

Calculations of Free Energy Surfaces for Small Proteins and a Protein-RNA Complex Using a Lattice Model Approach[†]

Eunsang Lee and YounJoon Jung*

Department of Chemistry, Seoul National University, Seoul 150-747, Korea

*E-mail : yjjung@snu.ac.kr

Received March 19, 2011, Accepted June 10, 2011

We calculate the free energy surfaces for two small proteins and a protein-RNA complex system by using a lattice model approach. In particular, we employ the Muñoz-Eaton model, which is a native-structure based statistical mechanical model for studying protein folding problem. The model can provide very useful insights into the folding mechanisms by allowing one to calculate the free energy surfaces efficiently. We first calculate the free energy surfaces of ubiquitin and BBL, using both approximate and recently developed exact solutions of the model. Ubiquitin exhibits a typical two-state folding behavior, while BBL downhill folding in our study. We then extend the method to study of a protein-RNA complex. In particular, we focus on PAZ-siRNA complex. In order to elucidate the interplay between folding and binding kinetics for this system we perform comparative studies of PAZ only, PAZ-siRNA complex and two mutated complexes. We find that folding and binding are strongly coupled with each other and the bound PAZ is more stable than the unbound PAZ. Our results also suggest that the binding sites of the siRNA may serve act as a nucleus in the folding process.

Key Words : Protein folding, Free energy, Ubiquitin, BBL, PAZ-siRNA complex

Introduction

Proteins play crucial roles in nearly all biological processes, including enzymatic activities, signal transmissions, and structural functions. They work as molecular machines that perform specific functions in the folded states. In order to understand molecular mechanisms of protein functions, knowledge on the structural and kinetic information of the protein is crucial. Protein folding problem, *viz.*, prediction of the three-dimensional structure of a protein from its amino acid sequence, has been an outstanding issue in protein sciences, and has been an intense area of research both from experimental and theoretical viewpoints. Although recent progresses made in the protein folding problem have been remarkable, *ab initio* prediction still remains very much as a challenge. In recent years, proteins not only in their folded states but in the misfolded states as well are drawing novel attentions due to their physiological importance. In particular, aggregations of certain peptides occurring in their unfolded states are believed to be intimately involved in the development of neurodegenerative diseases, such as the Alzheimer's and Parkinson's diseases. As such, not only the structural aspect but also the mechanism of protein folding has become key issues toward the understanding of specific functions that proteins carry out.¹⁻⁵

From a computational point of view, molecular dynamics (MD) simulations have provided important theoretical machinery in predicting protein structures and elucidating folding mechanisms. In order to fully understand folding

mechanisms of proteins, the information on free energy surfaces for the folding process is critical.^{6,7} However, constructing free energy surfaces from MD simulations are very numerically intensive. As such, even with fast developments in computational resources, obtaining free energy surfaces from all-atom MD simulations becomes prohibitively difficult in the case of medium-to-large sized proteins, consisting of more than a few hundred amino acids.

In this regard, coarse-grained models can provide a very attractive, alternative approach in the study of protein folding processes, when they are appropriately utilized.^{8,9} For example, Muñoz-Eaton (ME) model, originally developed by Wako and Saitō,^{10,11} and later extensively utilized by Muñoz and Eaton as well as others,¹²⁻¹⁶ provides an efficient method to calculate the free energy surfaces based upon a simple, lattice model, similar to the Ising model.

In this paper, we use the ME model to first calculate the folding free energy surfaces of two well-known proteins, ubiquitin and BBL, each of which represents a typical two-state and a downhill folder in the folding process, respectively. We choose these systems not only because they are representative members of their folding behavior but also they also perform physiologically important functions to be described later. Then, we extend the original ME model to the study of a protein-RNA complex, PAZ-siRNA complex, in order to study its own folding-binding behavior. We are not aware of any previous study on the folding and binding behavior of PAZ-siRNA complex based on the ME model.

This paper is organized as follows. After a brief introduction of the ME model, we explain how it can be used to calculate the folding free energy surface in the Computational Methods Section. Then, we first present our

[†]This paper is dedicated to Professor Eun Lee on the occasion of his honourable retirement.

results on two simple proteins, ubiquitin and BBL, and then move on to the protein-RNA complex in the Results and Discussion Section. We provide a concluding remark and discuss further studies in the Conclusion Section.

Computational Methods

Typically protein conformations are described by dihedral angles along peptide bonds, and they assume continuous values belonging to typical ranges. In order to obtain the free energy surface of proteins using MD simulations, it is often necessary to sample a huge number of protein conformations corresponding to different values of dihedral angles. This is usually a very computationally demanding procedure.

As an alternative approach, one can invoke a native-structure based model in the free energy calculation. One of such models, the ME model, describes protein conformations using a binary variable, m_i , for each peptide bond, instead of continuous values for dihedral angles. The binary variable m_i describes the conformational state of the i -th peptide bond and it takes a value $m_i=1$ when the i -th bond is in a native conformation while $m_i=0$ when it is in a non-native or unfolded conformation. Thus, the conformation of a protein is fully described by specifying an N -dimensional vector, $\{m_i\}$ ($i=1, 2, \dots, N$), with N being the number of residues in a given protein.

The effective Hamiltonian of the ME model is given by

$$H(\{m_k\}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^N \varepsilon_{ij} \Delta_{ij} \prod_{k=i}^j m_k - T \sum_{i=1}^N s_i m_i \quad (1)$$

Here, m_i is a binary, bond variable described earlier and Δ_{ij} is a matrix element of the so called contact map of a protein. In each native structure of proteins, if the distance between heavy atoms in the i -th and j -th residues is smaller than a pre-defined cut-off distance, Δ_{ij} is assigned a value 1; otherwise 0. ε_{ij} is the interaction energy between the i -th and j -th residues and s_i is the residual entropy cost for the i -th residue in the native state. T is the temperature. We note that the i -th and j -th residues can interact with each other only when all the peptide bonds between them are in their native conformations. For simplicity, we use 5.0 Å as a cut-off distance as used previously.¹³⁻¹⁶ We also use common interaction parameters $\varepsilon_{ij} = -0.5$ kcal/mol and the entropic cost $S_i = -1.327$ cal/mol·K for all interactions and residues of proteins as done previously.¹³⁻¹⁶

The statistical weight of each protein conformation is given by the usual Boltzmann factor,

$$w(\{m_k\}) = \exp\left[-\frac{H(\{m_k\})}{k_B T}\right] \quad (2)$$

where k_B is the Boltzmann constant. Even with such a simple model system, exact enumerations of all the possible configurations in the ME model remains as a challenge. For example, ubiquitin has a 76 peptide bonds and the number of possible microstates is $2^{76} = 7.6 \times 10^{22}$. Therefore, we first perform approximate calculations of the free energy based upon so called single, double, and triple sequence

approximations (SSA, DSA, and TSA) that were used before.¹⁷ Briefly, SSA includes only one consecutive native stretches while DSA and TSA two and three consecutive native stretches, respectively, when calculating the free energy surface.

Recently, an exact solution of the ME model was found via transfer matrix formalism.¹⁸ This method introduces a new variable $m_{j,i}$ by mapping the original one-dimensional variable into a two-dimensional one,

$$m_{j,i} = \sum_{k=i}^j m_k \quad (3)$$

We then define the generating function, $Z(\lambda)$, for the partition function of the system Z_j ,

$$Z(\lambda) = \sum_{j=0}^N Z_j \lambda^j \quad (4)$$

where Z_j is the partition function for all the configurations with j native bonds and λ is a dummy variable.

In the transfer matrix formalism, we first define v_k^j ($k > 0$) as a j dimensional row vector whose first $(j-k)$ elements are zero and the next k elements are one, and v_0^j as another j dimensional row vector whose elements are all zero except for the first element which is one. Then we can obtain the following matrix-vector equations for the transfer matrix elements $Q_{j+1}^j(\lambda)$,

$$Q_{j+1}^j(\lambda) v_k^{j+1} = v_{k-1}^j \text{ for } k=1, 2, \dots, j+1 \quad (5-1)$$

$$Q_{j+1}^j(\lambda) v_0^{j+1} = \sum_{k=0}^j \lambda^k w_{j,j+1-k} v_k^j \quad (5-2)$$

where $w_{j,i}$ is the Boltzmann weight of the configuration that has native states from the i -th to the j -th bond. The partition function given in Eq. (4) is then calculated as

$$Z(\lambda) = v_0^0 Q_{N+1}^N(\lambda) Q_N^{N-1}(\lambda) \dots Q_2^1(\lambda) Q_1^0(\lambda) v_0^{N+1} \quad (6)$$

The partition function for the configurations with j native bonds, Z_j , is given by the coefficient of λ^j in Eq. (4), and it is related to the free energy profile F_j as a function of the number of native bonds j via the following equation,

$$F_j = -k_B T \ln Z_j \quad (7)$$

Results and Discussions

To construct the Hamiltonian for each protein and the protein-RNA complex, we need information on the native structure of the system. We use PDB database to obtain structural information of the protein systems studied in this work, ubiquitin(PDB ID: 1UBQ), BBL(1BBL) and PAZ-siRNA complex(1SI2). As an example, we show a native structure of PAZ-siRNA complex and its contact map in Fig. 1.

(a) Ubiquitin. Ubiquitin, a small protein present in all eukaryotic cells, serves as a tag that marks a protein for destruction. Unfolded ubiquitin could induce serious problems in the protein destruction process, which makes it very important to have detailed molecular picture regarding its

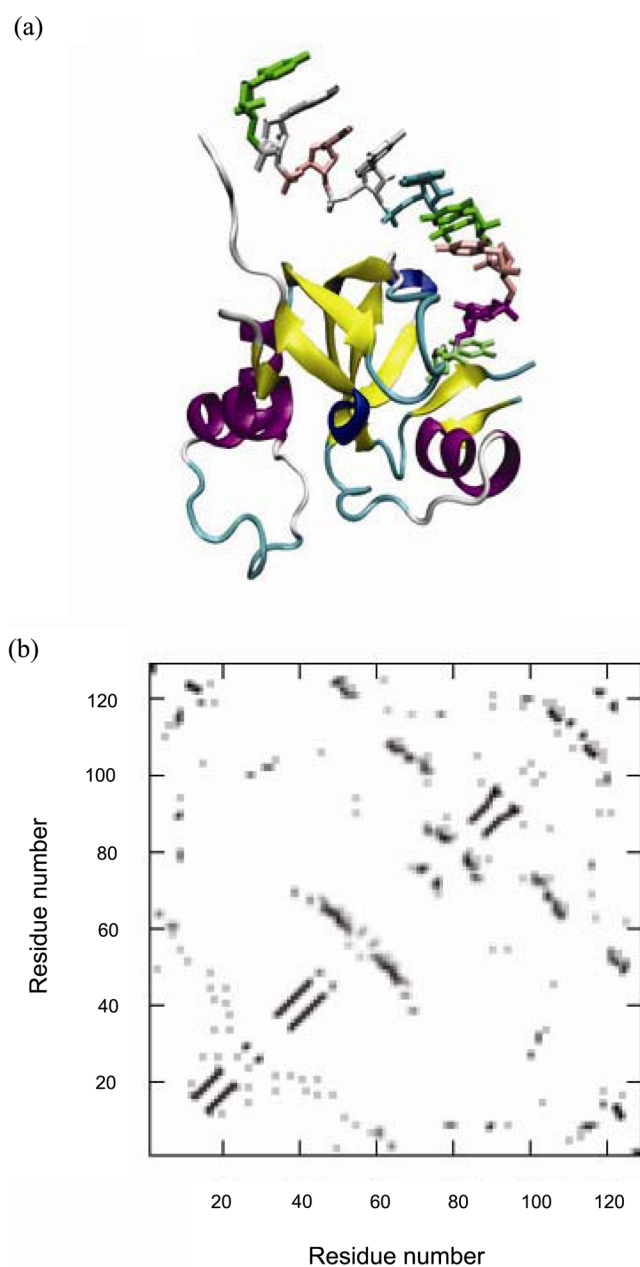


Figure 1. (a) Three-dimensional structure of PAZ-siRNA complex and (b) its contact map. Black dots in a contact map means that $\Delta_{ij} = 1$.

unfolding-folding behavior. Due to this reason there has been intense interest in the folding pathways of ubiquitin in recent years.¹⁹⁻²¹ Whether the ubiquitin undergoes a two-state or three-state folding process has been controversial in the folding kinetics studies.²⁰ Recent experimental studies seem to support the two-state folding behavior of ubiquitin, although there are still some debates, in particular, regarding experimental conditions such as temperature and pH.^{22,23}

We show in Fig. 2(a) one-dimensional free energy surfaces of ubiquitin. We calculated them using three approximate schemes and the exact method at the folding mid-temperature, $T_m = 458.7\text{K}$. As the reaction coordinate, or order parameter to be more precise, we chose to use the

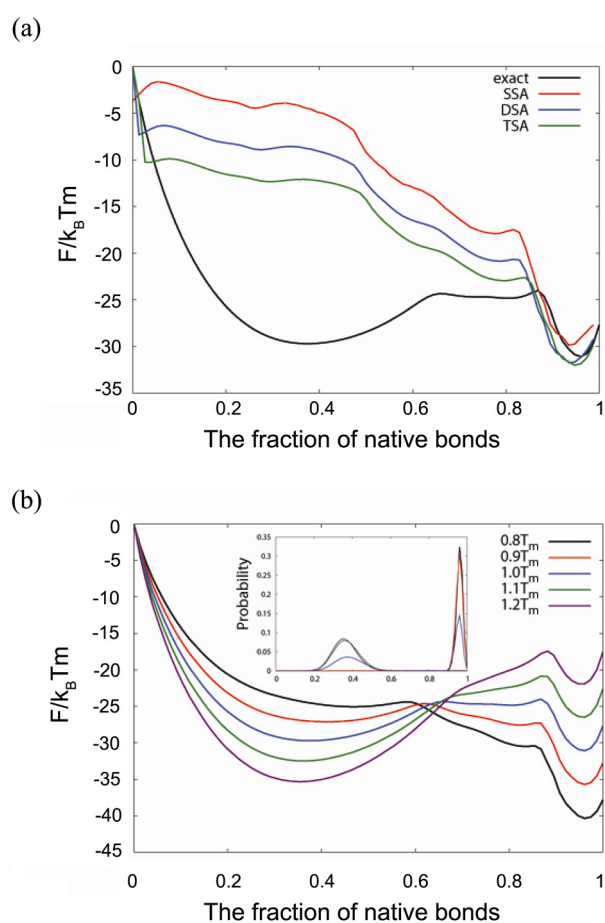


Figure 2. (a) Free energy profiles of ubiquitin at T_m calculated by SSA(red), DSA (blue), TSA (green), and exact solution(black) along the reaction coordinate chosen as the number of native bonds in the conformation. (b) Free energy profiles and probability distributions (inset) of ubiquitin at different temperatures.

fraction of native bonds in the protein. The folding mid-temperature, T_m , is defined by the temperature at which the populations of native and non-native structures are the same with each other. The free energy surface calculated using the exact method clearly exhibits a two-state behavior of ubiquitin that is characterized two distinct minima and a single barrier separating them in the free energy surface.

It is noteworthy that the free energy minimum for the native state does not correspond to the case where all the residues are folded perfectly, which would appear at the very end of the reaction coordinate. Rather, the minimum appears a little earlier. This phenomenon reflects the fact some residues, in particular, terminal ones, are partially folded in the native state, and it can be explained by the entropic cost effect. In other words, since the perfectly ordered state would be highly penalized by the entropic term, the free energy minimum appears at conformations with partially unfolded bonds in our calculation.

While the exact calculation results clearly shows a two-state folding behavior of ubiquitin, none of three approximate calculation results captures this feature. This is because

the entropic contributions to the free energy are not sufficiently considered in the approximate calculations. It is consistent with the observation that as the number of consecutive native sequences increases from SSA to TSA, the free energy of unfolded state decreases. Furthermore, free energy surfaces calculated by approximate methods display more rugged structures than the exact one does, which arises due to the fact that approximate schemes, even at the level of TSA, cannot take into account configurations with lower energy but more consecutive native stretches. Therefore, if a protein is trapped in a local minimum, it cannot escape from there without additional native stretches formed in a non-native state region in the case of approximate calculations. This suggests that transitions from non-native to native conformation of peptide bonds should occur simultaneously around energetically favorable regions.

Figure 2(b) shows free energy profiles of ubiquitin as the temperature changes from below to above T_m . In the inset also shown are the probability distributions of the conformations given by the Boltzmann weight of the free energy curves. One can see that the probability distributions show two distinct and relatively sharp peaks at non-native and native regions, which is a characteristic of a typical two-state folder.

(b) BBL. Another target protein is the peripheral subunit binding domain (PSBD) of the dihydrolipoamide succinyl transferase of *Escherichia coli*²⁴, which is called BBL. This domain has a small number of residues relative to other functional proteins, and it comprises two alpha helices connected by a loop. The domain has attracted much interest because of its barrierless transition of folding or unfolding kinetics, in other words, downhill behavior, revealed by many experiments.²⁵

In order to compare the free energy of this protein with that of ubiquitin, we also calculate free energy profiles and the probability distributions of BBL shown in Fig. 3. The probability distribution is obtained by Boltzmann factors of free energies and the folding mid-temperature turns out to be 262.0K.

In Figure 3, it is interesting to notice that the free energy curve of BBL exhibits a very low barrier at T_m . Also, even with a slight variation of the temperature the free energy barrier disappears completely, and the probability distribution of conformations develop a single broad peak as shown in the inset. This feature clearly distinguishes itself from the ubiquitin case. In the case of BBL, as the temperature decreases below T_m , the configuration with highest probability, that is, the equilibrium configuration, approaches to the native structure, consistent with the fact BBL being a downhill folder. Therefore, while being a simple lattice-based model, the ME model successfully captures both two-state folding and downhill folding behavior of each representative protein.

(c) PAZ-siRNA Complex. Finally, we choose to study PAZ domain consisting of eight β -sheets and three α -helices. The PAZ domain binds to siRNA to form a multimeric, RNA-induced silencing complex.²⁶ Recently an

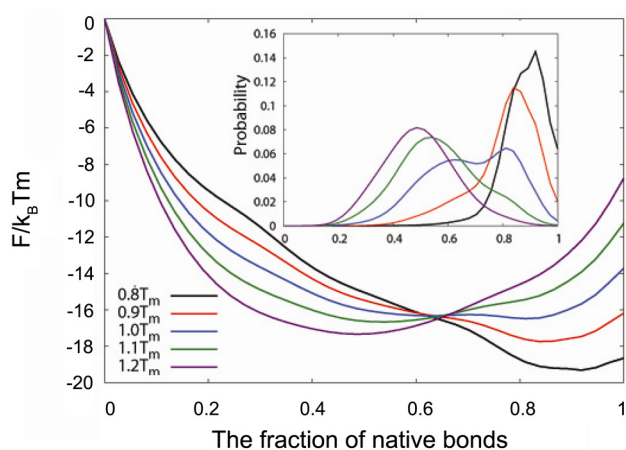


Figure 3. Free energy profiles and probability distributions (inset) of BBL at different temperatures, $0.8 T_m$, $0.9 T_m$, $1.0 T_m$, $1.1 T_m$, and $1.2 T_m$.

MD study showed that the folding and binding events occurring in the PAZ-siRNA complex are coupled with each other.²⁷ We extend our study on proteins to the complex case to consider the folding and binding behavior of the PAZ-siRNA complex based on the ME model.

PAZ domain has 120 peptide bonds, and an RNA having 9 nucleotides binds to the PAZ domain as shown in Fig. 1. We attached a 3'-end of the siRNA to an N'-terminal of the protein. To extend the ME model to this system we need to adjust model parameters used in the Hamiltonian. In general, the interaction between a nucleotide and an amino acid is mediated by more hydrogen bonds than that between amino acids, and the nucleotides have more orientational degrees of freedom than amino acids do. Accordingly, we set the energy parameters to be $\epsilon_{ij,np} = 1.4 \epsilon_{ij,p} = -0.7$ kcal/mol, where $\epsilon_{ij,np}$ and $\epsilon_{ij,p}$ are the interaction energies, the former between a nucleotide and an amino acid residue and the latter between amino acid residues, respectively. Defined in a similar way the entropy parameters, $s_{i,n}$ and $s_{i,p}$, are set to be $s_{i,r} = 2s_{i,p} = -2.654$ cal/mol·K.

The quaternary structure of a protein complex is determined by the relative orientation of bindings between domains which may be taken as an "auxiliary bond." If one models the binding through an auxiliary bond, it will have more degrees of freedom than real peptide bonds within a single protein. In turn, the entropy cost of the auxiliary bond upon binding should be greater than those of peptide bonds upon folding. Due to this reason, we use $s_{9,np} = 6.373$ cal/mol·K to take into account a higher entropic cost for the auxiliary bond between a N-terminal residue and a 3'-end of the nucleotide.

Using the parameters defined above we calculate the free energy curve of a PAZ-siRNA complex in Fig. 4 by using the exact method of ME model. In doing so we compare the free energy profiles of the PAZ domain alone (apo-PAZ) and the PAZ-siRNA complex, respectively. In Fig. 4, the fraction of native bonds in the complex system is defined as the ratio of the number of native bonds (including nucleotides) to the

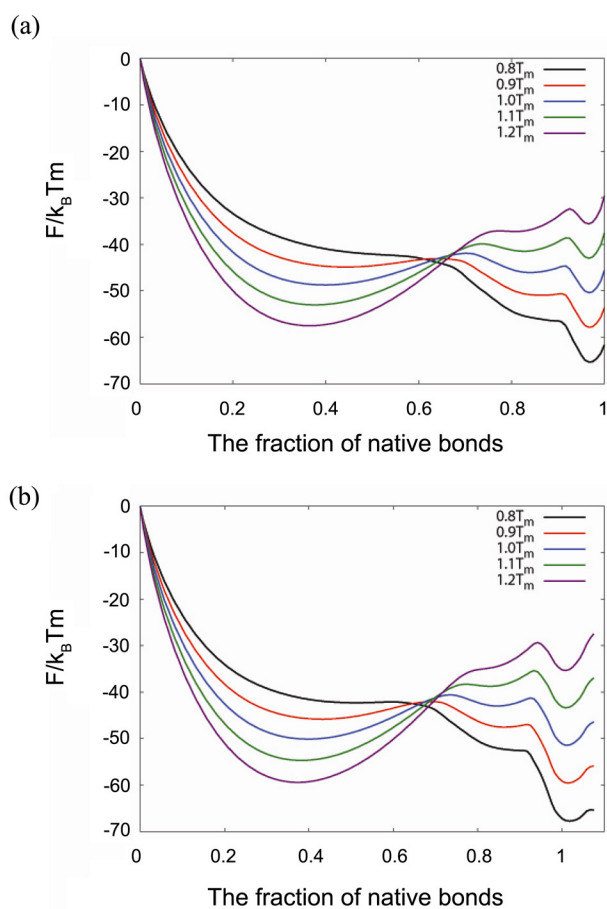


Figure 4. (a) Free energy profiles of apo-PAZ and (b) free energy profiles of bound-PAZ at different temperatures.

total number of residues in apo-PAZ. Therefore, in the case of the complex, the fraction of native bonds can be larger than unity. In Figure 4, the folding mid-temperature of the apo-PAZ and that of bound-PAZ are determined as $T_{m,a} = 428.3\text{K}$, $T_{m,b} = 451.6\text{K}$, respectively. The higher folding mid-temperature of the PAZ-siRNA complex implies that, at the same temperature, the bound-PAZ is more stable than that of the apo-PAZ. Moreover, both models show two-state folding behavior regardless of RNA binding, which is consistent with a recent MD simulation results.²⁷

To compare the stabilities in the native state and the transition state, we calculate free energy profiles of both apo-PAZ and PAZ-siRNA at the same temperature, $T = T_{m,a} = 428.3\text{K}$. From the Fig. 5, despite the large entropic cost of the auxiliary bonds between the protein and the siRNA, the free energy in the native state of the complex is much lower than that of the apo-PAZ and there is no additional free energy barrier. These results show that binding and folding are coupled for the PAZ-siRNA complex, which is also consistent with MD simulation.²⁷

In order to consider the effect of siRNA binding for the folding mechanism of the PAZ-domain, we introduce two mutant models of an apo-PAZ, mut1-PAZ and mut2-PAZ. In mut1-PAZ, we first identify 28 residues interacting with the

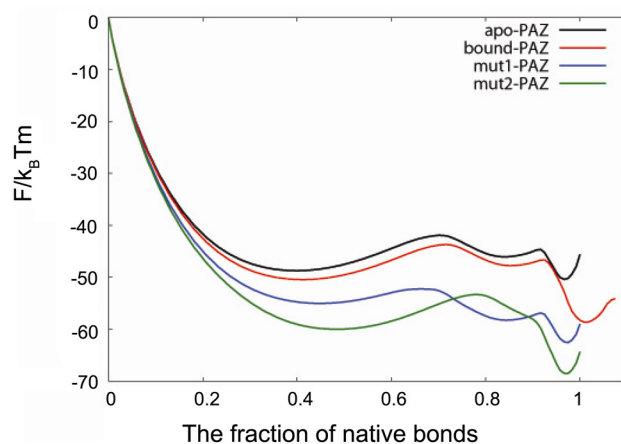


Figure 5. Comparison of free energy profiles of apo- (black), bound- (red), and mut1- (blue), and mut2- (green) PAZ models.

siRNA in the bound-PAZ complex, and let their entropic parameters to be zero. This will stabilize the siRNA binding sites in their native states. As a control case, we introduce another mutant model, mut2-PAZ, where the entropy parameters are set to be zero for other 28 residues, randomly chosen from the non-binding sites of a protein. Free energy curves of apo-, bound-, mut1-, and mut2-PAZs are compared in Fig. 5.

We notice that while the magnitude of the free energy barrier of a mut2-PAZ is similar to that of an apo-PAZ, the free energy profile of a mut1-PAZ has a very low free energy barrier. This implies that the interaction between the PAZ domain and the siRNA contributes significantly to lowering the free energy barrier in the folding kinetics of the PAZ domain, and that the binding sites of the siRNA, which corresponds to a beta-sheet region, may serves as a nucleus in folding.

Concluding Remarks

In this work we used the ME model to demonstrate that essential free energetic characteristics can be captured by a simple, lattice based model for of three different class of protein systems. Although it is a simple, lattice based model, the ME model can be useful in successfully explaining basic mechanisms underlying the folding and unfolding processes, not only of a simple proteins but a protein-RNA complex. It can adequately capture different classes of folding behaviors, two-state folding for ubiquitin and downhill for BBL. We also demonstrated that the model can be simply extended to explain the folding and binding event of a PAZ-siRNA complex system, and showed that binding and folding events are coupled to each other. In the near future, we plan to study nonequilibrium nature of the folding processes based on the ME model, and will elucidate diverse folding pathways of proteins in more detail.

Acknowledgments. This work was supported by the National Research Foundation of Korea (Grant Nos. 2011-

0001212, 2011-0003555, and 2011-0018038), the KISTI Supercomputing Center (KSC-2009-S02-0003 and KCS-2011-C1-07), the BK21 Program, and the Seoul Science Fellowship awarded to Eunsang Lee.

References

1. Roychaudhuri, R.; Yang, M.; Hoshi, M. M.; Teplow, D. B. *J. Biol. Chem.* **2009**, *284*, 4749.
 2. Sipe, J. D.; Cohen, A. S. *J. Struct. Biol.* **2000**, *130*, 88.
 3. Jackson, S. E.; Fersht, A. R. *Biochemistry* **1991**, *30*, 10428.
 4. Onuchic, J.; Luthey-Schulten, A.; Wolynes, P. G. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545.
 5. Plaxco, K. W.; Simos, K. T.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 985.
 6. Lee, J.; Shin, S. *Bull. Kor. Chem. Soc.* **2008**, *29*, 741.
 7. Lee, W.; Park, H.; Lee, S. *Bull. Kor. Chem. Soc.* **2008**, *29*, 363.
 8. Urbanc, B.; Cruz, L.; Ding, F.; Sammond, D.; Khare, S.; Buldyrev, S. V.; Stanley, H. E.; Dokholyan, N. V. *Biophys. J.* **2004**, *87*, 2310.
 9. Schnabel, S.; Bachmann, B.; Janke, W. *Phys. Rev. Lett.* **2007**, *98*, 48103.
 10. Wako, H.; Saitô, N. *J. Phys. Soc. Japan* **1978**, *44*, 1931.
 11. Wako, H.; Saitô, N. *J. Phys. Soc. Japan* **1978**, *44*, 1939.
 12. Muñoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. *Nature(London)* **1997**, *390*, 196.
 13. Imparato, A.; Pelizzola, A.; Zamparo, M. *Phys. Rev. Lett.* **2007**, *98*, 148102.
 14. Muñoz, V.; Henry, E. R.; Hofrichter, J.; Eaton, W. A. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 5872.
 15. Zamparo, M.; Pelizzola, A. *Phys. Rev. Lett.* **2006**, *97*, 68106.
 16. Henry, E. R.; Eaton, W. A. *J. Chem. Phys.* **2004**, *307*, 163.
 17. Muñoz, V.; Eaton, W. A. *Proc. Natl. Acad. Sci. USA.* **1999**, *96*, 11311.
 18. Bruscolini, P.; Pelizzola, A. *Phys. Rev. Lett.* **2002**, *88*, 258101.
 19. Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. *J. Mol. Biol.* **1987**, *194*, 531.
 20. Jackson, S. E. *Org. Biomol. Chem.* **2006**, *4*, 1845.
 21. Gräter, F.; Grubmüller, H. *J. Struct. Biol.* **2007**, *157*, 557.
 22. Krantz, B. A.; Sosnick, T. R. *Biochemistry* **2000**, *39*, 11696.
 23. Went, H. M.; Benitez-Cardoza, C. B.; Jackson, S. E. *FEBS Lett.* **2004**, *567*, 333.
 24. Packman, L. C.; Perham, R. N. *FEBS Lett.* **1986**, *206*, 193.
 25. Garcia-Mira, M. M.; Sadqi, M.; Fischer, N.; Sanchez-Ruiz, J. M.; Muñoz, V. *Science* **2002**, *298*, 2191.
 26. Ma, J. B.; Ye, K.; Patel, D. J. *Nature* **2004**, *429*, 318.
 27. Hai-Feng, C. *J. Chem. Theory Comput.* **2008**, *4*, 1360.
-