Parallel Tempering Simulation of a Miniprotein: Trp-Cage

Chia-Lin Lo, Chun-Ping Yu and Hoong-Chien Lee
Department of Physics and Center for Complex Systems
National Central University, Chungli, Taiwan 320

December 22, 2004

Abstract

Parallel tempering combined with molecular dynamics simulations in explicit solvent have been carried out to study the folding thermodynamics of designed 20-residue peptide Trp-Cage. The GROMOS96 43a1 force field is used for the molecular dynamics. The actual computer simulations are performed on the distributive computing facility Protein@CBL. Initial conformations of peptide are extended random coils and the simulation temperature range is from 250 to 478 degrees K. After an accumulated simulation time of $4.16 \mu$s, one peptide sampled a conformations in which its $C_\alpha$ have an rms positional deviation 0.23 nm relative to the native structure of the peptide.

Introduction

Prediction of the structure and function of proteins solely from their amino acids sequences is still an unsurmountable challenge. Attempts to investigate the sequence-structure relationship by means of computer simulations are hampered by the problem that the energy landscape is characterized by a vast number of local minima separated by hard-to-overcome energy barriers. At low temperatures, simple canonical Monte Carlo or molecular dynamics (MD) simulations of realistic protein models will not thermalize within a finite amount of available CPU time, and physical quantities cannot be calculated accurately. A second problem is the reliability of protein models. Experimental evidence suggests that the biologically active state of a protein is in its global minimum in free energy at room temperature (which for sufficiently large molecules can be approximated by the global minimum in potential energy). However, this is not necessarily true for the available energy functions that only approximate the interactions between atoms within a protein and between the protein and surrounding solvent. For this reason, structure prediction of proteins is limited by the accuracy of the energy functions.
A number of novel techniques aiming at optimizing the sampling procedure in protein simulations have been developed over the last few years [1]. For instance, generalized ensemble methods [2] proved to be successful in calculating reliable low-temperature estimates of thermodynamic quantities. Here we use parallel-tempering method [3] to analyze the protein folding processes. The simulated target is a “mini-protein”, the Trp-Cage.

**Trp Cage.** The mini-protein Trp Cage (PDB ID: 1L2Y) [4] has the amino acid sequence NLYIQWLKDGPPSSGRPPPS and is one of the smallest proteins known to date that exhibit stable, spontaneous, two-state folding properties. The folding appears to be cooperative based on the hydrophobic effect of a Tryptophan side chain that is surrounded by Proline rings. This 20-residue mini-protein is a truncated and optimized version of the predominantly helical, 39-residue peptide exendin-4 (EX4) from Gila monster saliva. It is derived from C-terminal fragments of exendin-4 [5]. The backbone structure is shown in Fig. 1. A recent paper by Neidigh et al. [6] describes its design. Several constructs of increasing stability were made, gradually introducing stabilizing features like helical N-capping residues and a solvent-exposed salt bridge. The final 20-residue peptide exhibits a cooperative melting transition with a midpoint of 315 K in aqueous solution at pH 7. It shows a well structured hydrophobic core where the indoor side chain of a Trp residue is buried between the rings of two Pro residues. The small size and stability of this protein make it an ideal choice for simulation studies of protein folding.

Simmerling et al. [7] have reported a blind structure prediction of the TC5b sequence based on a simulated annealing protocol using a combination of implicit and explicit solvent with a modified version of the Assisted Model Building with Energy Refinement (AMBER) force field [8]. Their prediction reproduced the experimental structure to within $\leq 1.0 \ \text{Å}$ $C_\alpha$ rms positional deviation (RMSD). After these experimental and computational characterizations of the folded structure, significant work was done to study the kinetics of folding in this protein. Laser temperature jump relaxation experiments
have shown that it is the most rapidly folding protein known, with a folding
time of $\approx 4.1\mu s$ [9]. These experiments were complemented by simulations of
the folding kinetics that indicated a folding time between 1.5 and 8.7 $\mu s$ [10].
Here we try to use another force field to investigate this mini protein.

## Tools and Methods

In this work, the MD simulation was carried out using the GROMACS package [11, 12], which comes with its own force field. GROMACS allows the effect of solvation to explicitly included in the simulation. Here the explicit solvation model was implemented where the spc216 water structure is applied. We added one chlorine atom to neutralize the solvated protein. There are a total of 2560 atoms in our system, 304 in the peptide and 2256 in the solvation, all of which were periodically constrained in a cubic box with sides 3.0 nm long. Every timestep is set to 2 fs. The cutoff radius was set to be 1.2 nm. The Coulomb interaction was implemented using the generalized-reaction-field algorithm. Our experiment showed that this algorithm can reduce the value of RMSD. The dielectric constant outside the cutoff radius was set to be 78.5. For reducing the temperature fluctuation during simulation, the nose-hoover method was applied to affect temperature coupling. We used three temperature groups for protein, chlorine and solvation. The temperatures of these groups are set to the same bath. The LINGS constraint [13] is used to limit all bonds. This solvated protein system is build with energy minimization for over 200 step in the tolerance of 2000 KJ/mole. For each peptide, an initial position restraining simulation was run for 200 ps at 300 degree K. The resulting conformation was used as the initial conformation for every simulation of the subsequent replica-exchange calculation. We test these setting parameters by simulating the native state for 20 ns. The result is shown in Fig. 8. The RMSD variation is less than 0.3 nm during the run. [note: This paragraph needs to further clarified].

In GROMACS the energy of the simulated system is expressed as

$$E_{\text{tot}} = E_{\text{bond}} + E_{\text{Coul}} + E_{\text{LJ}} + E_{\text{solvation}},$$  \hspace{1cm} (1)

where

$$E_{\text{Coul}}(r_{ij}) = \frac{q_i q_j}{\epsilon r_{ij}}$$  \hspace{1cm} (2)

$$E_{\text{LJ}}(r_{ij}) = c_{ij}^{(12)} r_{ij}^{-12} - c_{ij}^{(6)} r_{ij}^{-6}$$  \hspace{1cm} (3)

In atomic energy calculation, the first and second neighbors (atoms $i+1$ and $i+2$) are therefore excluded from the Lennard-Jones interaction list of atom $i$. For third neighbors the normal Lennard-Jones repulsion is sometimes still too strong, which means that when applied to a molecule the molecule would deform or break due to internal strain. Therefore for some of these interactions the Lennard-Jones repulsion has been reduced in the GROMOS force field, which is implemented by keeping a separate list of 1-4 and normal Lennard-Jones parameters. This so-called “1-4 interaction” is included in Eq. (1).
Much of the computation was executed on our recently built distributive computing facility Protein@CBL, or PAC. PAC enables us to share the CPU resources of a large number of “clients”. Any PC owner can become a client by visiting the website of PAC (http://protein.ncu.edu.tw) and completing a hand-shake protocol, which includes downloading an MD package. A detailed description of PAC will be given elsewhere [14].

**Parallel-tempering.** The energy landscape of proteins in an all-atom description is characterized by a huge number of local minima separated from each other by “high-energy” barriers. The success of a search algorithm to converge to the global energy minimum (assuming that it exists) hinges on the ability of the algorithm to allow the peptide to visit many local minima but not be locked into any one of them. Generalized ensemble methods have been increasingly recognized as a way to overcome the problem of slow convergence. In this study we used the method of parallel tempering [15], also known as replica-exchange method or Multiple Markov chains, which has been shown to have good convergence properties in protein studies [3, 16, 17, 18].

Our simulation data are gathered as follows. For a peptide sequence, simulations of many peptides in possibly different initial states are executed separately and independently (in our case, on the PC’s of different clients) at different temperatures. We call the folding state of a peptide and its temperature collectively a configuration. MD changes the folding state of a peptide but not its temperature. A data manager sends out peptides to clients (independently) for simulation and collects the peptides from the clients after a predetermined run time.

Before a peptide is sent out again for simulation it is subject to one or more “temperature swappings” which may or may not change its temperature (but not its folding state). Two peptides, with energies and temperatures $E_1$, $T_1$ and $E_2$, $T_2$, respectively, are involved in a temperature swapping. The temperatures of the two states are swapped or not according to the swap probability

$$P = \min \left\{ 1, \exp \left[ -\frac{(E_2 - E_1)}{kT_1} - \frac{1}{kT_2} \right] \right\}$$

Here total potential (protein, solvation and chlorine) is included in $E_i$. The purpose of temperature swapping is to enhance the probability of a peptide for crossing an energy barrier in case it is trapped in a local energy minimum.

When the number of peptides involved is large, the data manager may be thought of as a grain elevator. Freshly gathered peptides - after their respective most recent simulation runs - enter the elevator from the top. In the mid section of the elevator the peptides go through temperature swapping. Then they proceed on to the bottom of the elevator and wait there to be sent to separate clients for their respective next simulation runs. Thus, after going through one manager-client cycle a peptide acquires a new configuration. In a complete run that may involve many cycles a history of configurations for each peptide is kept.
Figure 2: Best result for a peptide in a simulation totaling 4000 replica exchanges, for a total simulation time of 4.16 µs. Left: the initial extended conformation (showing backbone only). Middle: the best conformation with RMSD(Cα)=0.23 nm. Right: the native conformation of the 38th model [19].

For the results reported here, we used twenty six different temperatures: 250, 256, 263, 269, 276, 283, 290, 298, 305, 313, 321, 329, 337, 346, 355, 364, 373, 382, 392, 402, 414, 426, 439, 451, 465 and 478 degrees K. The MD has a unit time step of Δt=0.002 ps and each simulation run lasts 40 ps, or 20,000 unit time steps. Data on twenty six peptides (the number of peptides need not necessarily be equal to the number of different temperatures) were collected after 4000 simulation runs, so that each peptide had an aggregate simulation time of 160 ns. The required total clock time was 28 days. This yielded a total of 104,000 conformations for statistical analysis.

Result and Discussion

Fig. 2 shows the conformation (middle figure) closest to the native conformation (right figure) of the Trp Cage obtained in this study from an initial extended random coil conformation (left figure). This conformation belongs to a configuration reached by a peptide after 153 ns of simulation and the temperature of the configuration was 364 K. The best conformation has an RMSD(Cα) value of 0.238 nm, where RMSD(Cα) is defined as

$$\text{RMSD}(C_\alpha) = \left[ \frac{1}{M} \sum_{i=1}^{N} m_i [r_i - r'_i]^2 \right]^{\frac{1}{2}}$$

where $r_i$ and $r'_i$ is the position of the $i^{th}$ $C_\alpha$ atom of the peptide in the simulated and native conformations, respectively. The simulation structures were compared against all 38 NMR structures in the Protein Data Bank file and the RMSD($C_\alpha$) value given above refers to the 38th structure. The initial extended structure has an RMSD greater than 1 nm.

Fig. 3 shows the variation of temperature, potential and RMSD(Cα) for one swap history.
Figure 3: Variables in a replica history. Left: Temperature variation. Middle: Energy variation. Right: the RMSD variation. The best structure (RMSD(Cα)=0.238 nm) is indicated by the black line at the 3814th step.

Figure 4: The RMSD(Cα) distribution among the 104,000 samples for analyzed.

Among the 104,000 conformations we analyzed, the RMSD(Cα) distribution is shown in Fig. 4. We find out most state are located in the area RMSD(Cα) = 0.55 nm - 0.45 nm. It is expectedly lower RMSD(Cα) comes with less structure number.

The normalized RMSD(Cα) distribution is plotted and shown in Fig. 5. It shows no matter how you accumulate simulation data, the distribution will hold the same. And we enlarge the area with low RMSD(Cα). Fit curve shows the distribution at low RMSD(Cα) is exponential decreasing and the slope will become fixed. This provides us a way to find out how much time it should take to get a certain RMSD(Cα). In this research, the estimating time to reach RMSD(Cα)=0.2 nm is 54 µs. And 19 ms for RMSD(Cα)=0.1 nm.

Fig. 6 shows folded population for different accumulative simulation time.
Figure 5: Left: The normalize RMSD distribution for every accumulating simulation time. Right: Enlarage the distribution between RMSD(Cα) 0.2 nm - 0.35 nm, and plotted as a logarithm.

Figure 6: Folded fraction varies with temperature. The definition of “folded” is RMSD(Cα) is less than 0.35 nm. As accumulation data is increasing. The folded population is moving toward low temperature area.

We define this peptide ”folded” if its RMSD(Cα) is less than 0.35 nm. In the begining of replica-exchange, the most folded conformations located at high temperature (400K) area. After replica-exchange is proceeding, folded population is moving toward to room temperature (300K) area. This phenomenon can also been seen under best RMSD(Cα) and corresponding temperature. During 0 1 µs simulation, the lowest RMSD(Cα) is 0.25 nm and corrsponding temperature is 439K. During 3 4 µs simulation, the lowest RMSD(Cα) is 0.23 nm and corrsponding temperature is 364K.

We also calculate the values of heat capacity with constant volume(C_V) as a function of temperatures. The heat capacity value is defined by

$$C_V = \beta^2 (\langle E_{tot}^2 \rangle - \langle E_{tot} \rangle^2)$$

(6)
We only collect conformations that RMSD($C_\alpha$) is less than 0.35 nm. As to this definition of heat capacity, it should be added a factor of $\frac{1}{\sqrt{N}}$, where $N$ is the molecular numbers. But the system composed of solvation and protein is inhomogenous, and exact molecular number cannot be defined. Although the uncertainty is not small, this analyse shows a transition temperature – 414 K. It is close the result of the other force field [20]. Experiment shows this peptide is deconstructed at 305K. This unreasonable melting temperature indicates that gromacs force field need to be improved to fit the real system.

The large standard deviations (shown as error bars) seen in the plots are a reflection of the relatively small size of the data set. These plots suggest that although the best conformations were reached preferably at high temperatures, a majority of the native-like conformations still occurred at the more physical temperatures. The high-temperature preference is a source of concern. Our simulations are carried out in the explicit solvation model where the entire system, including the very large number of water moleculars, is bathed in the same temperature, and we have noticed that the temperatures of solvent has a deciding effect in the energy of the entire system. This contrasts with the case in [20], where the simulations were carried out using the implicit solvation model in which the mean solvation field is independent of temperature.

**Conclusion**

We have run the native state simulation in 300 K for 200 ns. The RMSD varies from 0.15 nm to 0.30 nm. The variation is near our result – 0.23 nm. Parallel-tempering method can accelerate MD simulation and get a reasonable result in acceptable time.
It has been mentioned earlier that most of RMSD(Cα) distribution are located in 0.6 nm 0.4 nm. The reasons we cannot get more structures with low RMSD are the temperature range and replica-exchange times. Setting the temperature range is the key point in replica-exchange method. The interval between two neighbors cannot be too large, otherwise the probability between two states will too small to swap. The suggestion is exponentially increasing the interval. The high temperature area will be a problem either. In this study, We use explicit solvation model. It is always happened the whole system blow up if temperature exceed 500K. This is expected because no protein can survive in such high temperature in real world. Molecular dynamics is not well parameterized in such condition. From the view of replica-exchange, limit of temperature could let the global energy minimum unreachable. Different protein has different energy landscape. We use 478K as out highest temperature and could not fit the protein we invest here.

Two way to solve the system exploding problem. One is use implicit solvation model. The difference between explicit solvation and implicit solvation model is the solvation temperature dependence. Implicit solvation model is not really calculating the distribution of solvation energy. And there is no problem about exploring. Jed W. Pitera [20] has done such research. The other way is reducing time step. Two fs is recommend for normal MD simulation. Reducing time step to 0.5 fs or lower is a way to solve problem. And it also causing a time consuming problem.

Extending replica-exchange much longer is for sure to get better statistical property. It also comes with the time wasting situation. Here we suggest a distributive computing way for further study. The normal replica-exchange is calculating the probability after a short time simulation for two neighbor proteins. They are always compared in the same simulation time. We suggest a modified strategy to fit distributive computing facility. Every conformation we gather will be put into one database. Choosing a suitable conformation
randomly in the database and calculating the probability. Swap could happen in different simulation time. It offers the chance to explore the other area in phase space. More clients participate this system, and more the accumulating data.

References


