All-atom generalized-ensemble simulations of small proteins

Brian S. Kinnear a, Martin F. Jarrold b, Ulrich H.E. Hansmann c,∗

a Portland Technology Development, Intel Corp., Hillsboro, OR 97124, USA
b Chemistry Department, Indiana University, Bloomington, IN 47405, USA
c Department of Physics, Michigan Technological University, Houghton, MI 49931, USA

Abstract

We give an overview of some generalized-ensemble techniques that have proven successful in all-atom simulations of proteins. We show that these techniques enable efficient investigations of secondary structure formation and folding in peptides and small proteins. Results are presented for various alanine-based artificial peptides and a small protein, the 36-residued villin headpiece subdomain (HP-36). Our results indicate that all-atom simulations of proteins may be more restricted by the accuracy of the present energy functions than by the efficiency of the search algorithms.

© 2003 Elsevier Inc. All rights reserved.

Keywords: Generalized-ensemble simulations; Proteins; HP-36

1. Introduction

The successful deciphering of whole genomes has aggravated an old challenge in protein science: for most of the resolved protein sequences one does not know the corresponding structures and functions. Since proteins are only functional if they fold into their specific shape, and misfolded proteins can even cause a variety of diseases, it is important to understand how the structure and function of proteins emerge from their sequence of amino acids. 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.

A number of novel simulation techniques have been developed that promise to alleviate the above-stated multiple-minima problem (for a review, see [1]). One successful method is the so-called generalized-ensemble approach [2] that was first applied to protein simulations in [3]. In the following we will present a short review of this 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

The complex form of the intramolecular forces and of the interaction with the solvent, containing both repulsive and attractive terms, leads for all-atom models of proteins to a very rough energy landscape with a huge number of local minima. Hence, a typical thermal energy of the order $k_BT$ is much less than the energy barriers that the protein has to overcome in the low-temperature region. This is because the probability to cross an energy barrier is proportional to $\exp(-\Delta E/k_BT)$ where $\Delta E$ is the heights of the energy barrier. Hence, simple canonical Monte Carlo or molecular dynamics simulations will get trapped in a local minimum.

The essence of all generalized-ensemble simulations is to replace the canonical weights by such that do allow escape from any local minimum. From this simulation, one
can calculate the thermodynamic average of any physical quantity $A$ by re-weighting [5]

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

Here, $x$ stands for configurations and $k_B$ is the Boltzmann constant. The weight $w(E)$ is not a priori known in generalized ensembles, and estimators have to be determined by an iterative procedure described in [4,6]. We remark that if it is straightforward to define ensembles that lead to flat distributions in more than one variable [7].

Even within the generalized-ensemble approach simulations of proteins can still be hampered by large correlations between the sampled conformations. Improved updates are one way to overcome this obstacle. One example is parallel tempering. The method is also known as replica exchange or Multiple Markov Chain method [8], and was first introduced to protein folding in [9]. In its most common form, one considers in this technique 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 that effect only one copy, parallel tempering introduces now a new global update: the exchange of conformations between two copies $i$ and $j = i + 1$ with probability

$$w(C^{\text{old}} \rightarrow C^{\text{new}}) = \exp(-\beta_i E(C_j) - \beta_j E(C_i) + \beta_i E(C_i) + \beta_j E(C_j)).$$

The exchange of conformations will lead especially at low temperatures to a faster convergence of the Markov chain toward the stationary distribution than is observed in regular canonical simulations with only local moves. 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 [9].

The emphasis in today's computational protein studies is on the prediction of the three-dimensional shape in which a protein is biologically active. Most proteins are at room temperature thermodynamically stable, i.e. the biologically active configuration is the global minimum in free energy at $T = 300 \text{K}$. Since this state is unique, one can identify the free-energy global minimum with the lowest potential energy conformation. Hence, structure prediction of proteins can be considered as a global optimization problem.

A general characteristic of successful optimization techniques is that they avoid entrapment in local minima and continue to explore the energy landscape for further solutions. Noticing that this also a characteristics of generalized-ensemble algorithms, Wilke and Hansmann proposed recently a new optimization method, energy landscape paving (ELP), that proved very promising in protein studies [10].

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:

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

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. With time, 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. Obviously, for $f(H(q, t)) = f(H(q))$ the method reduces to the various generalized-ensemble methods [2] (for instance for $f(H(q, t)) = \ln H(E)$ to multicanonical sampling).

The efficiency of the new global optimization method was tested and compared with simulated annealing and tabu search in simulations of small and medium-sized peptides [10]. ELP was also applied to the problem of crystal structure prediction from X-ray diffraction data [11].

3. Helix formation for non-polar amino acids

In the following, we want to demonstrate that generalized-ensemble techniques are well suited for protein research. While our understanding of the folding physics has increased considerably over the last few years, many questions remain unsolved. One example is the role of secondary structure formation in the folding process, and we focus here on the formation of $\alpha$-helices.

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 [12]. Helix formation was also researched in [13] for polyalanine, polyvaline and polglycine chains of chain length $N = 10$. The computational results of the latter work indicate that alanine has a strong propensity to form helices in gas phase while glycine shows no such tendency. Helix formation was observed for valine, but is much weaker than for alanine. However, recent experimental results of alanine and valine based peptides rather indicate that valine has in gas phase a higher helix propensity than alanine [14]. For this reason, we have repeated the simulations of [13] and extended them to homopolymer chains with up to 20 monomers.

Our investigation of helix formation for these three homopolymers is based on a detailed, all-atom representation of these molecules. The interactions between the atoms are...
described by a standard force field, ECEPP/2 [15], as implemented in the program package SMMP [16]:

\[ E_{\text{ECEPP/2}} = E_C + E_{\text{LJ}} + E_{\text{HB}} + E_{\text{freq}}. \]  

\[ E_C = \sum_{i,j} 32\epsilon_{ij} q_{ij} \left( \frac{r_{ij}}{\xi} \right)^6. \]  

\[ E_{\text{LJ}} = \sum_{i,j} \left( \frac{A_{ij}}{r_{ij}} - \frac{B_{ij}}{r_{ij}^{12}} \right). \]  

\[ E_{\text{HB}} = \sum_{i,j} \sum_{\alpha,\beta} C_{\alpha\beta} \left( \frac{r_{ij}}{\xi} \right)^{12} - \left( \frac{r_{ij}}{\xi} \right)^6. \]  

\[ E_{\text{freq}} = \sum_{i} U_i (1 \pm \cos(n_i \phi_i)). \]  

Here, \( r_{ij} \) (in Å) is the distance between the atoms \( i \) and \( j \), and \( \phi_i \) is the \( i \)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.

Our results rely on multicanonical simulations of 2,000,000 sweeps for Ala\(_{10}\), Gly\(_{10}\) and Val\(_{10}\), and of 4,000,000 sweeps for Ala\(_{20}\), Gly\(_{20}\) and Val\(_{20}\). The simulations start with a random configuration and the first 10,000 sweeps are discarded to allow for equilibration. One of the quantities that we monitor in our simulations is \( n_H \) (in H/\( \% \)) backbone hydrogen bonding that is characteristic for \( \alpha \)-helices.

We have displayed \( n_H \) in Fig. 1a–c for polyvaline, polyalanine and polyglycine, respectively. Data for chains of length \( N = 10 \) and 20 are plotted. The corresponding values for the specific heat

\[ C(T) = \frac{\langle (E^2) \rangle(T) - \langle E \rangle^2(T)/N}{(k_B T)^2} \]  

are drawn in the inlets. In Fig. 1a and b, we observe for both polyvaline and polyalanine a helix–coil transition that separates a high-temperature phase with disordered coil configurations from a low-temperature phase that is dominated by helical configurations. This transition becomes more pronounced as the chain length increases from \( N = 10 \) to 20. The corresponding transition temperatures are determined from the location of the peaks in the specific heat shown in the inlets. For polyalanine the helix–coil transition temperature increases from \( T_c = 423(20) \) K for \( N = 10 \) to \( T_c = 510(10) \) K for \( N = 20 \). The corresponding temperatures for polyvaline are lower: \( T_c = 405(25) \) K for \( N = 10 \) and \( T_c = 435(15) \) K for \( N = 20 \). On the other hand, no helix formation is observed for polyglycine. The peaks in the specific heat at \( T_c = 360(30) \) K (\( N = 10 \)) and \( T_c = 330(30) \) K (\( N = 20 \)) rather correspond to a collapse transition associated with a sharp decrease in the radius of gyration (data not shown).

Fig. 1. Helicity \( q_{\text{H}} \) of (a) polyvaline, (b) polyalanine and (c) polyglycine as function of temperature. The corresponding values for the specific heat \( C(T) \) are shown in the inlets.

Table 1 lists for the three homopolymers the differences in energy \( \Delta E \), free energy \( \Delta G \) and entropy \( T \Delta S \) between...
We conclude that our data reaffirm qualitatively the previous results of [13] in that we find a strong helix forming tendency for polyalanine, no helix propensity for polyglycine and intermediate behavior for polyvaline. This ranking in helix propensities is different from the one reported in [14] where experimental results for homopolymers of the three amino acids are reported. Measurements of the three homopolymers indicate that in gas phase valine is a stronger helix former than alanine while again no helix formation is observed for glycine [14].

We remark that the ranking of helix propensities for the three types of homopolymers as calculated from our computer simulations is in agreement with the one measured in experiments on the solvated molecules. Hence, one possible explanation for the disagreement between our computational data and the results from the gas phase experiments in [14] is that the ECEPP/2 force field does not describe accurately molecules in gas phase. However, we have currently not sufficient data to test this conjecture. Another possible explanation for the discrepancy is that our computer simulations do not describe correctly the experiments that required partial charges on the ends of the homopolymers. Such charged polymers cannot be studied with the current version of SMMP, and the different charge distribution maybe responsible for the differences in our results.

### Table 1

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>N = 10</th>
<th></th>
<th>N = 20</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ΔG</td>
<td>ΔE</td>
<td>TΔS</td>
<td></td>
<td>ΔG</td>
<td>ΔE</td>
<td>TΔS</td>
</tr>
<tr>
<td>Alanine</td>
<td>−4.1 (2)</td>
<td>−11.8 (8)</td>
<td>−7.7 (8)</td>
<td>−13.0 (1.2)</td>
<td>−39.2 (2.8)</td>
<td>−7.2 (1.6)</td>
<td></td>
</tr>
<tr>
<td>Valine</td>
<td>−1.1 (7)</td>
<td>−6.4 (1.2)</td>
<td>−5.3 (1.5)</td>
<td>−5.6 (1.6)</td>
<td>−16.6 (5.1)</td>
<td>−11.0 (6.1)</td>
<td></td>
</tr>
<tr>
<td>Glycine</td>
<td>1.5 (1)</td>
<td>−2.6 (1)</td>
<td>−3.6 (2)</td>
<td>2.0 (7)</td>
<td>−3.3 (1.6)</td>
<td>−5.3 (2.0)</td>
<td></td>
</tr>
</tbody>
</table>

The numbers in parentheses represent errors.
a drop in the average end-to-end distance \( \langle d_{ee} \rangle_T \) as plotted in Fig. 3b. This quantity is a measure for the compactness of a protein conformation and decreases with lowering temperature. Below the helix-coil transition \( T_{hc} \), the decrease slows down and the curve becomes almost flat at a value of \( \langle d_{ee} \rangle \approx 10 \, \text{Å} \) indicating that there is little further change in the compactness of the molecule. However, at temperature \( T_h \) the end-to-end distance decreases again sharply toward a new value \( \langle d_{ee} \rangle = 6.1 \, \text{Å} \). Hence, \( T_h \) 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 our lowest-energy configuration when protein-solvent interactions are neglected and the peptide is simulated in gas phase, the native structure differs has a \( \approx 30 \text{ kcal/mol} \) higher energy than the lowest-energy configuration (Fig. 5c) that was found in this simulation.

In order to understand more the differences between the gas-phase results and that with a solvent accessible surface term, parallel tempering simulations of HP-36 were performed on 20 nodes of a cluster of IBM four-ways 375 MHz SMP Thin Nodes with temperatures \( T = 1000, 900, 800, 700, 610, 560, 530, 510, 495, 485, 475, 465, 450, 420, 390, 360, 330, 275, 250, 230 \) \[21\]. Both gas-phase simulations and such relying on a solvent-accessible surface term with the parameter set of \[20\] were performed.

From these parallel tempering simulations we have calculated the number of helical residues as function of and displayed in Fig. 6. 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 simulation of the solvated peptide while it stays small in gas phase. As one can see from Fig. 7, configurations in the two simulations differ also in their compactness. Here, we show the average radius of gyration \( \langle r_{gy} \rangle_T \) as a function of temperature. The corresponding values for the total number of contacts \( n_{TC} \) are shown in the inlet. Both quantities indicate that configurations in gas phase are substantially more compact than the ones for the solvated molecule. For instance, at \( T = 300 \text{ K} \), we find \( r_{gy} = 9.6(1) \text{ Å} \) in gas phase compared to \( r_{gy} = 9.6(1) \text{ Å} \) in gas phase.
Fig. 5. (a) Experimentally determined structure of HP-36 as deposited in the PDB data bank, (b) lowest-energy configuration of HP-36 as obtained from an ELP simulation of the solvated peptide and (c) lowest-energy configuration of HP-36 as obtained from an ELP simulation of the peptide in gas phase. The figures are taken from [21].

12.5(1) Å in simulations of the solvated molecule. We conjecture that this bias toward compact configurations inhibits the formation of α-helices, and that the low-energy states of HP-36 in gas phase are characterized by large density and low helicity. We also remark that the high transition temperature of \( T = 490 \) K agrees well with some recent findings in simulations of the C-terminus β-hairpin from protein G using the OPLSAA and AMBER force field [24].

Fig. 6. Average number of helical residues \( \langle n_H \rangle (T) \) as function of temperature as obtained from simulations of HP-36 in gas phase and solvated. The figure is taken from [21].

6. Conclusion

We gave a brief introduction into generalized-ensemble techniques and their applications to the protein folding problem. Our examples demonstrate that these techniques are well-suited for investigations into the physics of proteins. With the advent of these and other modern sampling techniques, all-atom simulations of proteins may now be more
restricted by the accuracy of the present energy functions than by the efficiency of the search algorithms.

Acknowledgements

Parts of the results presented in this article are also published in [17,21]. Financial supports from a research grant (CHE-9981874) of the National Science Foundation (USA) is gratefully acknowledged.

References