Published in final edited form as:

Europhys Lett. 2014 February 1; 105(3): 30008–. doi:10.1209/0295-5075/105/30008.

Multicanonical Molecular Dynamics Simulations of the N-terminal Domain of Protein L9

Fatih Ya ar^{1,(a)}, Ping Jiang^{2,(b)}, and Ulrich H. E. Hansmann^{2,(c)}

¹Department of Physics Engineering, Hacettepe University, Ankara, 06800, TURKEY

²Department of Chemistry & Biochemistry, University of Oklahoma, Norman, OK 73019-5251, USA

Abstract

We describe multicanonical molecular dynamic simulations of the N-terminal domain of the protein L9. Analyzing free energy landscapes and thermal ordering, we propose a possible folding mechanism for the protein. By comparing our results with that of molecular dynamics runs of the protein at constant temperature, we find that multicanonical molecular dynamics leads to orders of magnitude higher sampling of folding transitions.

Introduction

While Shaw and co-workers [1] have demonstrated that it is possible to study reversible folding of small proteins in atomistic detail at the time scale observed in experiments, biomolecular processes still often cover time scales which exceed the ones achievable in atomistic simulations. This is not only because the work of the Shaw group relies on specialized hardware and an extent of CPU usage that is out of reach for most academic institutions, but most of all because the computational costs grow exponentially with the size of the protein or protein complex. This scaling can be reduced to a power law by using replica exchange sampling [2, 3], also known as parallel tempering [4], which was first introduced to protein simulations in Ref. 5. However, while the use of this technique has become ubiquitous, its usefulness deteriorates for first-order like transitions. We have recently demonstrated [6] that in cases where folding/unfolding transitions are rare events, techniques such as multicanonical sampling may be more appropriate. For Go-model simulations of the 75 residue protein MNK6 and all-atom simulations of the 36 residue DS119 we show that multicanonical molecular dynamics lead to improvements of factors ≈ 30 over that of replica exchange molecular dynamics in the sampling of folding/unfolding transitions.

Motivated by these improvement factors we have decided to use multicanonical molecular dynamics in a study of a protein with a more complex folding mechanism. Extending our series of investigations into the folding of proteins with end-to-end β -sheets [6,7] we choose as our test system the 39-residue N-terminal domain of ribosomal protein L9 (Protein Data

^(a)fatih@hacettepe.edu.tr. ^(b)pingj@ou.edu. ^(c)uhansmann@ou.edu.

Bank Identifier 2HBB); as this protein allows comparison of our work with that of other groups that used different sampling techniques [8].

The L9 protein from *Bacillus stearothermophilus* binds to specific sequences of naked ribosomal RNA and to other binding proteins of the ribosome [9,10]. The structure of the 149 amino-acid protein has been determined by X-ray crystallography and NMR spectroscopy (Protein Data Bank Identifier 1DIV) [11]. A 39-residue N-terminal domain of the protein, referred to as NTL9 (1-39) is able to fold by itself assuming the same structure as it does in the full-length protein [12–15]. This structure (PDB-identifier 2HBB) is shown in Fig. 1. It is made of a three-stranded anti-parallel β -sheet and an α -helix. The first strand reaches from residue 1 to residue 6 and is connected by a lysine-glycine-rich loop to the second strand consisting of residues 17 to 22. This β -hairpin is followed by an α -helix (residues 27 to 34) and the C-terminal strand (residues 35-39) that completes the three-stranded β -sheet. Folding of the protein happens on the millisecond time scale and has been studied previously by the Pande group [8] at 300 K, 330 K, 370 K and 450 K using the Amber force field 96 [16] and the Generalized Born Surface Area (GBSA) implicit solvent model of Onufriev, Bashford and Case [17]. The purpose of the present paper is to compare for this protein the efficiency of multicanonical molecular dynamics with that of regular (fixed temperature) molecular dynamics, and to use our sampling technique to probe the folding mechanism of this protein.

Methods

Multicanonical molecular dynamics

Multicanonical sampling [18, 19] replaces Boltzmann weights $e^{E(x)/RT}$, where x labels the configurations and R is the gas constant, with weights $w(E)$ designed to lead to a distribution

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

with $n(E)$ the density of states. Such a multicanonical simulation leads to a random walk in energy, and therefore can overcome energy barriers easily. This allows one to calculate thermodynamic averages of any physical quantity $\mathcal{A}$ at a temperature T by re-weighting [20,21]:

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

Application of the method has been limited by the need to determine iteratively estimators of the weights $w(E)$ [18,19]. We have shown recently [6] that the additional workload may be worthwhile for proteins with strong cooperative transitions. Another problem is that the implementation for molecular dynamics [21, 22] is less straightforward than that for Monte Carlo sampling [19,23]. Hence, applications of multicanonical molecular dynamics in protein studies are few (see Ref. 24 and references therein).

We have described the implementation of multicanonical sampling in molecular dynamics simulations in detail in Ref. 6, but will summarize it again here for convenience. In the context of molecular dynamics it is convenient to rewrite the multicanonical weights as

$$w_{muca}(E(x)) = e^{-E_{muca}(x)/R\hat{T}}, \quad (3)$$

where $\hat{T}$ is a pre-chosen thermostat temperature. The “effective” energy E_{muca} is calculated iteratively by

$$E_{muca}^{(n)} = E_{muca}^{(n-1)} + R\hat{T} \ln P^{(n)}(E) \quad (4)$$

with $E_{muca}^{(0)} = E$, the “physical” potential energy of a given configuration, and $P^n(E)$ the measured histogram of energies in the n -th iteration. With this definition, the forces are written as

$$\vec{F}_{muca}(\vec{r}) = -\nabla E_{muca}(\vec{r}) \quad (5)$$

$$= -\nabla E(\vec{r}) \frac{\partial E_{muca}}{\partial E} \quad (6)$$

$$= \vec{F}_{can}(\vec{r}) \Lambda(E) \quad \text{with} \quad \Lambda(E) = \frac{\partial E_{muca}}{\partial E}, \quad (7)$$

where ∇ is the gradient operator and $\vec{r}$ marks the positions of all atoms in the system.

Hence, the regular forces $\vec{F}_{can}(\vec{r})$ (that one would integrate in constant temperature simulations) are scaled in multicanonical molecular dynamics with an energy-dependent factor $\Lambda(E)$, calculated iteratively according to

$$\Lambda^{(n)}(E) = \Lambda^{(n-1)}(E) + R\hat{T} \frac{d \ln P^{(n)}(E)}{dE}, \quad \text{with} \quad \Lambda^{(0)}(E) = 1. \quad (8)$$

Note that such scaling of forces is equivalent to scaling all masses by $m_i^{muca} = m_i / \Lambda(E)$, implying for the viscosity:

$$\eta^{muca}(E) = \eta / \sqrt{\Lambda(E)}. \quad (9)$$

With $\Lambda(E) = R\hat{T} \partial S(E) / \partial E = \hat{T} / T(E)$, and assuming that in first approximation the viscosity $\eta \propto 1 / \sqrt{T}$, the effective viscosity $\eta^{muca}(E)$ becomes constant in multicanonical molecular dynamics.

In order to maximize the number of folding/unfolding transitions, one needs to maximize the current j of configurations flowing from a low-energy E_l (where folded structures are expected) to high-energy E_h (where configurations are unfolded), and back. Following Trebst et al [25] we attach a tag $t = 1$ to a configuration when its energy becomes equal or

larger than E_h , and set the tag $t = 0$ when the energy becomes equal or less than E_l . For energies between E_l and E_h , the tag is not changed. Defining now

$$f(E) = \frac{\sum_i^{H(E)} t_i}{H(E)}, \quad (10)$$

where $H(E)$ is the number of times that the system has an energy E , one can calculate the current as

$$j = D(E) P(E) \frac{d f(E)}{d E} \quad (11)$$

with the energy-dependent diffusion coefficient $D(E)$. Maximizing this current leads to a scheme for iteratively optimizing multicanonical weights:

$$\ln w_{mu}^{(n)}(E) = \ln w_{mu}^{(n-1)}(E) + \frac{1}{2} \left(\ln \frac{d f(E)}{d E} - \ln P^{(n)}(E) \right), \quad (12)$$

or in terms of the “multicanonical energies” E_{muca} :

$$E_{muca}^{(n)} = E_{muca}^{(n-1)} + \frac{1}{2} R \hat{T} \left(\ln P^{(n)}(E) - \ln \frac{d f(E)}{d E} \right). \quad (13)$$

Hence, with N iterations to generate a flat distribution and M iterations to optimize further the number of folding/unfolding transitions, we choose as final “multicanonical energies” :

$$E_{muca} = E + R \hat{T} \sum_{n=1}^N \ln P^{(n)}(E) + \frac{1}{2} R \hat{T} \sum_{m=1}^M \left(\ln P^{(m)}(E) - \ln \frac{d f^{(m)}(E)}{d E} \right). \quad (14)$$

From a multicanonical molecular dynamics simulation with forces scaled by the resulting

energy-dependent factor $\Lambda(E) = \frac{\partial E_{muca}}{\partial E}$, one can calculate the average of a physical quantity $\mathcal{A}$ at a temperature T by re-weighting:

$$\langle \mathcal{A} \rangle_T = \frac{\int dx \mathcal{A}(x) e^{E_{muca}(E(x))/R\hat{T}} e^{-E(x)/RT}}{\int dx e^{E_{muca}(E(x))/R\hat{T}} e^{-E(x)/RT}}, \quad (15)$$

where x labels the configurations and $\hat{T}$ is again the thermostat temperature.

Systems and simulation protocol

We use the Amber 96 force field [16] combined with the Generalized-Born implicit solvent model of Onufriev-Bashford-Case (OBC) [17] as this choice allows a comparison with the earlier study of Ref. 8. All simulations are in double precision using GROMACS 4.5.5. [26], with a modified subroutine do force that implements Eq. 7. The force-scale factors (λE) are included through a look-up table in source file sim_util.c, and have to be updated for each iteration. Our time step is 2 fs. Hydrogen atoms are constrained to their bonded heavy atoms by the LINCS algorithm [27] and the twin range cut-offs method is used for VDW and

coulomb energy calculation. The thermostat temperature is controlled by the Nosé–Hoover method [28].

Iterations start with a short replica exchange molecular dynamic (REMD) simulation covering 17 temperatures between 271 K to 648 K. From the so-obtained energy distribution we calculate an initial estimate for the multicanonical weights, that are then updated in a series of iterations using the scheme described above. Our thermostat temperature is $\hat{T}=300$ K. We start with an initial replica exchange molecular dynamics simulations instead of a single high-temperature canonical simulation as this leads to faster convergence. In the final iteration, we used the optimization procedure of Eq. 13 to maximize the transition rate. While this optimization step can be repeated as often as needed, we found that one iteration is sufficient. The so-obtained multicanonical weights are used in three independent sets of long simulations, starting from either folded, unfolded or random configurations, from which all physical quantities at the desired temperatures are calculated by re-weighting [20,21] by means of Eq. 15. Table 1 summarizes the utilized simulation resources.

Results

We start our analysis by showing in Fig. 2 the time evolution of energy in our multicanonical molecular dynamics simulations. As expected, the systems move on random walks between low and high energies, allowing them to find local minima while at the same time avoiding to get trapped. The corresponding plot of the root-mean-square deviation (with the pdb-structure 2HBB as reference) illustrates that this random walk in energy is indeed connected with large structural changes in the protein, i.e. that we observed folding/unfolding events. In order to evaluate the efficiency of multicanonical molecular dynamics for our protein we list in table 2 the number of such folding/unfolding transitions for both multicanonical molecular dynamics and the regular canonical molecular dynamics simulations of the Pande group [8]. These canonical simulations are at temperatures 300 K, 330 K, 370 K and 470 K and rely on the Folding@Home platform. From a total of 3000 trajectories starting from either extended or random coil structures, only 8 trajectories lead to configurations with a C_α root-mean-square deviation smaller than 4 Å to the PDB-structure, i.e. have a folding transition. Hence, by comparing the frequency of folding/unfolding transitions we find that in our case sampling of folding/unfolding transitions in multicanonical molecular dynamics is increased by four orders of magnitude over constant temperature simulations.

Using re-weighting, one can calculate from the multicanonical trajectories thermodynamic averages over a range of temperatures. As an example, we show in Fig. 3 the specific heat as function of temperature. This quantity is a measure for the change of energy with temperature. In the present case, we find two peaks, at $T_H \approx 480$ K, $T_L \approx 230$ K, and a shoulder at $T_S \approx 280$ K, signaling that there are three critical temperatures where the average energy of configurations is sensitive to temperature changes.

The higher temperature, $T_H \approx 480$ K, is associated with a collapse transition and the formation of secondary structure. This can be seen from Fig. 4 where we display the average radius of gyration (a measure for compactness of configurations), and in the inset the

average number of hydrogen bonds which is correlated with secondary structure formation. Both quantities change dramatically at this temperature T_H , but below T_H the size of configurations changes only little with decreasing temperature. On the other hand, the number of hydrogen bonds continues to increase as the temperature is lowered.

Another quantity, the average end-to-end distance shown in Fig. 5, exhibits changes that correlate with both peak temperature T_H and T_L . This quantity is a signal for the overall global ordering of the protein and the completion of the three-stranded β -sheet. While this quantity drops rapidly at T_H , there is a second, smaller, drop at $T_L \approx 230$ K indicating that at this lower temperature there is final arrangement of configurations corresponding to the folding into the native state with a fully formed three-stranded sheet.

The two peak temperatures indicate a separation of the energy scales for collapse and secondary structure formation on one side, and tertiary structure formation on the other side. A more pronounced picture appears when one plots the frequency of configurations with the central helix (defined by hydrogen bonds involving residues 25-28), the N-terminal β -hairpin (defined by hydrogen bonds between residues 2-5 (β_1) and 17-20 (β_2)), and the configurations where the C-terminal β -strand is attached to the N-terminus by hydrogen bonds between residues 2-5 and 36-38 (β_3). Fig. 6 shows all three curves as functions of temperature. The frequency of configurations with the N-terminal β -hairpin grows quickly around T_H , and continues to grow with decreasing temperature. On the other hand, there is no signal for the formation of the central helix at T_H . Instead, the helix seems to appear only below $T \approx 350$ – 400 K, and its growth seems to be correlated with the shoulder in the specific heat curve at $T_S \approx 280$ K. Consistent with the plot of the end-to-end distance in Fig. 5 we find that the frequency of configurations where the N-terminal strand is attached to the C-terminal β -strand also has a signal at T_H . The growth of this quantity with decreasing temperature first plateaus, but later increases again at the lower peak temperature T_L , while at the same temperature the frequency of configurations with $\beta_1 - \beta_2$ decreases. This suggests that the lower peak temperature T_L is associated with rearrangements of the secondary structure elements, and can be identified as a folding temperature.

Hence, from the above described thermal ordering we conjecture a folding mechanism that starts with a hydrophobic collapse of the molecule and parallel formation of the N-terminal β -hairpin with only a loosely attached C-terminal strand. Once the sheet is formed the helix appears. In turn, the helix seems to stabilize the three-stranded sheet. We do not find a signal for the transient caching of the C-terminal residues as part of the central α -helix that was observed by the Pande group [8]. One possibility for the discrepancy is that our assumption of the correspondence between thermal and temporal ordering is not correct. Hence, in order to get a more direct picture of the folding mechanism we have projected the free energy landscape of the protein on a set of suitable reaction coordinates, the end-to-end distance (signaling global ordering), the number of hydrogen bonds (signaling formation of secondary structure), and the number of native contacts (quantifying to what degree the native structure is formed)

Projecting the free energy on the end-to-end distance and the number of hydrogen bonds leads to the landscape in Fig. 7 characterized by two attractors, both with similar number of

hydrogen bonds but differing in their end-to-end distance. This reflects a transition from a state where the secondary structure is essentially formed but the β -sheet not yet completed into the native state. This picture is supported by Fig. 8 which displays a projection of the free energy landscape onto number of native contacts and number of hydrogen bonds. Again, we find a landscape with two minima, both with the same number of hydrogen bonds but a different number of native contacts. Note that the number of hydrogen bonds is slightly larger for the minimum, corresponding to the folded state. The barriers are small between the two structures suggesting easy interconversion. Hence, our landscape plots are consistent with the mechanism from the proposed mechanism that we have derived from its thermal ordering.

Conclusions

We have described multicanonical molecular dynamic simulations of the 39-residue N-terminal domain of protein L9. Comparing our data with earlier work by the Pande Lab [8] which relied on constant temperature molecular dynamics with large statistics we observe that multicanonical molecular dynamics leads to much higher sampling of folding/unfolding events. The analysis of free energy landscapes and of thermal ordering suggests a folding mechanism that starts with a hydrophobic collapse of the molecule and parallel formation of the N-terminal β -hairpin with loosely attached C-terminal β -strand. Once the sheet is formed the helix appears and, in turn, stabilizes the three-stranded sheet. However, unlike the Pande group, we do not observe any signal for transient caching of the C-terminal residues in the central helix. By construction, generalized ensemble techniques such as multicanonical sampling lead to the same thermodynamic averages as regular molecular dynamics. However, their artificial kinetics may lead to short-cuts or alternate paths in the the folding process. Hence, caution must be applied when deriving folding mechanisms from generalized-ensemble simulations as these techniques may not be able to catch all of the details of the folding mechanism.

Acknowledgments

We acknowledge support from the National Institutes of Health (grant number GM62838) and the Hacettepe University Scientific Research Fund under project number 012.D12.602.-001. The simulations were done on the BOOMER cluster of the University of Oklahoma. F.Y. thanks the Department of Chemistry and Biochemistry for kind hospitality during his sabbatical stay at University of Oklahoma.

REFERENCES

- [1]. Klepeis JL, Lindorff-Larsen K, Dror RO, Shaw DE. Curr. Opin. Struct. Biol. 2009; 19:1. [PubMed: 19217770]
- [2]. Hukushima K, Nemoto K. J. Phys. Soc. (Japan). 1996; 65:1604.
- [3]. Geyer GJ, Thompson EA. J. Am. Stat. Assn. 1995; 90:909.
- [4]. Swendsen R, Wang J. Phys. Rev. Lett. 1986; 57:2607. [PubMed: 10033814]
- [5]. Hansmann UHE. Chem. Phys. Lett. 1997; 281:140.
- [6]. Jiang P, Yasar F, Hansmann UHE. J. Chem. Theory Comput. 2013; 9:3816.
- [7]. Jiang P, Hansmann UHE. J. Chem. Theory Comput. 2012; 8:2127. [PubMed: 24039551]
- [8]. Voelz VA, Bowman GR, Beauchamp K, Pande VS. J. Am. Chem. Soc. 2010; 132:1526. [PubMed: 20070076]
- [9]. Lillemoen J, Cameron CS, Hoffman DW. J. Mol. Biol. 1997; 268:482. [PubMed: 9159485]

- [10]. Adamski FM, Atkins JF, Gesteland RF. J. Mol. Biol. 1996; 261:357. [PubMed: 8780779]
- [11]. Hoffman DW, Davies C, Gerchman SE, Kycia JH, Porter SJ, White SW, Ramakrishnan V. EMBO J. 1994; 13:205. [PubMed: 8306963]
- [12]. Kuhlman B, Boice JA, Fairman R, Raleigh DP. Biochemistry. 1998; 37:1025. [PubMed: 9454593]
- [13]. Horng JC, Moroz V, Rigotti DJ, Fairman R, Raleigh DP. Biochemistry. 2002; 45:13360. [PubMed: 12416980]
- [14]. Horng JC, Moroz V, Raleigh DP. J. Mol. Biol. 2003; 326:1261. [PubMed: 12589767]
- [15]. Sezgin TH, Chung J, Patsalo V, Miyake-Stoner SJ, Miller AM, Brewer SH, Mehl RA, Green DF, Raleigh DP, Carrico I. Biochemistry. 2009; 48:9040. [PubMed: 19658436]
- [16]. Wang J, Cieplak P, Kollman PA. J. Comput. Chem. 2000; 21:1049.
- [17]. Onufriev A, Bashford D, Case D. Proteins. 2004; 55:383. [PubMed: 15048829]
- [18]. Berg B, Neuhaus T. Phys. Lett. B. 1991; 267:249.
- [19]. Hansmann UHE, Okamoto Y. J. Comput. Chem. 1993; 14:1333.
- [20]. Ferrenberg AM, Swendsen RH. Phys. Rev. Lett. 1988; 61:2635. [PubMed: 10039183]
- [21]. Hansmann UHE, Okamoto Y, Eisenmenger F. Chem. Phys. Lett. 1996; 259:321.
- [22]. Nakajima N, Nakamura H, Kidera A. J. Phys. Chem. 1997; 817:817.
- [23]. Hansmann UHE, Okamoto Y. Phys. Rev. E. 1996; 54:5863.
- [24]. Higo J, Ikebe J, Kamiya N, Nakamura H. Biophys. Rev. 2012; 4:27. [PubMed: 22347892]
- [25]. Trebst S, Huse D, Troyer M. Phys. Rev. E. 2004; 70:046701.
- [26]. Hess B, Kutzner C, van der Spoel D, Lindahl E. J. Chem. Theory Comput. 2008; 4:435.
- [27]. Hess B. J. Chem. Theory Comput. 2008; 4:116.
- [28]. Nosé S. J. Chem. Phys. 1984; 81:511.

Fig. 1.
The lowest-RMSD structure (green) superimposed on the X-ray structure, PDB-identifier 2HBB (purple).

Fig. 2.

Time evolution of potential energies a) and root-mean-square deviations (RMSD) b) in multicanonical molecular dynamics. The multicanonical trajectories are plotted every 40 data points.

Fig. 3.
Specific heat capacity as a function of temperature.

Fig. 4. Average of radius of gyration (RG) as a function of temperature. The inset displays the average number of hydrogen bonds (N_{HB}) as a function of temperature.

Fig. 5.
Average of end-to-end distance (D_{e-e}) as a function of temperature.

Fig. 6. Percentage of the configurations with central helix, $\beta_1-\beta_2$ and $\beta_1-\beta_3$ sheets, as a function of temperature.

Fig. 7.
Free energy landscape at 300 K using the number of hydrogen bonds and the end-to-end distance as reaction coordinates.

Fig. 8. Free energy landscape at 300 K using the number of hydrogen bonds and the number of native contacts (N_{NC}) as reaction coordinates.

Table 1

Breakdown of computational resources required by multicanonical simulations.

System	REMD	Pre-production	Production	Total
NTL9	425 ns (25 ns $\times$ 17) ¹	1080 ns (20 ns $\times$ 6 $\times$ 9) ²	6.0 μ s (2.0 μ s $\times$ 3) ³	7.5 μ s

¹ simulation length of one replica $\times$ number of replicas.

² simulation length of one multicanonical simulation $\times$ number of parallel runs $\times$ number of pre-production iterations

³ simulation length of one multicanonical simulation $\times$ number of production runs

Table 2

Transition rate comparison between multicanonical and canonical molecular dynamics simulations.

System	Production	Total	Transition events	Transition per μs
MuCa	6.0 μs	7.5 μs	46*	6.1
CMD	> 20 ms	> 20 ms	8	0.0004

* A transition is defined by a round trip from a configuration with backbone root-mean-square deviation larger than 20 Å to smaller than 5 Å, or reverse.