

# **B 11 How to fold Proteins on Computers?**

U.H.E. Hansmann

John v. Neumann Institut für Computing,  
Forschungszentrum Jülich GmbH

Dept. of Physics, Michigan Technological  
University

## **Contents**

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Introduction</b>                            | <b>2</b>  |
| <b>2</b> | <b>Energy Landscape Paving</b>                 | <b>3</b>  |
| <b>3</b> | <b>Parallel Tempering</b>                      | <b>5</b>  |
| <b>4</b> | <b>Multicanonical Sampling</b>                 | <b>7</b>  |
| <b>5</b> | <b>Other generalized-ensemble techniques</b>   | <b>8</b>  |
| <b>6</b> | <b>Structure Predictions of Small Proteins</b> | <b>9</b>  |
| <b>7</b> | <b>Conclusion</b>                              | <b>12</b> |

# 1 Introduction

One of the most common and important class of molecules in living systems are proteins. Muscles and connective tissues are formed by them, and as enzymes, they catalyze and regulate biochemical reactions in the cell. Greatly differing in size and structure, all proteins are chemically linear chain molecules with the twenty naturally occurring amino acids as monomers. Locally, regular elements like helices, sheets and turns are formed, but the biological function of a protein is decided by its unique overall three-dimensional shape that is specified solely by the sequence of amino acids.

The sequence of amino acids that make up a protein is set in the genome. Hence, after the successful completion of the human genome project one knows in principal the chemical composition of all proteins in the human body. However, for most of the resolved protein sequences one does not know the corresponding structures. Since proteins are only functional if they fold into their specific shape, it is important to understand how the structure and function of proteins emerge from their sequence of amino acids.

One possibility to unveil the sequence-structure (function) relationship are computer experiments. Most proteins exist at room temperature in a *unique* structure that one can identify it with the lowest *potential* energy conformation [1]. Hence, structure prediction of proteins is a global optimization problem. Both deterministic methods such as the  $\alpha\mathbf{BB}$  algorithm [2] and stochastic methods like Monte Carlo minimization [3], simulated annealing [4] or genetic algorithms[5] are often exploited.

As with all optimization problems, choice of an appropriate energy function is of crucial importance. As calorimetric measurements show that a protein in its native state is only marginally more stable (by a free-energy difference of  $\approx 10 - 20$  kcal/mol) than the ensemble of the denatured conformations it is important to use realistic models where the interactions among all atoms are taken into account. The resulting potential energy  $E_{tot} = E_{protein} + E_{solv}$  (given in kcal/mol) is often written as a sum of two terms. The first term,  $E_{protein}$ , describes the interactions between all atoms within a protein, and the second term,  $E_{solv}$ , the interaction of a protein with the surrounding water. Since explicit inclusion of water molecules is computationally demanding one often has to rely on implicit solvent models. One example is the introduction of a solvent-accessible surface term that approximates the hydrophobic forces on the protein [6]

$$E_{solv} = \sum_i \sigma_i A_i . \quad (1)$$

Here  $A_i$  is the solvent-accessible surface area of the  $i$ -th atom in a given configuration, and  $\sigma_i$  is the empirically determined solvation parameter of the atom  $i$ .

As an example for the atomic force fields that model the interactions within a protein I show here the ECEPP energy function [7]. It is defined by the sum of an electrostatic term  $E_{es}$ , a van der Waals energy  $E_{vdW}$ , and a hydrogen-bond term  $E_{hb}$  for all pairs of atoms in the peptide together with a torsion term  $E_{tors}$  for all torsion angles:

$$\begin{aligned} E_{ECEPP} &= E_{es} + E_{vdW} + E_{hb} + E_{tors} , \\ E_{es} &= \sum_{(i,j)} \frac{332q_i q_j}{\epsilon r_{ij}} , \\ E_{vdW} &= \sum_{(i,j)} \left( \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) , \end{aligned}$$

$$E_{hb} = \sum_{(i,j)} \left( \frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right), \quad (2)$$

$$E_{tors} = \sum_l U_l (1 \pm \cos(n_l \alpha_l)). \quad (3)$$

Here,  $r_{ij}$  is the distance between the atoms  $i$  and  $j$ , and  $\alpha_l$  is the torsion angle for the chemical bond  $l$ . The parameters ( $q_i, A_{ij}, B_{ij}, C_{ij}, D_{ij}, U_l$  and  $n_l$ ) are calculated from crystal structures of amino acids. Since the bond lengths and bond angles are set constant, the true degrees of freedom are rotations around these bonds characterized by dihedral angles  $\phi, \psi, \omega$ , and  $\chi_i$ .

Unfortunately, computer simulations are notoriously difficult for such detailed protein models. Containing both repulsive and attractive terms, all-atom models of proteins lead to a very rough energy landscape with a huge number of local minima separated by high energy barriers. For this reason, 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. Only recently has been progress in alleviating the above stated multiple-minima problem. For a review, see, for instance, Ref. [8]. In the following, I will describe some of these methods that proved to be successful in numerical simulations and in whose development I was involved. I will further present some recent applications that illustrate the success and limitations of current protein simulations.

## 2 Energy Landscape Paving

A general characteristic of successful optimization techniques is that they avoid entrapment in local minima and continue to search for further solutions. One example that proved very promising in protein studies is ENERGY Landscape Pavingg (ELP) [9]. In this technique, one performs low-temperature Monte Carlo simulations with an effective energy designed to steer the search away from regions that have been already explored:

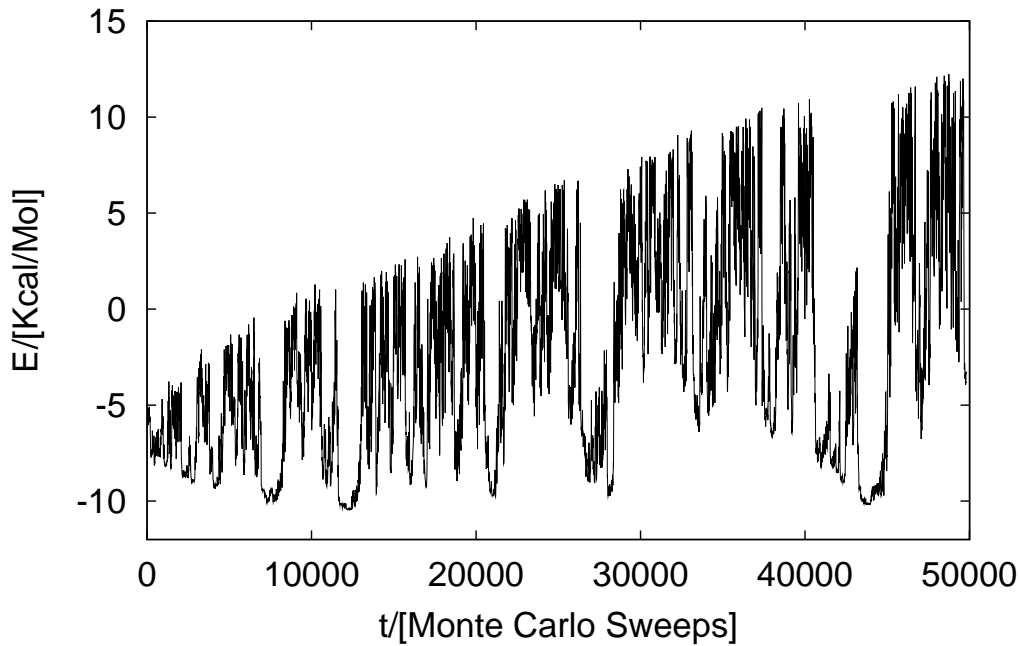
$$w(\tilde{E}) = e^{-\tilde{E}/k_B T} \quad \text{with} \quad \tilde{E} = E + f(H(q, t)). \quad (4)$$

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”  $q$ . This may be a “natural” quantity for the system under study or the energy itself.

The weight of a local minimum state decreases with the time the system stays in that minimum, i.e. ELP deforms the energy landscape locally till the local minimum is no longer favored, and the system will explore higher energies. It will then either fall in a new local minimum or walk through this high energy region till the corresponding histogram entries all have similar frequencies, and the system again has a bias toward low energies. Since the weight factor is time dependent it follows that ELP violates detailed balance. Hence, the method can not be used to calculate thermodynamic averages. Note, however, that for  $f(H(q, t)) = f(H(q))$  detailed balance is fulfilled, and ELP reduces to the *generalized-ensemble* methods [10] discussed later. The small peptide Met-enkephalin is used to illustrate the search process in ELP [9]. This pentapeptide has the sequence Tyr-Gly-Gly-Phe-Met and is a frequently used benchmark model to examine new algorithms. Its ground state is known for the ECEPP/2 field (see Eq. 3), as implemented in the computer code SMMP [11], and has an energy  $E_0 = -10.7 \text{ kcal/mol}$ . Since the next higher local minimum has an energy of  $E_1 = -9.8 \text{ kcal/mol}$  [12], one can easily identify any configuration with energy below  $E = -9.8 \text{ kcal/mol}$  as a representative of

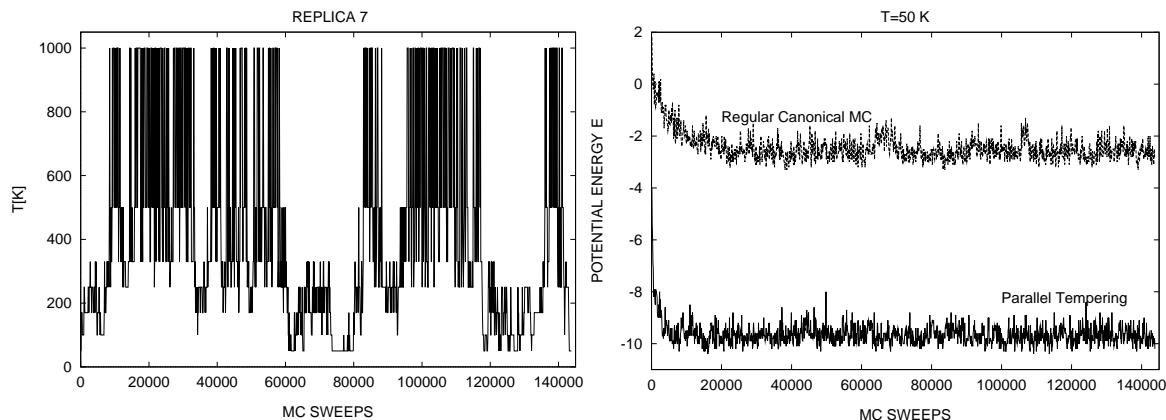
the ground state. As in our algorithmic presentation of ELP we use the potential energy itself as an order parameter. Thus the deformed energy landscape of Met-enkephalin is generated by  $\tilde{E} = E + H(E, t)$ , where  $H(E, t)$  is the histogram in energy at MC sweep  $t$ . We chose a bin size  $E_{bin} = 0.25 \text{ kcal/mol}$  in the histogram and set the temperature to  $T = 50 \text{ K}$ .

Fig. 1 illustrates the search process in energy landscape paving. The starting configuration has an energy of  $E_{start} = -5.1 \text{ kcal/mol}$  and was obtained from a random configuration through quenching in initial 100 sweeps. The simulation soon gets trapped in a local minimum of  $E \approx -7.8 \text{ kcal/mol}$  (after only 250 MC sweeps). Through the following MC sweeps entries in the corresponding histogram bin are accumulated and the energy landscape locally deformed, until after about 750 MC sweeps the simulation escapes this local minimum to find a lower local minimum after 2000 MC sweeps. This process is repeated till the simulation finds the global minimum conformation for the first time after 7260 sweeps. Within the 50,000 sweeps of our simulation the ground state region ( $E < -9.8 \text{ kcal/mol}$ ) was reached 5 times each time separated by explorations in the high energy region. Note that the range of energies covered increases with MC time: ELP starts with filling up the small ‘potholes’ in the energy landscape, but fills up also large valleys as the simulation continues.



**Fig. 1:** “Time series” of energy for a ELP simulation of the peptide Met-enkephalin. The figure is taken from Ref. [9]

We have tested the efficiency of ELP by performing 20 independent ELP runs of each 50,000 MC sweeps. The results of the ELP runs are compared with 20 simulated annealing [4] runs of equal statistics using the annealing schedule that proved to be optimal for Met-enkephalin in Ref. [13]. However, even with this optimized annealing schedule, the ground state is found only in  $8/20 = 40\%$  of the simulations and the average value of the lowest energy conformation ( $\langle E_{min} \rangle = -8.5 \text{ kcal/mol}$ ) is above our threshold for ground state configurations ( $-9.8 \text{ kcal/mol}$ ). On the other hand, with ELP we find the ground state in each of the 20



**Fig. 2:** “Time series” of temperature for one copy of Met-enkephalin and energy at temperature  $T = 50\text{K}$  as obtained from a parallel tempering simulation. The figure is taken from Ref. [16]

runs. As a consequence the average of lowest energy states  $\langle E_{min} \rangle = -10.3 \text{ kcal/mol}$  is well below our threshold for ground state configuration.

Note also that ELP allows even the possibility of zero-temperature simulations [14]. For  $T \rightarrow 0$  only moves with  $\Delta \tilde{E} \leq 0$  will be accepted. If we choose:  $\tilde{E} = E + cH(E, t)$ , we find as acceptance criterion:

$$\Delta E + c\Delta H(q, t) \leq 0 \leftrightarrow c\Delta H(q, t) \leq -\Delta E \quad (5)$$

where  $E$  is the physical energy. Hence, within ELP the system can overcome even at  $T = 0$  any energy barrier. The waiting time for such a move is proportional to the height of the barrier that needs to be crossed. Note that the factor  $c$  sets now only the time scale and in this sense the  $T = 0$  form of ELP is parameter-free.

### 3 Parallel Tempering

Structure prediction by means of global optimization requires the use of an energy function that describes the interactions within a protein and between the protein and the surrounding water. Hence, any global optimization approach to structure prediction of proteins is limited by the accuracy of the force fields. Global optimization techniques are also not suitable for investigations of the structural transitions in proteins that are a key issue for understanding the folding and biological function of a number of proteins. As with structure prediction, it is necessary to go beyond global optimization techniques such as ELP and to measure thermodynamic quantities, i.e. to sample a set of configurations from a canonical ensemble and take an average of the chosen quantity over this ensemble.

Such sampling is hampered by the roughness of the energy landscape. One popular method to overcome the resulting extremely slow thermalization at low temperatures is parallel tempering [15] (also known as replica exchange method or Multiple Markov chains), a techniques that was first applied to protein studies in Ref. [16]. In its most common form, one considers in parallel tempering an artificial system built up of  $N$  *non-interacting* replicas 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 replica (i.e. the molecule at a fixed temperature), an exchange of

conformations between two copies  $i$  and  $j = i + 1$  is allowed with probability

$$w(\mathbf{C}^{old} \rightarrow \mathbf{C}^{new}) = \min(1, \exp(-\beta_i E(C_j) - \beta_j E(C_i) + \beta_i E(C_i) + \beta_j E(C_j))) . \quad (6)$$

The exchange of conformations will at low temperatures lead 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. [16].

Met-enkephalin is used again to illustrate the parallel tempering algorithm. Simulations with seven copies were performed [16]. The corresponding temperatures are  $T_1 = 1000$  K,  $T_2 = 500$  K,  $T_3 = 330$  K,  $T_4 = 250$  K,  $T_5 = 170$  K,  $T_6 = 100$  K and  $T_7 = 50$  K. The simulation consists of 144,000 sweeps for each copy. After each sweep, an exchange of conformations between pairs of copies at neighboring temperatures was tried. The “time series” of temperatures for one of the seven copies is shown in Fig. 2. Due to the exchange move the configuration walks randomly between low temperatures and high temperatures. The resulting random walk in energy ensures - as in the case of ELP - that any energy barrier can be overcome, and the molecule will thermalize at all 7 temperatures. The faster convergence can be seen in Fig. 2 where also the “time series” in energy is displayed for both a regular canonical simulation at  $T = 50$  K and for the copy with  $T = 50$  K of a parallel tempering simulation. Obviously the regular canonical Monte Carlo got trapped in a local minimum and was not able to thermalize. From previous simulations (see Ref. [17]) it is known that even 1,000,000 sweeps are not enough to thermalize Met-enkephalin at  $T = 50$  K. On the other hand, with the exchange of configurations by parallel tempering the simulation thermalizes at  $T = 50$  K in less than 10,000 sweeps.

An interesting variant of the parallel tempering idea is “model hopping” [18] where the random walk in temperatures is replaced by one through an ensemble of models with slightly altered energy functions. For this we assume that the energy function can be separated in two terms:  $E = E_A + aE_B$ . As in parallel tempering, MH considers  $N$  non-interacting copies of the molecule, but copies are now exchanged according to

$$w(\mathbf{C}^{old} \rightarrow \mathbf{C}^{new}) = \min(1, \exp\{-\beta [E_A(C_j) + a_i E_B(C_j) + E_A(C_i) + a_j E_B(C_i) \quad (7)$$

$$- E_A(C_i) - a_i E_B(C_i) - E_A(C_j) - a_j E_B(C_j)]\}) \quad (8)$$

$$= \min(1, \exp\{\beta \Delta a \Delta E_B\}) . \quad (9)$$

Here,  $\Delta a = a_j - a_i$  and  $\Delta E_B = E_B(C_j) - E_B(C_i)$ . Due to this exchange move configurations perform a random walk on a ladder of models with  $a_1 = 1 > a_2 > a_3 > \dots > a_N$  that differ by the relative contributions of  $E_B$  to the total energy  $E$  of the molecule. For instance, barriers in the energy landscape of proteins often arise from van der Waals repulsion between atoms that come too close. In MH the protein walks randomly up and down on a ladder of models with successively smaller contributions from the van der Waals energy. While the “physical” system is on one side of the ladder (at  $a_1 = 1$ ), the (non-physical) model on the other end of the ladder (at  $a_N \ll 1$ ) may allow atoms to share the same position in space. As the protein “tunnels” in this way through energy barriers, sampling of low-energy configurations will be enhanced in the “physical” model (at  $a_1 = 1$ ).

## 4 Multicanonical Sampling

Generalized-ensemble simulations [10] offer another possibility to overcome the multiple minima problem and to calculate reliable low-temperature quantities. The idea is again to ensure that a simulation does not get trapped in local minima but samples both low and high energy states with sufficient probability. Such movement in and out of local minima is obtained by requiring that a Monte Carlo or molecular dynamics simulation shall lead to a uniform distribution of a pre-chosen physical quantity. Probably the earliest realization of this idea is *umbrella sampling* [19], but it has been lately re-discovered in various forms such as multicanonical sampling [20], simulated tempering [21], ect. The first application of these new techniques to protein simulations can be found in Ref. [22] where a Monte Carlo technique was used. Later, a formulation for the molecular dynamics method/ was also developed [23].

In the *multicanonical algorithm* [20] configurations with energy  $E$  are assigned a weight  $w(E)$  such that the distribution of energies

$$P_{mu}(E) \propto n(E)w_{mu}(E) = \text{const}, \quad (10)$$

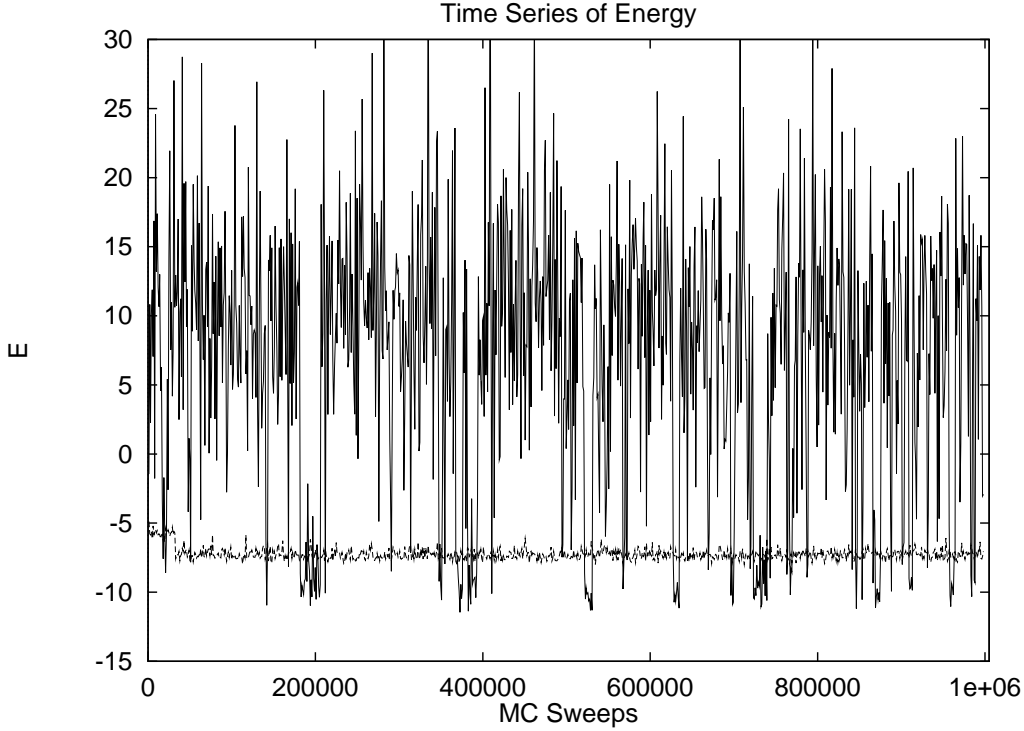
where  $n(E)$  is the spectral density. Since all energies appear with the equal probability, a free random walk in the energy space is enforced: the simulation can overcome *any* energy barrier and will not get trapped in one of the many local minima. In order to demonstrate the latter point the “time series” of energy is shown in Fig. 3 as a function of Monte Carlo sweeps for both a regular canonical Monte Carlo simulation at temperature  $T = 50$  K (dotted curve) and a multicanonical simulation. The displayed data are again from a simulation of the pentapeptide Met-enkephalin using a slightly modified version [13] of the ECEPP/2 force field. Starting from a random configuration the two simulations continued for 1,000,000 Monte Carlo sweeps. 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  $\langle E \rangle = -11.1$  kcal/mol as found in Ref. [13]. On the other hand, the multicanonical simulation 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. From such a multicanonical simulation one can not only locate the energy global minimum, but also calculate the expectation value of any physical quantity  $\mathcal{O}$  at temperature  $T$  by re-weighting techniques [24]

$$\langle \mathcal{O} \rangle_T = \frac{\int dE \mathcal{O}(E) P_{mu}(E) w_{mu}^{-1}(E) e^{-E/k_B T}}{\int dE P_{mu}(E) w_{mu}^{-1}(E) e^{-E/k_B T}} \quad (11)$$

$$= \frac{\int dx \mathcal{O}(x) w_{mu}^{-1}(E(x)) e^{-\beta E(x)}}{\int dx w_{mu}^{-1}(E(x)) e^{-\beta E(x)}} \quad (12)$$

where  $x$  stands for configurations.

Unlike in the canonical ensemble the weights  $w_{mu}(E) \propto n^{-1}(E)$  are not *a priori* known and one needs their estimates for a numerical simulation. Hence, multicanonical sampling consist of three steps: Calculation of the multicanonical (and other generalized-ensemble weights) is usually done by an iterative procedure [22, 13]. The following algorithmic presentation describes a simple version of this procedure. In it, one uses that the histogram of a multicanonical simulation can be written as  $H(E) = n(E)w_{mu}^i(E)$  where  $w_{mu}^i(E)$  is the  $i$ -th estimate of the canonical weight. Setting  $w_{mu} = 1/n(E)$ , one obtains the iterative relation  $w_{mu}^{i+1} = w_{mu}^i(E)/H(E)$ . *Iter* is the number of iterative improvements of the weights  $w_{mu}(i)$ , *sweeps* is the number of Monte



**Fig. 3:** “Time series” of energy for the pentapeptide Met-enkephalin. Shown are both the results from a canonical simulation at  $T = 50$  K (dotted line) and a multicanonical simulation.

Carlo sweeps in each cycle, and  $nbin$  is the number of energy bins. We remark that calculation of the weights can be slow (about 40% of the total CPU time was spent in Ref. [22] on this point) and several attempts were made to obtain generalized-ensemble weights in a faster way; see, for instance, Ref. [25].

## 5 Other generalized-ensemble techniques

In multicanonical simulations the computational effort increases with the number of residues like  $\approx N^4$  (when measured in Metropolis updates) [26]. In general, the computational effort in simulations increases with  $\approx X^2$  where  $X$  is the variable in which one wants a flat distribution. This is because generalized-ensemble simulations realize by construction of the ensemble a  $1D$  random walk in the chosen quantity  $X$ . In the multicanonical algorithm the reaction coordinate  $X$  is the potential energy  $X = E$ . Since  $E \propto N^2$  the above scaling relation for the computational effort  $\approx N^4$  is recovered. Hence, multicanonical sampling is not always the optimal generalized-ensemble algorithm in protein simulations. A better scaling of the computer time with size of the molecule may be obtained by choosing more appropriate reaction coordinate for our ensemble than the energy.

One often used choice is *simulated tempering* [21] where the temperature itself becomes a dynamic variable and is sampled uniformly. Temperature and configuration are both updated with a weight:

$$w_{ST}(T, E) = e^{-E/k_B T - g(T)} . \quad (13)$$



Here, the function  $g(T)$  is chosen so that the probability distribution of temperature is given by

$$P_{ST}(T) = \int dE n(E) e^{-E/k_B T - g(T)} = \text{const} . \quad (14)$$

Physical quantities have to be sampled for each temperature point separately and expectation values at intermediate temperatures are calculated by re-weighting techniques [24].

As common in generalized-ensemble simulations, the weight  $w_{ST}(T, E)$  is not *a priori* known (since it requires knowledge of the parameters  $g(T)$ ) and their estimator has to be calculated. They can be again obtained by an iterative procedure as described in section 4. In the simplest version the improved estimator for  $g^{(i)}(T)$  for the  $i$ -th iteration is calculated from the histogram of temperature distribution  $H_{ST}^{(i-1)}(T)$  of the preceding simulation as follows:

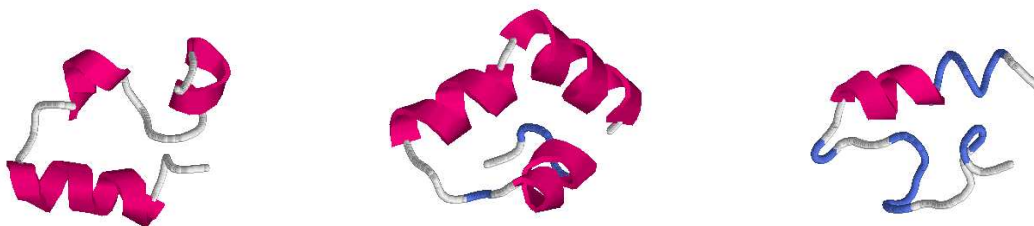
$$g^{(i)}(T) = g^{(i-1)}(T) + \log H_{ST}^{(i-1)}(T) . \quad (15)$$

In this procedure one uses that the histogram of the  $i$ -th iteration is given by

$$H_{ST}(T) = e^{-g_{i-1}(T)} Z_i(T) , \quad (16)$$

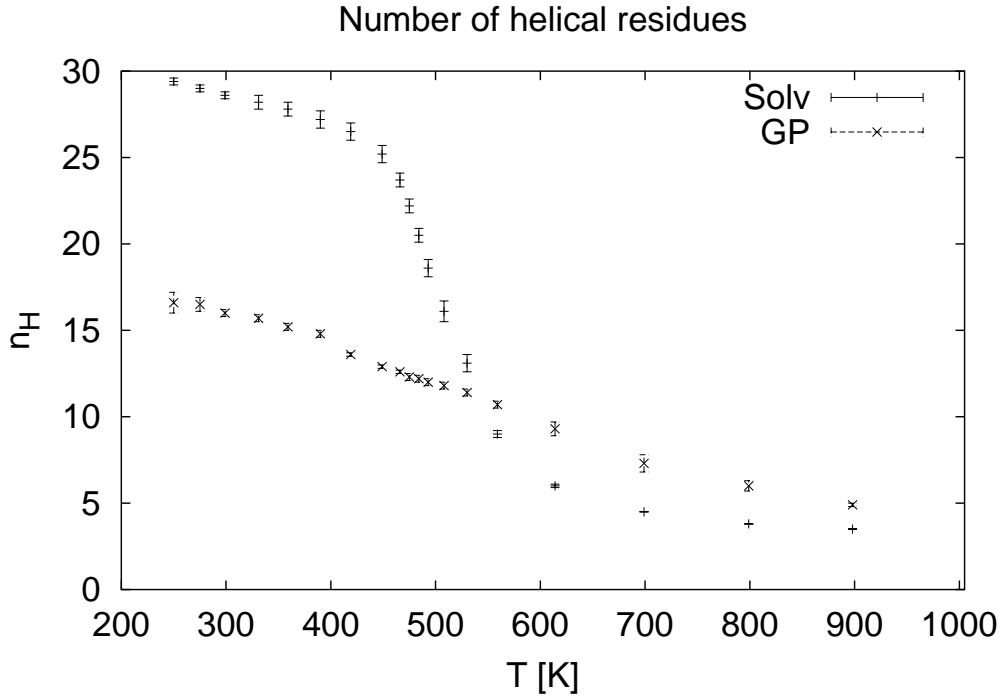
where  $Z_i(T) = \int dE n(E) \exp(-E/k_B T)$  is an estimate for the canonical partition function at temperature  $T$ . Setting  $\exp(g_i(T)) = Z_i(T)$  leads to the iterative relationship of Eq. 15.

## 6 Structure Predictions of Small Proteins



**Fig. 4:** Left: *Experimental structure of HP-36 as deposited in the PDB data-bank.* Middle: *Lowest energy structure as obtained in a simulation of the solvated peptide.* Right: *Lowest energy structure of HP-36 as obtained in a simulation in gas phase.*

The second example is the 36-residue villin headpiece subdomain HP-36, one of the smallest peptides that can fold autonomously. HP-36 was chosen by Duan and Kollman for a 1-microsecond molecular dynamics simulation of protein folding [27]. The experimental structure was determined by NMR analysis [28]. Luc Wille (Florida Atlantic University) and I have used this protein to study the efficiency of the ELP algorithm. We have used the approach of



**Fig. 5:** Average number of helical residues  $\langle n_H \rangle(T)$  of HP-36 as a function of temperature for both the solvated protein and in gas-phase. The figure is taken from Ref. [30]

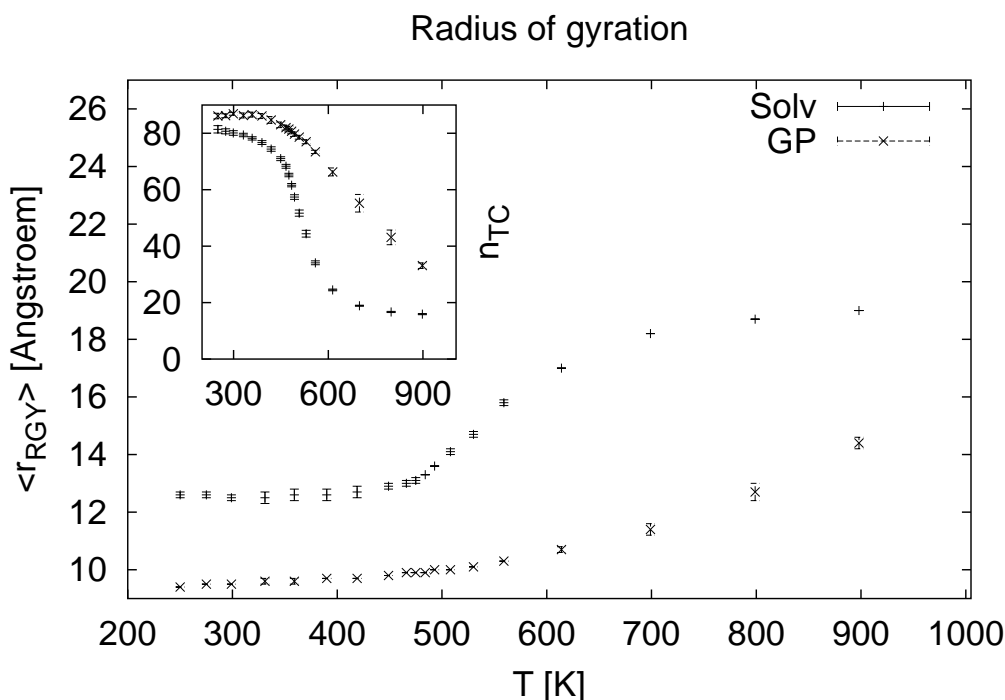
Eq. 1 to approximate the interaction between protein and water with the parameters  $\sigma_i$  chosen from Ref. [29].

Build up only out of  $\alpha$ -helices as secondary structure elements, HP-36 allows in a simple way the definition of an order parameter to characterize configurations other than by their energy. This natural order parameter is the number  $n_H$  of residues in the peptide which are part of an  $\alpha$ -helix. Throughout the search process we try to deform the energy landscape by means of a histogram  $H(E, n_H, t)$  in *both* helicity and energy:  $\tilde{E} = E + H(E, n_H, t)$ . Operating again at a temperature  $T = 50$  K, we find as weights for the search algorithm

$$w(E, n_H, t) = e^{-\beta(E + H(E, n_H, t))} . \quad (17)$$

Using this weight we performed simulations with 50,000 MC sweeps (starting from random configurations) keeping track of the lowest energy configuration during the search process.

The structure of HP-36 as obtained from the Protein Data Bank (PDB code 1vii) is shown in Fig. 4. The structure consists of three helices between residues 4-8, 15-18, and 23-32, respectively, which are connected by a loop and a turn. We find for this structure in our model an energy (ECEPP/2 + solvation term)  $E_{nat} = -276$  kcal/mol. Our approach led to a configuration with the lowest energy  $E_{min} = -277$  kcal/mol which we show also in Fig. 4 [9]. The above structure consists of three helices where the first helix stretches from residue 2 to residue 11 and is more elongated than the corresponding one in the native structure (residues 4-8). The second helix consist of residues 13-17 (compared to residue 15-18 in the native structure) and the third helix stretches from residue 23-33 (residues 23-32 in the PDB structure). The structure has 95% of the native helical content and a radius of gyration  $R_g = 10.1$  Å which indicates that the numerically obtained structure is slightly less compact than the experimental structure



**Fig. 6:** Average radius of gyration  $\langle r_{gy} \rangle(T)$  of HP-36 as a function of temperature for both the solvated protein and in gas-phase. The figure is taken from Ref. [30]

( $R_\gamma = 9.6\text{\AA}$ ). 60% of native contacts are formed. These values are comparable with the results in Ref. [27] (but required orders of magnitude less computer time) where the optimal structure of a  $1\text{ }\mu\text{s}$  molecular dynamic folding simulation showed 80% of native helical content and 62 % of native contacts. Similarly comparable were the values of the root-mean-square deviation (RMSD) of both numerically determined conformers to the native structure:  $5.8\text{ }\text{\AA}$  versus  $5.7\text{ }\text{\AA}$  in Ref. [27] (counting only backbone atoms). On the other hand, an ELP simulation of 50,000 sweeps relying only on the ECEPP/2 force field led to a structure with an ECEPP energy of  $E_{GP} = -192\text{ kcal/mol}$ . That structure, shown in the bottom of Fig. 4, is build out of two helices (between residues 2-16 and 23-33) connected by a loop, differs significantly from the regularized PDB structure with the higher potential energy  $E_{nat} = -176\text{ kcal/mol}$ . Hence, the native structure of the peptide HP-36 is *not* the global minimum configuration in ECEPP/2 in gas phase.

In order to understand more the differences between the gas-phase results and that with a solvent accessible surface term, Chai-Yu Lin, Chin-Ku Hu (both Academia Sinica, Taiwan) and I have simulated recently HP-36 with parallel tempering on 20 nodes of a cluster of IBM 4-ways 375MHZ SMP Thin Nodes [30]. We have chosen as temperatures  $T = 1000, 900, 800, 700, 610, 560, 530, 510, 495, 485, 475, 465, 450, 420, 390, 360, 330, 300, 275, 250\text{ K}$ . On each node, we performed 150,000 MC sweeps, and a replica exchange move was attempted after each sweep. Both gas-phase simulations and such relying on a solvent-accessible surface term with the parameter set OONS of Ref. [6] were performed.

From these parallel tempering simulations we have calculated the number of helical residues as function of temperature. Fig. 6 displays our results. Little difference is found at high temperatures. However, below the transition temperature  $T \approx 490\text{ K}$  the data for both simulations

diverge. The helicity grows rapidly with decreasing temperature in the OONS simulation while it stays small in gas phase. Configurations in gas phase and in OONS simulations differ also in their compactness. We display in Fig. 6 for HP-36 two quantities that measure the compactness of protein configurations. The main graph is a plot of the average radius of gyration  $\langle r_{gy} \rangle(T)$  as a function of temperature. The corresponding values for the total number of contacts  $\langle n_{TC}(T) \rangle$  are shown in the inset. Both plots indicate that configurations in gas-phase are substantially more compact than the ones in the OONS simulation. For instance, at  $T = 300$  K, we find  $r_{gy} = 9.6(1)$  Å in gas phase compared to  $r_{gy} = 12.5(1)$  Å in OONS simulations. Note that even at  $T = 1000$  K, the peptide in gas phase has a radius-of gyration  $r_{gy} = 15.6(1)$  Å and is substantially more compact than in OONS simulation ( $r_{gy} = 19.2$  Å). We conjecture that this bias toward compact configurations inhibits the formation of  $\alpha$ -helices, and that the low-energy states of HP-36 in gas phase are characterized by large density and low helicity.

Our simulations of HP-36 demonstrate that the simulation techniques described in this review allow one not only to predict the structure of small peptides but also to evaluate the limitations of the utilized energy functions. For instance, in our example, we were able to determine the reasons behind the failure of gas-phase simulations when compared to such with simple solvent approximations. Since presently available energy functions are often parametrized for small molecules, their limitations will become more obvious as one proceeds toward larger systems. Modern simulation techniques may open ways to unveil and finally overcome these limitations.

## 7 Conclusion

I gave a brief introduction into some techniques used in simulations of the protein folding problem. These examples demonstrate that modern simulation algorithms are well-suited for investigations both of the thermodynamics of proteins and the prediction of their structure. It seems now that all-atom simulations of proteins are rather restricted by the accuracy of the present energy functions than by the efficiency of the search algorithms.

### Acknowledgements:

The presented work was partially supported by research grants of the National Science Foundation (CHE-0313618) and the National Institutes of Health (GM62838), both USA.

## References

- [1] C.B. Anfinsen, *Principles that govern the folding of protein chains*, Science **181**, 223 (1973).
- [2] I.P. Androulakis, C.D. Maranas and C.A. Floudas, Comparative Study of Global Minimum Energy Conformations of Hydrated Peptides, J. Glob. Opt. **11**, 1 (1997).
- [3] Z. Li and H.A. Scheraga, Monte Carlo-minimization approach to the multiple-minima problem in protein folding, Proc. Natl. Acad. Sci. U.S.A. **84**, 6611 (1987).
- [4] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, *Optimization by simulated annealing*, Science **220**, 671 (1983).
- [5] J. Holland, *Adaption in Natural and Artificial Systems*. University of Michigan Press 1975.
- [6] T. Ooi, M. Obatake, G. Nemethy, and H.A. Scheraga, Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides, Proc. Natl. Acad. Sci. USA **8**, 3086 (1987).
- [7] M.J. Sippl, G. Némethy, and H.A. Scheraga, *Intermolecular potentials from crystal data. 6. Determination of empirical potentials for O-H $\cdots$ O=C hydrogen bonds from packing configurations*, J. Phys. Chem. **88**, 6231 (1984), and references therein.
- [8] U.H.E. Hansmann and Y. Okamoto, *New Monte Carlo Algorithms for Protein Folding*, Curr. Opin. Struc. Biol. **9**, 177 (1999).
- [9] U.H.E. Hansmann and L.T. Wille, *Global Optimization by Energy Landscape Paving*, Phys. Rev. Let. **88**, 068105 (2002).
- [10] U.H.E. Hansmann and Y. Okamoto, *The Generalized-Ensemble Approach for Protein Folding Simulations*, In: D. Stauffer (ed.), *Annual Reviews in Computational Physics VI*, 129 (World Scientific, Singapore, 1999).
- [11] F. Eisenmenger, U.H.E. Hansmann, Sh. Hayryan, C.-K. Hu, *[SMMP] A Modern Package for Simulation of Proteins*, Comp. Phys. Comm. **138**, 192 (2001).
- [12] F. Eisenmenger and U.H.E. Hansmann, *Variation of the Energy Landscape of a Small Peptide under a Change from the ECEPP/2 Force Field to ECEPP/3*, J. Phys. Chem. B **101**, 3304 (1997).
- [13] U.H.E. Hansmann and Y. Okamoto, *Comparative Study of Multicanonical and Simulated Annealing Algorithms in the Protein Folding Problem*, Physica A **212**, 415 (1994).
- [14] A. Schug, W. Wenzel and U.H.E. Hansmann, *Energy Landscape Paving Simulations of the trp-cage Protein*, J. Chem. Phys., **122** 194711(2005).
- [15] K. Hukushima and K. Nemoto, *Exchange Monte Carlo Method and Applications to Spin Glass Simulations*, J. Phys. Soc. (Jpn.) **65**, 1604 (1996); G.J. Geyer, *Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference*, J. Am. Stat. Assn **90** (431), 909 (1995)

- [16] U.H.E. Hansmann, *Parallel Tempering Algorithm for Conformational Studies of Biological Molecules*, Chem. Phys. Lett. **281**, 140 (1997).
- [17] U.H.E. Hansmann and Y. Okamoto, *Generalized-Ensemble Monte Carlo Method for Systems with Rough Energy Landscape*, Phys. Rev. E **56**, 2228 (1997).
- [18] W. Kwak and U.H.E. Hansmann, *Efficient Sampling of Protein Structures by Model Hopping*, Phys. Rev. Lett. **95** 138102 (2005).
- [19] G.M. Torrie and J.P. Valleau, *Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling*, J. Comput. Phys. **23**, 187 (1977).
- [20] B.A. Berg and T. Neuhaus, *Multicanonical algorithms for first order phase transitions*, Phys. Lett. B **267**, 249 (1991).
- [21] A.P. Lyubartsev, A.A. Martinovski, S.V. Shevkunov, P.N. Vorontsov-Velyaminov, *New approach to Monte Carlo calculations of the free energy: Method of expanded ensembles*, J. Chem. Phys., **96**, 1776 (1992); E. Marinari, G. Parisi, *Simulated Tempering: A new Monte Carlo Scheme*, Europhysics Letters, **19**, 451 (1992).
- [22] U.H.E. Hansmann and Y. Okamoto, *Prediction of Peptide Conformation by Multicanonical Algorithm: A New Approach to the Multiple-Minima Problem*, J. Comp. Chem. **14**, 1333 (1993).
- [23] U.H.E. Hansmann, Y. Okamoto, and F. Eisenmenger, *Molecular dynamics, Langevin and hybrid Monte Carlo simulations in a multicanonical ensemble*, Chem. Phys. Lett. **259**, 321 (1996).
- [24] A.M. Ferrenberg and R.H. Swendsen, *New Monte Carlo technique for studying phase transitions*, Phys. Rev. Lett. **61**, 2635 (1988); *Optimized Monte Carlo Data Analysis*, Phys. Rev. Lett. **63**, 1658(E) (1989), and references given in the erratum.
- [25] F. Wang and D.P. Landau, *Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States*, Phys. Rev. Lett. **86**, 2050 (2001).
- [26] U.H.E. Hansmann and Y. Okamoto, *Finite-size scaling of helix-coil transitions in poly-alanine studied by multicanonical simulations*, J. Chem. Phys. **110**, 1267 (1999); **111**, 1339(E) (1999).
- [27] Y. Duan and P.A. Kollman, *Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution*, Science **282**, 740 (1998).
- [28] C.J. McKnight, D.S. Doehring, P.T. Matsudaria and P.S. Kim, *A Thermostable 35-Residue Subdomain within Villin Headpiece*, J. Mol. Biol. **260**, 126 (1996).
- [29] L. Wesson and D. Eisenberg, *Atomic solvation parameters applied to molecular dynamics of proteins in solution*, Protein Science **1**, 227 (1992).
- [30] C.-Y. Lin, C.-K. Hu and U.H.E. Hansmann, *Parallel Tempering Simulations of HP-36, Proteins* **52**, 436 (2003).