempering does not require Boltzmann weights. The method can be combined easily with other generalized-ensemble techniques as was demonstrated first in Ref. [13].

The common idea behind all generalized-ensemble techniques is that they avoid by construction of the algorithm entrapment in local minima. Another realization of this idea is *energy landscape paving* (ELP) a new optimization method that proved very promising in protein studies [14, 15].

In ELP, one performs low-temperature Monte Carlo simulations with a modified energy expression designed to steer the search away from regions that have been already explored:

$$w(\tilde{E}) = e^{-\tilde{E}/k_B T} \quad \text{with} \quad \tilde{E} = E + f(H(q, t)) . \quad (4)$$

Here,  $T$  is a (low) temperature,  $\tilde{E}$  serves as a replacement of the energy  $E$  and  $f(H(q, t))$  is a function of the histogram  $H(q, t)$  in a pre-chosen ‘‘order parameter’’  $q$ . It follows that within ELP the weight of a local minimum state decreases with the time the system stays in that minimum till the local minimum is no longer favored. The system will then explore higher energies till it falls into a new local minimum. Obviously, for  $f(H(q, t)) = f(H(q))$  the method reduces to the various generalized-ensemble methods [10] (for instance for  $f(H(q, t)) = \ln H(E)$  to multicanonical sampling).

### 3 Secondary Structure and Folding

In order to demonstrate that generalized-ensemble simulations are well suited for protein research we will focus in the following on the role of secondary structure formation in the folding process.

#### 3.1 Helix Formation in Water

The two most common secondary structure elements are  $\alpha$ -helices and  $\beta$ -sheets. More accessible to numerical simulations are  $\alpha$ -helices as they involve only contacts between residues that are close in the protein chain. It is long known that  $\alpha$ -helices undergo a sharp transition toward a random coil state when the temperature is increased. The characteristics of this so-called helix-coil transition have been studied extensively [16]. An example is Ref. [17] where the order of the helix-coil transition in polyalanine in gas phase was studied. While their results, including a helix-coil transition

temperature of more than 500 K, are in agreement with recent experiments of Jarrold and collaborators [18], biologically more relevant is the question of helix formation for solvated proteins.

Preliminary investigations of this question were described in Refs.[19, 20] where polyalanine chains of length 10 have been studied. Because the size of these chains is too small to determine the order of the helix-coil transition for solvated polyalanine, we have performed multicanonical simulations of chains of length up to 30 residues [21] using a detailed, all-atom representation of these molecules. The interactions between the atoms are described by a standard force field, ECEPP/2 [22], as implemented in the program package SMMP [23]:

$$E_{ECEPP/2} = E_C + E_{LJ} + E_{HB} + E_{tor} , \quad (5)$$

$$E_C = \sum_{(i,j)} \frac{332q_i q_j}{\epsilon r_{ij}} , \quad (6)$$

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

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

$$E_{tor} = \sum_l U_l (1 \pm \cos(n_l \chi_l)) . \quad (9)$$

Here,  $r_{ij}$  (in Å) is the distance between the atoms  $i$  and  $j$ , and  $\chi_l$  is the  $l$ -th torsion angle. The peptide bond angles are set to their common value  $\omega = 180^\circ$ . We further assume that  $\epsilon = 2$  in the protein interior. All energies are in kcal/mol, hence the factor ‘332’ in the electrostatic energy term  $E_C$ .  $E_{LJ}$  is a Lennard-Jones term,  $E_{HB}$  the hydrogen-bond energy and  $E_{tor}$  accounts for the torsion energy of the molecule.

The interactions between the peptide and the surrounding water are approximated by adding a solvent accessible surface term  $E_{solv}$  [24] to the energy function:

$$E = E_{ECEPP/2} + E_{solv} \quad \text{with} \quad E_{solv} = \sum_i \sigma_i A_i . \quad (10)$$

In this approximation, named by us ASA, one assumes that the free energy difference between atomic groups immersed in the protein interior and groups exposed to water is proportional to the solvent accessible surface area  $A_i$  of the  $i$ th atom with the parameters  $\sigma_i$  as experimentally determined proportionality factors.

Simulations are performed for polyalanine with chain length of 10, 15, 20 and 30. The calculation of the multicanonical weight required between 100,000 ( $N = 10$ ) and 800,000 ( $N = 30$ ) sweeps. All thermodynamic quantities are estimated then from one production run of 6,000,000 Monte Carlo sweeps starting from a random initial conformation. Thermodynamic quantities that we have calculated from these multicanonical simulations include the average energy, specific heat, helicity, susceptibility and the complex partition function zeros whose analysis was introduced by us recently to protein studies [17].

In Fig. 1 we display  $q_H = \langle n_H(T) \rangle / (N - 2)$ , which is a natural order parameter for the helix-coil transition. Here,  $\langle n_H \rangle$  is the average number of helical residues. The normalization factor  $N - 2$  is chosen instead of  $N$ , the number of residues because the terminal residues are flexible and are usually not part of an  $\alpha$ -helix. A clear separation is observed between a high-temperature phase with few helical residues and a low-temperature phase that is characterized by a single  $\alpha$ -helix. The transition temperature can be determined from the corresponding plots in the specific heat  $C_N(T)$  that are shown in Fig. 2. As already observed in Ref. [20], the transition temperatures are for simulations in an implicit solvent lower than in gas phase. However, the differences decrease with chain length. If we extrapolate the listed temperatures to the infinite chain limit by  $T_c(L) = T_c(\infty) - a e^{-bN}$ , we find for ASA simulations as the critical temperature  $T_c(\infty) = 480$  K, which is only 30 K lower than the corresponding value for gas-phase simulations:  $T_c(\infty) = 514$  K. We note that the transition temperatures are outside of the range of physiologically relevant temperatures indicating limitations of our energy functions.

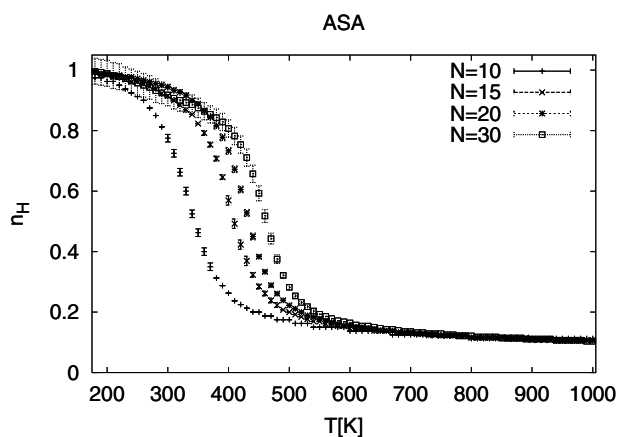


Figure 1. Temperature dependence of the helicity order parameter  $q_H = \langle n_H \rangle / (N - 2)$  as calculated from simulations of polyalanine of chain length  $N = 10, 15, 20, 30$  in an implicit solvent.

In order to research the strength of the observed helix-coil transition for solvated polyalanine, we analyze its partition function zeros using the approach by Janke and Kenna [25]. Our data (see Ref. [21] for details) lead to an estimate for the specific heat exponent,  $\alpha = 0.10(9)$ , that is small and within the errorbars compatible with zero. Hence, our analysis indicates that the helix-coil transition in solvated polyalanine is second order with exponents that are consistent with  $\alpha = 0$ , and by means of the hyperscaling relation  $\alpha = 2 - d\nu$  with  $d\nu = 2$ . These exponents are fundamentally different from the one in gas phase ( $\alpha = 0.86(14)$ ,  $\gamma = 1.06(10)$ ) and  $d\nu = 0.93(7)$  [17] that indicate a (weak) first order transition or a strong second order transition.

Our results are not unexpected. A large part of the energy gain through helix formation comes from the formation of hydrogen bonds between a residue and the fourth following one that characterizes an  $\alpha$ -helix. Within water this process competes with the entropically more favorable formation

of hydrogen bonds with the surrounding water. Hence, one can expect that helix-formation is more favored in gas-phase than in solvent.

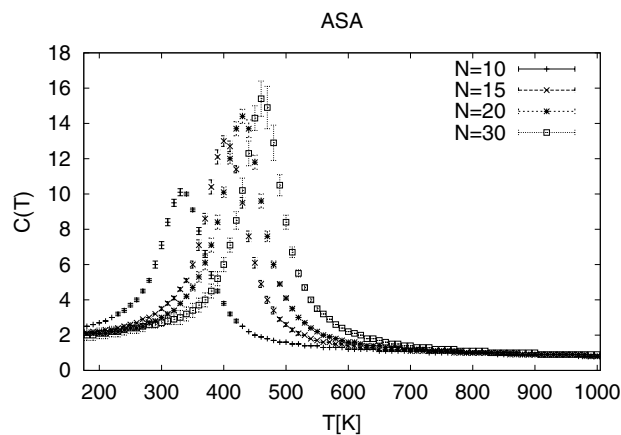


Figure 2. Specific heat  $C(T)$  as a function of temperature  $T$ .

### 3.2 Helix vs. Sheet Formation

It has become clear over the last years that mis-folding of proteins, often involving formation of  $\beta$ -sheets instead of  $\alpha$ -helices and subsequent aggregation, is the cause of various illnesses including Alzheimer's disease, BSE and other Prion diseases. In order to research the  $\alpha \rightarrow \beta$  we have chosen a peptide, whose sequence of amino acids in one letter code is EKAYLRT and that appears in natural occurring proteins with significant frequency at positions of both  $\alpha$ -helices and  $\beta$ -sheets. As in our previously described work, our results rely on multicanonical simulations of peptides in a detailed representation where the interactions between all atoms are taken into account. EKAYLRT is simulated both in gas phase and with our implicit solvent. We needed between 100,000 and 200,000 sweeps for the weight factor calculations. All thermodynamic quantities are then estimated from one production run of 2,000,000 Monte Carlo sweeps that followed 10,000 sweeps for "thermalization". A more detailed account of our results is published in Ref. [26].

We start with presenting our results for an EKAYLRT peptide that is not interacting with other molecules and display in Fig. 3 its average helicity  $\langle n_H \rangle$  as a function of temperature. Shown are data obtained in gas-phase (GP) and for the soluted peptide (ASA). We observe in both cases a steep helix-coil transition that separates a high-temperature region with little helicity from a low-temperature region where most of the residues are part of an  $\alpha$ -helix. An example for these helical configurations is shown in Fig. 4. The location of this helix-coil transition can be determined from the corresponding peaks in the specific heat  $C(T)$  that are drawn in the inset of Fig. 3. The more pronounced peak for the solvated molecule indicates a temperature  $T_{hc}^{ASA} = 340 \pm 10$  K that is considerably lower than the one in gas phase:  $T_{hc}^{GP} = 445 \pm 15$  K.

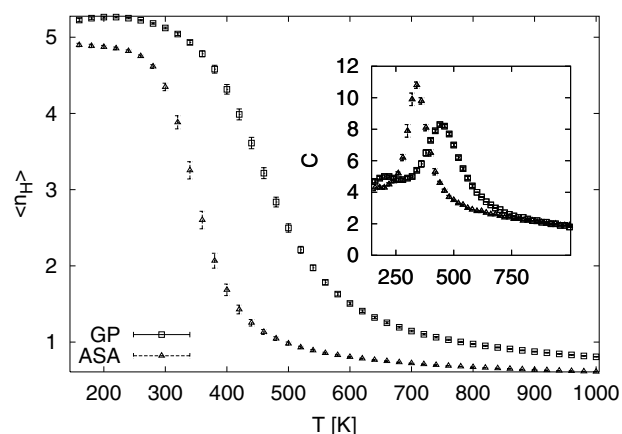


Figure 3. The average number  $\langle n_H \rangle$  of helical residues as a function of temperature  $T$  for EKAYLRT in gas phase (GP) and simulated with an implicit solvent term (ASA). The specific heat  $C(T)$  as function of temperature  $T$  is displayed in the inset.

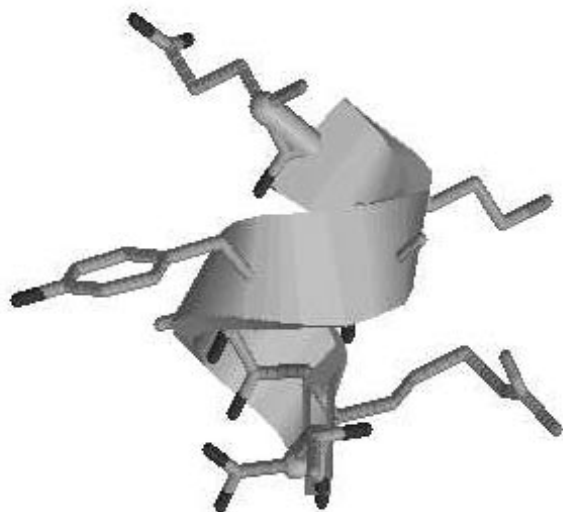


Figure 4. Lowest energy configuration of EKAYLRT.

Our results indicate that the peptide EKAYLRT has an intrinsic tendency to form helices. Strands have of order  $\approx 30$  kcal/mol higher free energies and are rarely observed (data not shown). This result is independent on whether the molecule is in gas phase or simulated with an implicit solvent. Since EKAYLRT appears *within* proteins both in helices and  $\beta$ -sheets it follows that sheet formation is due to the interaction of the peptide with its surrounding. We conjecture that EKAYLRT forms a  $\beta$ -sheet if it is in the proximity of another strand, for instance if it is close to another EKAYLRT peptide that is already in a strand configuration. Unfortunately, the present version of SMMP does not allow the simulation of two interacting proteins. Hence, in order to test our conjecture, we have studied instead the peptide

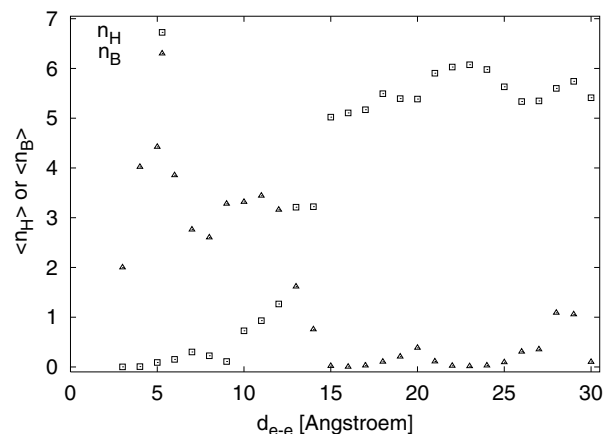


Figure 5. The average helicity  $\langle n_H \rangle$  and sheetness  $\langle n_B \rangle$  at  $T = 300$  K of the N-Terminal EKAYLRT residues as a function of the end-to-end distance  $d_{e-e}$ .

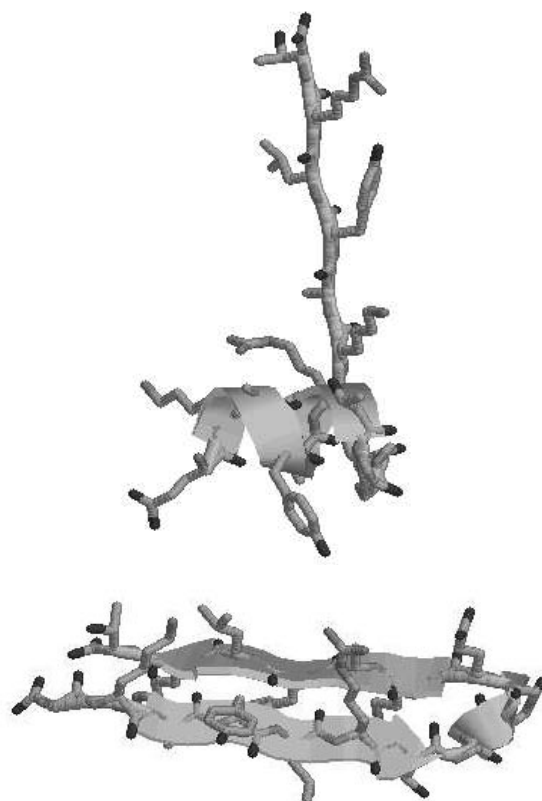


Figure 6. Low-energy configurations of molecule 'A'. The one in (a) is the lowest-energy configuration where the N-terminal EKAYLRT-residues form an  $\alpha$ -helix; the one in (b) where they form a  $\beta$ -sheet.

EKAYLRT-GGGG-EKAYLRT with the C-terminal EKAYLRT residues kept as  $\beta$ -strand. The four glycine residues form a flexible chain that holds the two EKAYLRT-units together but allows their relative positions to vary. We refer to the so constructed peptide as molecule 'A'.

In Fig. 5, we display the helicity and sheetness of the N-Terminal EKAYLRT at  $T = 300$  K as calculated from

the multicanonical simulation of molecule 'A'. Both quantities are shown as functions of the end-to-end distance  $d_{e-e}$  which is a measure for the separation of the two EKAYLRT chains. Two regions are observed. For  $d_{e-e} > \sim 16$  Å the N-terminal EKAYLRT chain forms a complete helix and strands are rarely observed. Hence, for these distances the N-terminal chain has a similar behavior as the isolated EKAYLRT-peptide. However, for decreasing end-to-end distance, the helicity also decreases and vanishes for  $d_{e-e} < \sim 10$  Å. At the same time, the sheetness increases and the peptide forms a  $\beta$ -sheet for  $d_{e-e} \approx 5 - 6$  Å. Examples of configurations that correspond to the two minima are shown in Fig. 6. Both minima have comparable free energies and are separated by barriers of only 2 kcal/mol allowing an easy interchange between the two forms.

Our results [26] suggest auto-catalytic properties for EKAYLRT: if the peptide forms a strand, it becomes favorable for other nearby EKAYLRT molecules to transform themselves into a sheet (instead of the normally preferred helix), and eventually to aggregate with the first one. We find that this behavior is due to more favorable Lennard-Jones and electrostatic interactions (data not shown) [26].

The behavior of EKAYLRT is similar to the mechanism thought to be responsible for the outbreak of neurodegenerative illnesses such as Alzheimer's or the Prion diseases. Outbreak of these illnesses is associated with the appearance of a mis-folded structure that differs from the correctly folded one by a  $\beta$ -sheet instead of an  $\alpha$ -helix. The mis-folded structure is thought to be auto-catalytic, that is its presence leads to a structural transition by which the correctly folded (helical) structure changes into the harmful  $\beta$ -sheet form. Hence, peptides based on the sequence of amino acids EKAYLRT can serve as simple models to study  $\alpha \rightarrow \beta$ -transitions and the mechanism of Prion diseases. For instance, our investigation suggests that the formation of  $\beta$ -sheets can be minimized by shielding the surface area of already existing  $\beta$ -sheet forms, minimizing in this way the van der Waals interaction. Another possibility may be to introduce metal ions that alter the electrostatic interaction, decreasing in this way the energy bias toward  $\beta$ -sheets.

## 4 Conclusion

We gave a brief introduction into generalized-ensemble techniques and their applications to the protein folding problem. Using one of these techniques we have probed the influence of water on the characteristics of the helix-coil transition in polyalanine and investigated the  $\alpha \rightarrow \beta$  transition in the peptide EKAYLRT. Our results underline that generalized-ensemble algorithms are well-suited for researching the thermodynamics of proteins and may lead to a deeper understanding of the protein-folding problem.

### Acknowledgments:

This article developed out of the invited talk that U.H. presented at the III Brazilian Workshop on Simulational Physics (Aug. 13-15, 2003). Parts of the results presented in this article are also published in Refs. [21, 26]. N.A. Alves gratefully acknowledges support by CNPq (Brazil),

and U.H. Hansmann support by a research grant from the National Science Foundation (CHE-9981874).

## References

- [1] U.H.E. Hansmann and Y. Okamoto, *Curr. Opin. Struc. Biol.* **9**, 177 (1999).
- [2] U.H.E. Hansmann, *Comp. Sci. Eng.* **5**, 64 (2003).
- [3] U.H.E. Hansmann and Y. Okamoto, *J. Comp. Chem.* **14**, 1333 (1993).
- [4] B.A. Berg and T. Neuhaus, *Phys. Lett.* **B267**, 249 (1991); *Phys. Rev. Lett.* **68**, 9 (1992).
- [5] P.M.C. de Oliveira, T.J.P. Penna and H.J. Herrmann, *Braz. J. Phys.* **26**, 677 (1996); P.M.C. de Oliveira, *Int. J. Mod. Phys. C* **9**, 497 (1998).
- [6] F. Wang and D.P. Landau, *Phys. Rev. Lett.* **86**, 2050 (2001)
- [7] C. Tsallis, *J. Stat. Phys.* **52**, 479 (1988).
- [8] T.J.P. Penna, *Phys. Rev. E* **51**, R1 (1995).
- [9] K. Hukushima and K. Nemoto, *J. Phys. Soc. (Jpn.)* **65**, 1604 (1996); G.J. Geyer, *Stat. Sci.* **7**, 437 (1992).
- [10] U.H.E. Hansmann and Y. Okamoto, In *Annual Reviews in Computational Physics VI*. Edited by Stauffer D. Singapore: World Scientific; 1999, 129-157.
- [11] A.M. Ferrenberg and R.H. Swendsen, *Phys. Rev. Lett.* **61**, 2635 (1988); *Phys. Rev. Lett.* **63**, 1658(E) (1989), and references given in the erratum.
- [12] U.H.E. Hansmann and Y. Okamoto, *Physica A* **212**, 415 (1994).
- [13] U.H.E. Hansmann, *Chem. Phys. Lett.* **281**, 140 (1997).
- [14] U.H.E. Hansmann and L.T. Wille, *Phys. Rev. Lett.* **88**, 068105 (2002).
- [15] H.P. Hsu, S.C. Lin and U.H.E. Hansmann, *Acta Cryst. A* **58**, 259 (2002).
- [16] D. Poland and H.A. Scheraga, *Theory of Helix-Coil Transitions in Biopolymers* (Academic Press, New York, 1970).
- [17] N.A. Alves and U.H.E. Hansmann, *Phys. Rev. Lett.* **84**, 1836 (2000).
- [18] Hudgins, R.R.; Ratner, M.A.; Jarrold, M.F. *J. Am. Chem. Soc.* **1998**, 120, 12974.
- [19] A. Mitsutake and Y. Okamoto, *Chem. Phys. Lett.* **309**, 95 (1999).
- [20] Y. Peng and U.H.E. Hansmann, *Biophys. J* **82**, 3269 (2002).
- [21] Y. Peng, U.H.E. Hansmann and N.A. Alves, *J. Chem. Phys.* **118**, 2374 (2003).
- [22] M.J. Sippl, G. Némethy, and H.A. Scheraga, *J. Phys. Chem.* **88**, 6231 (1984), and references therein.
- [23] F. Eisenmenger, U.H.E. Hansmann, Sh. Hayryan, C.-K. Hu, *Comp. Phys. Comm.* **138**, 192 (2001).
- [24] T. Ooi, M. Oobatake, G. Némethy, and H.A. Scheraga, *Proc. Natl. Acad. Sci. USA* **84**, 3086 (1987).
- [25] W. Janke and R. Kenna, *J. Stat. Phys.* **102**, 1211 (2001).
- [26] Y. Peng and U.H.E. Hansmann, submitted for publication.