New algorithms and the physics of proteins

Ulrich H.E. Hansmann

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

Abstract

We review recent developments of modern simulation techniques that allow study of structural transitions and folding in peptides and small proteins. As one example for applications of these techniques, we study the helix formation in homopolymers. These investigations are then extended to research into the relation between helix–coil transition and folding in a simple artificial peptide. Finally, we present recent results on the 36-residue villin headpiece peptide HP-36 as an example for structure prediction of proteins with our techniques.

© 2002 Elsevier Science B.V. All rights reserved.

PACS: 02.70.Lq; 87.15By

Keywords: Generalized-ensemble simulations; Helix–coil transition; Structure prediction; Protein-folding problem

1. Introduction

The biological functions of proteins are closely related to their three-dimensional shape. Hence, knowledge on how this shape depends on the sequence of amino acids that form the protein would allow us to understand better the function of enzymes, working of the immune system, and could lead to more efficient ways of drug design. In principle, such knowledge can be obtained by means of computer experiments. However, simulations of realistic protein models are extremely difficult. Containing both repulsive and attractive terms, such models lead to a very rough energy landscape and a huge number of local minima separated by high energy barriers. As a consequence, simple canonical Monte Carlo or molecular dynamics simulations will at low temperatures not thermalize within the finite amount of available CPU time, and physical quantities cannot be calculated accurately.

E-mail address: hansmann@mtu.edu (U.H.E. Hansmann).

0378-4371/03/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0378-4371(02)01764-8
A number of novel simulation techniques were developed over the last few years that promise to alleviate the above-stated multiple-minima problem (for a recent review, see Ref. [1]). One of the more successful methods is the so-called generalized ensemble approach [2] that was first applied to protein simulations in Ref. [3]. In the following we will present a short review of this approach and demonstrate its usefulness for both structure prediction and folding studies of proteins. We concentrate on two examples. First, we study the helix–coil transition in homopolymers and study the relation between helix formation and folding for a simple artificial peptide, Ala$_{10}$–Gly$_{5}$–Ala$_{10}$. Secondly, we have chosen the more complicated 36-residue villin headpiece peptide HP-36 to demonstrate usefulness of our techniques for structure prediction of proteins.

2. Generalized-ensemble techniques

A generalized-ensemble simulation is characterized by the condition that a Monte Carlo or molecular dynamics simulation will lead to a uniform distribution of a pre-chosen physical quantity. We will discuss this idea for a prominent example, multi-canonical sampling [4], where the weights $w(E)$ are chosen such that the distribution of energies $P(E)$ is given by

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

($n(E)$ is the spectral density). Hence, a simulation with such weights generates a 1D random walk in the energy space, allowing itself to escape from any local minimum. Since a large range of energies are sampled, one can calculate the thermodynamic average of any physical quantity $A$ by the re-weighting technique [5]:

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

Here, $x$ stands for configurations, $w(x)$ is the generalized-ensemble weight of configuration $x$ and $k_B$ the Boltzmann constant. We remark that extensions of this idea to ensembles that lead to distributions in more than one variable are straightforward [6], as are combinations with annealing techniques [7].

However, unlike in the canonical ensemble, the weights are not a priori known for simulations in generalized ensembles. Instead estimators for these weights have to be determined. This is often done by an iterative procedure described in detail in Refs. [8,9] and can require up to 50% of the available computer time.

Some attempts were made to construct generalized ensembles where the estimators can be written in a simple functional form or where the weights are even a priori known. One example is the weight proposed in Ref. [10]. In this method, one uses that the probability distribution of energy in protein simulations should resemble that of an ideal low-temperature canonical distribution, but with a tail to higher energies. In this way, not only the low-energy region can be sampled efficiently but the simulation can also overcome energy barriers and escape from local minima. Such an ensemble can be obtained if configurations are updated according to the following probability
weight:

\[ w(E) = \left(1 + \frac{\beta(E - E_0)}{n_F}\right)^{-n_F}, \]

(3)

where \( E_0 \) is an estimator for the ground-state energy and \( n_F \) the number of degrees of freedom of the system. Note that this weight can be understood as a special case of the weights used in Tsallis generalized mechanics formalism [11] (the Tsallis parameter \( q \) is chosen as \( q = 1 + 1/n_F \)). The weight reduces in the low-energy region to the canonical Boltzmann weight \( \exp(-\beta E) \) for \( \beta(E - E_0)/n_F \ll 1 \). On the other hand, high-energy regions are no longer exponentially suppressed but only according to a power law, which enhances excursions to high-energy regions.

Another ensemble with similar properties was proposed in Ref. [13]. Here, conformations enter with a weight \( w(E) = \exp(f(E)/k_BT) \) where \( f(E) \) is a non-linear transformation of the potential energy onto the interval \([0, 1]\) and \( T \) a low temperature. The physical idea behind such an approach is to allow the system to “tunnel” through energy barriers in the potential energy surface [12]. Such a transformation can be realized by \( f(E) = e^{-(E-E_0)/n_F} \), where \( E_0 \) is again an estimate for the ground state and \( n_F \) the number of degrees of freedom of the system. Note that the location of all minima is preserved. Hence, at a given low temperature \( T \), the simulation can pass through energy barriers of arbitrary height, while the low energy region is still well resolved. An exploratory study on the efficiency of this algorithm for protein-folding simulations can be found in Ref. [13].

In both ensembles a broad range of energies is sampled. Hence, one can use again reweighting techniques [5] to calculate thermodynamic quantities over a large range of temperatures. In contrast to other generalized-ensemble techniques the weights are explicitly given for both new ensembles. One needs only to find an estimator for the ground-state energy \( E_0 \) which is easier than the determination of weights for other generalized ensembles.

A similar approach, energy landscape paving (ELP) [14], was recently proposed for the purpose of global optimization. In this approach, one performs low-temperature Monte Carlo (MC) simulations with a modified energy expression designed to steer the search away from regions that have already been explored:

\[ w(\tilde{E}) = e^{-\tilde{E}/k_BT} \quad \text{with} \quad \tilde{E} = E + f(H(q,t)) \]

(4)

and \( T \) a (low) temperature. \( \tilde{E} \) serves as an 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” \( q \). The histogram is updated at each MC step, hence the “time” dependence of \( H(q,t) \). As a result, the search process keeps track of the number of prior explorations of a particular region in order parameter space and biases against revisiting the same types of states. Rather than using the system states themselves in the histograms an appropriate order parameter is employed. This may be a “natural” quantity for the system under study or the energy itself. It follows that within ELP the weight of a local minimum state decreases with the time the system stays in that minimum, and consequently the probability to escape the minimum increases. With time, ELP deforms the energy landscape locally in such way that 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. Obviously, for \( f(H(q,t)) = f(H(q)) \) the method reduces to the various generalized-ensemble methods [2].

Even within the generalized-ensemble approach simulations of proteins can still be hampered by large correlations between the sampled conformations. This obstacle can be overcome by improved updates that allow a much faster sampling. One example, parallel tempering [15], proved especially valuable in generalized-ensemble simulations and was first introduced to protein simulations in Ref. [16] and studied there for both Monte Carlo and molecular dynamics. In this approach one considers an artificial system built up of \( N \) non-interacting copies of the molecule, each at a different temperature \( T_i \). In addition to standard Monte Carlo or molecular dynamics moves which effect only one copy, parallel tempering introduces now a new global update [15]: the exchange of conformations between two copies \( i \) and \( j = i + 1 \) which is accepted or rejected according to the Metropolis criterion with probability

\[
    w(C^{\text{old}} \rightarrow C^{\text{new}}) = \min(1, \exp(-\beta_iE(C_j) - \beta_jE(C_i) + \beta_iE(C_i) + \beta_jE(C_j))) .
\]

It is obvious that through the exchange of conformations the Markov chain converges at low temperatures much faster towards the stationary distribution than it does in the case of a regular canonical simulation with only local moves. Note that the parallel tempering technique does not require Boltzmann weights. The method works with any set of weights and parallel tempering can therefore easily be combined with generalized-ensemble simulations. This speed-up was first demonstrated in Ref. [16] and recent interesting applications of this idea can be found in Ref. [17].

3. Helix–coil transition and folding

In the following we want to demonstrate that generalized-ensemble techniques are well suited both for research in the physics of folding and for the prediction of the biologically active state of proteins. We start by investigating a common structural transition in proteins, the formation of \( \alpha \)-helices. It is long known that \( \alpha \)-helices undergo a sharp transition towards a random coil state when the temperature is increased. The characteristics of this so-called helix–coil transition have been studied extensively [18]. In Refs. [19] evidence was presented that the helix–coil transition in polyalanine exhibits a true thermodynamic phase transition when interactions between all atoms in the molecule are taken into account. However, these later results were obtained from gas-phase simulations of poly-alanine and the question remains how these results relate to the biologically more relevant case of solvated molecules.

For this reason, we have investigated now how the characteristics of helix–coil transition change with the details of the solvation term [9]. We have performed multicanonical simulations of polyalanine molecules of length 10 using a detailed all-atom representation of that peptide. Our results were estimated from a multicanonical run.
of 1,000,000 Monte Carlo sweeps. Additional 200,000 sweeps were needed for the weight factor calculation. The interaction between the atoms is described by a standard force field, ECEPP/2 [21] (as implemented in the program package SMMP [20]).

The interactions between our homo-oligomer and water are approximated by means of two implicit water models. In the first model (DDE) the electrostatic interactions in the presence of water rely on a distance dependent electrostatic permittivity [22]: \( \varepsilon(r) = D - (D - 2) / [(sr)^2 + 2sr + 2] e^{-sr} \). For the parameters \( D \) and \( s \) empirical values are chosen such that for large distances the permittivity takes the value of bulk water (\( \varepsilon \approx 80 \)), and the value \( \varepsilon = 2 \) for short distances (protein interior space). In the second solvent model, 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: \( E_{sol} = \sum_i \sigma_i A_i \), where \( E_{sol} \) is the solvation energy, \( A_i \) is the conformational dependent solvent accessible area of the surface of the \( i \)th atom and \( \sigma_i \) is the atomic solvation parameter for the atom \( i \). The sets of solvation parameter we study here are named by us OONS [23], JRF [24], W92 [25] and SCH [26], and are described in the respective references.

In previous gas-phase simulations of poly-alanine [7,19] we observed at \( T = 430 \) K a pronounced transition between a high-temperature phase dominated by disordered coil structures and an ordered phase with single, extended helices. A natural order parameter for this helix–coil transition is the average number \( \langle n_H(T) \rangle \) of residues in the oligomer that are part of an \( \alpha \)-helix. In Fig. 1 this order parameter is displayed as
function of temperature for a gas-phase simulation (GP) of Ala_{10} and simulations with the various solvation terms. The curves, representing the various simulations, fall into three groups. For the case where the protein–solute interaction was approximated by a distance-dependent permittivity (DDE), both $\langle n_H \rangle$ and $\chi$ have a similar temperature dependence as is observed in GP. However, the transition temperature $T_c$ is shifted from $T = 435 \pm 20$ K (gas-phase) to a higher value $T = 495 \pm 20$ K. To the same group belong the simulations in which the solvation energy was approximated by a solvent accessible surface term with either the OONS [23] or SCH [26] parameter set. The order parameter $\langle n_H \rangle$ shows again a temperature dependence similar to the one of gas-phase simulations. However, the transition temperature $T_c$ is shifted to lower temperatures. The shift towards lower temperatures was one of the main results reported in Refs. [27] for simulations with the OONS solvation energy, and our $T_c = 345 \pm 20$ K agrees well with their value $T_c = 340$ K (no errors quoted).

A somehow different behavior is observed in the simulation relying on the W92 [25] parameter set. Here, the form of $\langle n_H \rangle$ indicates only partial helix formation that occurs at much lower temperatures. No indication for a helix–coil transition is found. Yet another behavior is observed in simulations where the solvation energy is evaluated by means of the JRF parameter set. For this parameter set no formation of helices is observed in Fig. 1.

The above results indicate that the existence and characteristics of the helix–coil transition in polyalanine depend strongly on the details of the solvent representation. A detailed study of this relation can be found in Ref. [9]. While our results demonstrate both the importance of including solvation terms into protein simulations and the difficulties in choosing an adequate representation of the protein–water interactions, they also show that key elements of folding such as the formation of secondary structure elements (here $\alpha$-helices) can be studied in computer experiments. We have extended now our investigation in Ref. [28] to the relation between helix formation and folding. For this purpose we have simulated an artificial peptide, Ala_{10}–Gly_{5}–Ala_{10}. We started with simulations where we did not include explicitly the interaction of the peptide with the solvent into our simulations and set the dielectric constant $\epsilon$ equal to 2. Our results rely again on multicanonical simulations with large statistics: all thermodynamic quantities were estimated from one production run of 8,000,000 Monte Carlo sweeps. Estimators for the multicanonical weights were determined in 500,000 sweeps. Since polyalanine has a pronounced helix–coil transition in gas-phase, we expect formation of $\alpha$-helices in our peptide, and the average number of helical residues $\langle n_H \rangle$ is therefore one of the quantities that we have measured. $\langle n_H \rangle$ is displayed in Fig. 2 as a function of temperature, and we observe in this plot two temperature regions. At high temperature, few residues are found with backbone dihedral angles ($\phi, \psi$) typical for an $\alpha$-helix. On the other hand, at low temperatures we observe helix-formation, and almost all of the alanine residues are part of an $\alpha$-helix. The transition between the two temperature regions is sharp indicating the existence of a helix–coil transition.

We could show in earlier work [7,9] that the formation of $\alpha$-helices in polyalanine is related to a gain in potential energy. A pronounced change in energy with temperature corresponds to a peak in the specific heat. Such a peak can be indeed observed in
Fig. 2. The average helicity $\langle n_H \rangle (T)$ (left axis) and average end-to-end distance $\langle d_{e-e} \rangle (T)$ (right axis) of Ala$_{10}$–Gly$_5$–Ala$_{10}$ in gas phase as a function of temperature $T$.

Fig. 3 at a temperature $T = 480 \pm 10$ K. However, we find in addition a second, smaller peak in the specific heat at the lower temperature $T_f = 265 \pm 7$ K indicating yet another transition. In order to understand this second peak, we plot in Fig. 2 also the average end-to-end distance $\langle d_{e-e} \rangle_T$ as a function of temperature. This quantity is a measure for the compactness of a protein conformation and defined here by the distance between N of Ala$_1$ and O of Ala$_{25}$. We observe that this quantity decreases with decreasing temperature. Below the helix–coil transition $T_{hc}$ the decrease slows down and the curve becomes almost flat at a value of $\langle d_{e-e} \rangle \approx 10$ Å indicating that there is little further change in the compactness of the molecule. However, at temperature $T_f$ the end-to-end distance decreases again sharply towards a new value $\langle d_{e-e} \rangle = 6.1$ Å. Hence, $T_f$ marks the folding of the molecule into a defined compact structure with the two terminal ends of the peptide close together. This scenario is supported by Fig. 4 in which we display the configuration with lowest energy ever found in our multicanonical simulation of 8,000,000 sweeps. It consists of two helixes (made up out of the alanine residues) connected by a turn (build out of the flexible glycine residues) towards a U-turn-like structure that is consistent with the small value of the end-to-end distance $d_{e-e}$ observed in Fig. 2 for temperatures below $T_f$.

Our above analysis of the thermodynamics of our peptide suggests that in gas-phase, Ala$_{10}$–Gly$_5$–Ala$_{10}$ folds in a 2 step process. The first step is the formation of $\alpha$-helices and can be characterized by a helix–coil transition temperature $T_{hc} = 485 \pm 5$ K. The formation of $\alpha$-helices then restricts the possible configuration space. Energetically most
Fig. 3. The specific heat $C(T)$ of $\text{Ala}_{10} - \text{Gly}_5 - \text{Ala}_{10}$ in gas-phase as a function of temperature $T$.

Fig. 4. Lowest-energy conformation of $\text{Ala}_{10} - \text{Gly}_5 - \text{Ala}_{10}$ in gas-phase as found in our multicanonical simulations.
favorable is the folding of two \( \alpha \)-helices (made out of the alanine residues) into a hairpin. This second step can be characterized by a lower folding temperature \( T_f = 265 \pm 7 \) K.

4. Structure prediction by generalized-ensemble techniques

Generalized-ensemble techniques are also well suited for the prediction of the native state of proteins. This is demonstrated here for the example of the villin head-piece subdomain, a 36-residue peptide (HP-36). HP-36 is one of the smallest peptides that can fold autonomously and it was chosen recently by Duan and Kollman for a 1-\( \mu \)s molecular dynamics simulation of protein folding [29]. The experimental structure was determined by NMR analyses [30]. Since it is a solvated molecule we had to take into account the interaction between protein and solvent. We have again used the solvent accessible surface approach to approximate this contribution to the overall energy. The parameters \( \sigma_i \) were chosen from Ref. [25]. Our simulations rely on the generalized-ensemble technique described in Ref. [14].

The structure of HP-36 as obtained from the Protein Data Bank (PDB code 1vii) is shown in Fig. 5. The figure was created with RasMol [32]. The structure consists

![Fig. 5. Experimental structure of HP-36 as deposited in the PDB data bank.](image-url)
of three helices between residues 4–8, 15–18, and 23–32, respectively, which are connected by a loop and a turn. After regularizing this structure with the program FANTOM [31] we obtained as its energy (ECEPP/2+solvation term) $E_{\text{nat}} = -276$ kcal/mol. Our approach led to a configuration with the lowest energy $E_{\text{min}} = -277$ kcal/mol which we show in Fig. 6 [14]. The above structure has a radius of gyration $R_g = 10.1$ Å which indicates that the numerically obtained structure is slightly less compact than the experimental structure ($R_g = 9.6$ Å). It consists of three helices where the first helix stretches from residue 2 to 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 65% of the native contacts were formed in our structure. Both values are comparable with the results in Ref. [29] (but required orders of magnitude less computer time) where the optimal structure of a 1 μs 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 Å versus 5.7 Å in Ref. [29] when all backbone atoms were counted. We conclude that even for large peptides such as HP-36 our generalized-ensemble method is able to find structures that are close to the experimentally determined structures.

Fig. 6. Lowest energy structure of HP-36 as obtained in our computer simulation with a solvation term.
5. Conclusion

We gave a brief introduction into generalized-ensemble techniques and their application to the protein folding problem. Using one of these techniques we were able to study the relation between helix-formation and folding in homopolymers and an artificial peptide, and to demonstrate the effect of protein–solvent interaction on this relation. Generalized-ensemble simulations also allowed us to find in an unbiased simulation the correct structure of medium-sized peptides. Both examples demonstrate that generalized-ensemble algorithms are well-suited for investigations of the thermodynamics of proteins and prediction of their structure and may lead to an increased understanding of the protein folding problem.

Acknowledgements

Parts of the results presented in this article are also published in Refs. [28,14] and resulted from collaborations with N.A. Alves (FFCLRP, University of Sao Paulo, Brazil) and L.T. Wille (Florida Atlantic University, Boca Raton, FL). Financial supports from a research grant (CHE-9981874) of the National Science Foundation (USA) is gratefully acknowledged.

References