A Lattice Model Study of Native Contact Restraints in Protein Folding

Won Seok Oh and Jae-Min Shin*

Molecular Design Laboratory, Hanhyo Institute of Technology, San 6, Daeya-Dong, Shiheung-Shi, Kyungki-Do 429-010, Korea

Received April 22, 1996

To explore protein folding mechanism, we simulated a folding pathway in a simplified 3×3×3 cubic lattice. In the lattice folding Monte Carlo simulations, each of the 28 possible native packing pairs that exist in the native conformation was used as a conformational restraint. The native packing restraints in the lattice model could be considered as a disulfide linkage restraint in a real protein. The results suggest that proteins denatured with a small disulfide loop can, but not always, fold faster than proteins without any disulfide linkage and than proteins with a larger disulfide loop. The results also suggest that there is a rough correlation between loop size of the native packing restraint and folding time. That is, the order of native residue-residue packing interaction in protein folding is likely dependent on the residue-residue distance in primary sequence. The strength of monomer-monomer pairwise interaction is not important in the determination of the packing order in lattice folding. From the folding simulations of five strong folding lattice sequences, it was also found that the context encoded in the primary sequence, which we do not yet clearly understand, plays more crucial role in the determination of detailed folding kinetics. Our restrained lattice model approach would provide a useful strategy to the future protein folding experiments by suggesting a protein engineering for the fast or slow folding research.

Introduction

Protein folding problem, how primary amino acid sequence determines its three dimensional (3D) structure in a given environment, is one of the most fundamental questions in bio-molecular science and has been treated as one of the main research projects after the demonstration of *in vitro* refolding of ribonuclease A.[1] One of the most interesting questions in protein folding is whether the native conformation is fully accessible from any conformational states. It is believed from many experimental and theoretical studies that the native conformation for most proteins is likely a global free energy minimum conformation and is accessible from any conformational states, though there are some exceptions that strongly support kinetic control of protein folding.[2,3]

In order for a protein to get correctly folded into the native 3D conformation, it is no doubt that number of correctly matched residue-residue, or atom-atom, packing interactions that exist in and is believed to stabilize the native 3D structure should occur during the folding process. An interesting question here is whether there are any instructive rules that govern the sequence of these packing events during the folding process. If there are, these rules are encoded in the primary sequence and the next question would be how to figure out these instructions from the primary sequence. As an effort to answer about these questions, we have studied on the folding pathway of the 3×3×3 cubic lattice chain (Figure 1) by Monte Carlo (MC) simulation. In our lattice folding study, native packing (NP) is defined as a pair of beads that topologically contact each other in the native 3D conformation but are not nearest neighbor in primary sequence (dotted line in Figure 1). We have studied the folding mechanism of the lattice chain by using this NP as a conformational restraint. Thus, our NP restrained lattice model can be considered as a protein restrained by a native residue-residue contact, for an example, linked by a native disulfide bond.

Simplified lattice models have been used for the study of protein folding.[4~6] These lattice models could be considered as a simplified protein for the folding study when the lattice sequence is carefully designed.[7~9]

Thermodynamically, the denatured state of the NP restrained lattice chain is less stable than that of the lattice chain without any restraints because of the lack of chain flexibility.[10~16] Moreover, the restraints that form larger restrained loop generally destabilize the denatured states more

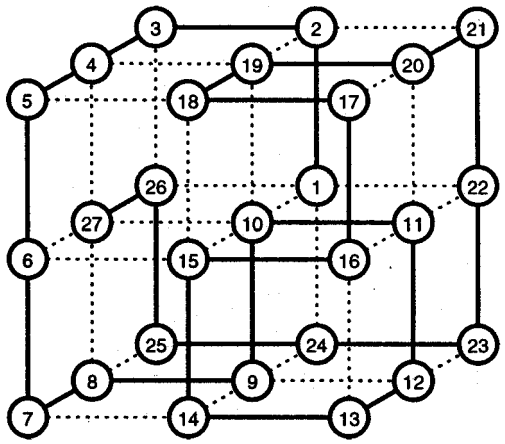

Figure 1. An example of the global energy minimum conformation of sequence number 1 in 3×3×3 cubic lattice. There are 27 beads linked through solid lines and 28 native packing (NP) pairs drawn by dashed lines. Considering symmetry, this lattice can have 103346 different conformations.[31]

than the smaller restrained loop. In this study, the former will be termed as a distal NP and the latter as a proximal NP restraint for the convenience, respectively.

Kinetically, proximal NP is likely to form earlier than distal NP during the folding pathway because of the spatial availability. No matter how is the detailed folding process in the lattice model, the 28 NPs (Figure 1) should be in contacts during folding process to form and to stabilize the native conformation. It is the basic idea of this study that if there are certain sequence of NP events in the folding process, then we could find useful information for the fast folding NP and slow folding NP by exhaustive search of the foldicity of the simple lattice chain restrained by every NP.

Our NP restrained model approach will provide insight for the protein folding study by suggesting an engineering of protein with a native-like conformational restraint such as an insertion of the disulfide linkage that forms a native contact.[17,18]

Methods

Lattice model. Simple 3×3×3 cubic lattice model (Figure 1)[19] is used. The parameters for the energy expression of cubic lattice chain are obtained by the same procedure that have been reported by Sali *et al.*,[7,8] *i.e.*,

$$E=\sum_{i<j} B_{ij}\Delta(r_i-r_j), \tag{1}$$

where E is total energy and B_{ij} is the interaction energy parameter between beads i and j located at position r_i and r_j, respectively. $\Delta(r_i-r_j)$ is 1 when beads i and j are in contact and is 0 otherwise. The interaction energy parameters B_{ij} are randomly assigned from a Gaussian distribution with a mean B_0 and standard deviation σ_B:

$$P(B_{ij})=\frac{1}{\alpha 2\pi\sigma_B}\, e^{-1/2(B_{ij}-B_0/\sigma_B)^2}. \tag{2}$$

Monte Carlo (MC) Lattice Simulation. In order to obtain strong folding lattice chains,[7,8] 200 random sequences are generated with $B_0=-2.0$ and $\sigma_B=1.0$ in the equation 2. From the partition function of all the 103346 different compact conformations[19] for each lattice chain, $Z=\sum\exp(-E_i/kT)$, the order parameter, $X=1-\sum p_i^2$, where $p_i=\exp(-E_i/kT)/Z$, is calculated as a function of temperature, T. Among these 200 random sequences, five strong folding lattice sequences are chosen for the NP restrained lattice folding study by examining the energy differences between the lowest energy conformation and second lowest energy conformations, $\Delta(E_1-E_0)$, and also by examining the simulation temperature, $T_X=T(X=0.8)$.[8] The energy spectra for 10 strongest folding sequences are shown in Figure 2. Monte Carlo (MC) simulation with Metropolis algorithm[20] of each strong folding sequence find its global energy minimum conformation from randomly generated initial conformations within 1-2×10^7 MC steps. Monte Carlo selection procedure for the moves of lattice chains is carried out by the same methods described in elsewhere.[8]

In the native conformation, lowest energy conformation, each lattice chain has a distinct set of the 28 NP[*i-j*] pairs which are drawn by dashed lines in Figure 1. In NP[*i-j*] restrained MC simulations, beads i and j are restrained to be in contact with each other. Initial random conformation for NP[*i-j*] restraint is obtained by selecting the random initial conformations that satisfy the NP[*i-j*] restraint. 20 independent MC runs are carried out for each NP[*i-j*] restraint to minimize the data fluctuation. Maximum of 1×10^8 Monte Carlo steps are allowed in all NP restrained MC folding simulations because the lattice chains with and without NP restraint usually fold into the native conformation within these steps. For some NP[*i-j*] restrained lattice chains, however, it is not possible to find out the native conformation within this maximum MC steps. In such cases, the maximum MC step is used for further analysis as if this run finds the native conformation at 1×10^8 MC step.

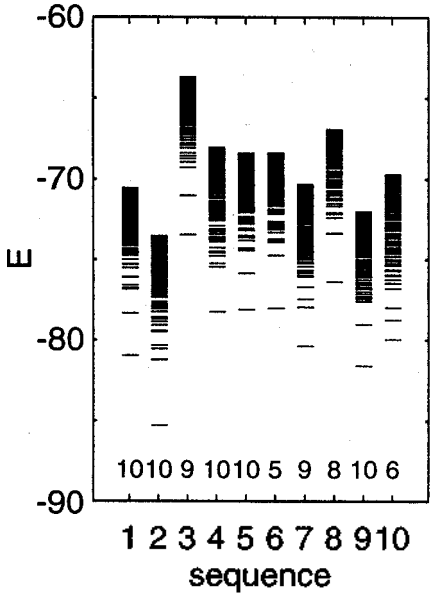

Figure 2. Energy spectra for 10 strong folding lattice sequences. 400 lowest energies are plotted. The numbers below the spectra are the number of successes runs among the 10 independent MC folding simulations, which find the native conformation from the random initial conformations within 5×10^7 MC steps.[7,8]

Folding Profile of NP[*i-j*] restrained lattice. The foldicity of NP[*i-j*] restrained lattice chain is defined as an average MC step of 20 independent MC runs at which the simulations of NP[*i-j*] restrained lattice chain find the native conformation from initial random conformation. The folding step is also considered as a refolding time of the lattice chain from the random conformation[7] because the Boltzmann selection rules by Metropolis MC algorithm basically follow thermodynamically accessible pathway. All the MC trajectories were saved at every 10^3 MC steps for the subsequent analysis. From these trajectory files, the density of the states and the free energy profile of lattice folding simulation can be calculated.[7] That is,

$$\langle E_Q\rangle=\sum_i^N n_{Qi}\varepsilon_i/\sum_i^N n_{Qi} \tag{3}$$

$$\langle S_Q\rangle=k\ln Z_Q+\frac{\langle E_Q\rangle}{T} \tag{4}$$

$$\langle F_Q\rangle=-kT\ln Z_Q \tag{5}$$

where $\langle E_Q \rangle$, $\langle S_Q \rangle$, and $\langle F_Q \rangle$ mean the statistical average of the energy, entropy, and free energy, respectively, at the reaction coordinate Q. In this study, Q is the number of NP, N_{NP}. Another convenient definition of reaction coordinates can also be defined as $(N_{NP})^2/CN$ where CN is number of contacts including non-native contacts. Index i means the bin of the slice of the ε axis where ε is the energy of lattice chain. In our study, the bin is 0.05 in reduced unit. n_{Q_i} is the number of conformation at (Q,i) position in (Q,ε) space. k and T are Boltzmann constant and temperature, respectively. In this study, k is set to 1 and T is 1.25-1.45, depending on the lattice sequence.[8] The Z_Q is defined as:

$$Z_Q = \sum_i^{\tilde{N}} \omega_{Q_i} e^{\varepsilon_i/kT} \tag{6}$$

where

$$\omega_{Q_i} = \frac{n_{Q_i}}{n_{Q_0\varepsilon_0}} e^{[-(\varepsilon_0-\varepsilon_i)/kT]} \tag{7}$$

Here, $n_{Q_0\varepsilon_0}$ is the number of population at native state. Accordingly, ε_0 is the energy of the native state.

Results and Discussion

Folding of NP restrained lattice. Five strong folding lattice sequences (see Methods section) are selected among 200 random sequences and studied throughout this study. In a 3×3×3 cubic lattice, the lowest energy conformation is considered as the native conformation, though it is more related to molten globule states[21] than the 'true' native conformation of real protein.[7,8,22] An example of the native conformation of the 3×3×3 cubic lattice chain is shown in Figure 1. The potential energy of this conformation is global minimum among the 103346 different compact conformations.[15,19] In the native conformation of the 3×3×3 lattice model, there are 28 native packing (NP) pairs which are drawn as dashed lines in Figure 1.

Figure 3 shows typical trajectory of lattice folding MC simulations for 10^8 steps with and without NP restraints. The lattice chain usually folds from the randomly generated initial conformation into the native conformation within 1-2×10^7 MC steps when there are no restraint.[7,8] The NP[*i-j*] restrained lattice chain is defined as the lattice chain that has a link between beads i and j. During lattice folding MC simulation, the number of NPs increase from near-zero to 28, *i.e.*, from random coil states or denatured states to the native conformation.[7] It is noted that NP[10-15] restrained lattice chain shows not only higher population of the native conformation but also faster folding than the free lattice chain.

Fast and slow folding NP restraints. The average MC folding steps of the NP[*i-j*] restrained lattice chains are shown in Figure 4. The more closer NP[*i-j*] to the diagonal in this figure corresponds to the more proximal NP[*i-j*] restraint in the primary sequence, thus more proximal NP[*i-j*] forms a smaller loop. It can be seen that the lattice chains restrained by distal NP[*i-j*], near the lower left corner in this figure, shows difficulties in finding the native conformation from the random initial conformation. As mentioned earlier, the free lattice chain usually find the native conformation within 1-2×10^7 MC steps. However, among the 28 NP[*i-j*] restrained lattices, about three to six NP[*i-j*] restrained lattices find the native conformation within 1×10^7 MC steps, while five to eight NP[*i-j*] restrained chains don't find it within even greater than 5×10^7 MC steps.

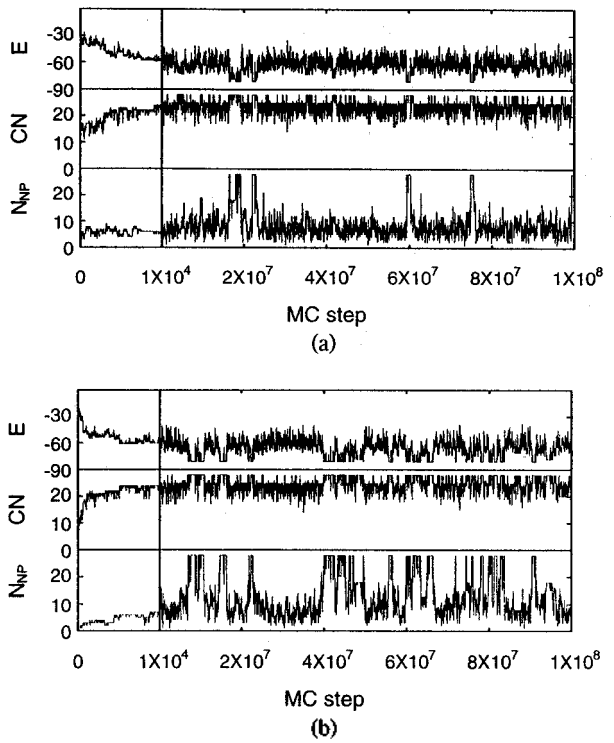

Figure 3. Typical trajectories for lattice folding simulations without any NP restraint (a) and with NP[10-15] restraint (b). E is energy in reduced unit of kT, where k is Boltzmann constant and T is temperature; CN is number of contacts including non-native one; N_{NP} is number of native contacts.

The fastest and slowest MC folding steps for the lattice sequence 1, on the average from 20 independent simulations, are 4.2×10^6 for NP[10-19] restraint and 92×10^6 for NP[3-26] restraint, respectively. The average MC folding step for the free lattice chain of the same sequence is 12.7×10^6. This means that in the lattice chains of same primary sequence (thus the same native conformation), NP[10-19] restrained lattice folds about three times faster than the free lattice chain, while NP[3-26] restrained lattice folds seven times slower than the free lattice chain. Because these results are obtained by averaging from 20 independent MC runs and the reliable distributions of the fast and slow NP[*i-j*] restrained MC folding are observed in all five strong folding lattice chains, it generally means that there are fast and slow folding NP[*i-j*] restraints in the lattice folding pathway.

The average folding steps for the five strong folding sequences are analyzed as a function of the NP distance in primary sequence, $|i-j|$, which are shown in Figure 5. It is surprising that there seems to exist rough correlation between the foldicity and the NP distance. Two general features can be drawn from this figure. Firstly, there are optimal NP[*i-j*] restraints that result in fast folding kinetics where NP distances are between 5 and 15. Secondly, distal NP

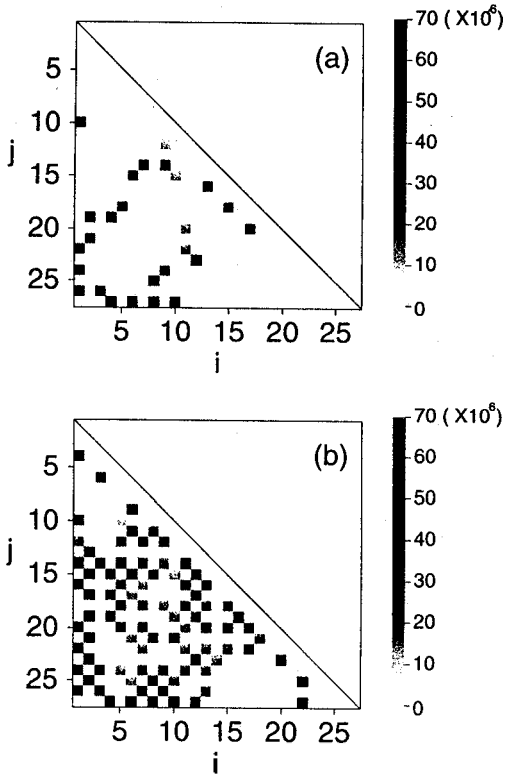

Figure 4. Plot of average MC folding steps of NP[*i-j*] restrained lattice simulations for sequence number 1 (a) and for all five strong folding sequences (b). *i* and *j* are the bead numbers and there are 28 NP[*i-j*] pairs for each strong folding sequence. Folding step for each NP[*i-j*] is obtained from 20 independent MC runs (see text). In (b), if there are two or more NP[*i-j*] restraints overlapped at the same (*i,j*) position among the five sequences, these values are averaged.

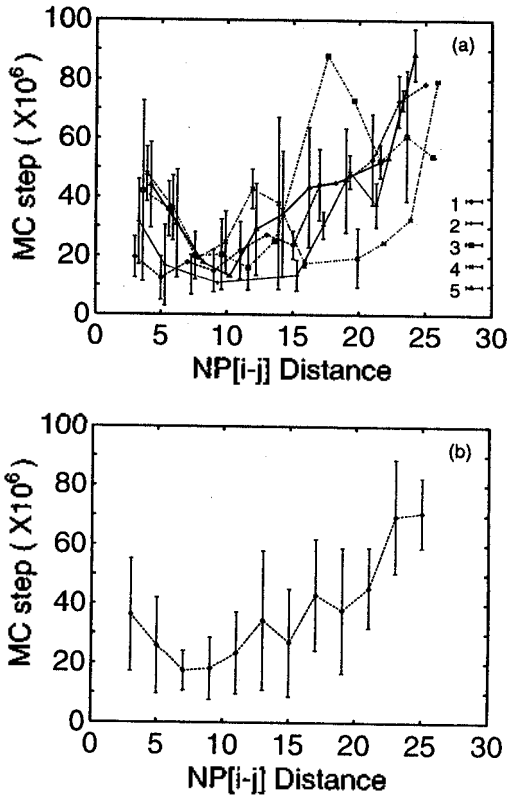

Figure 5. Correlation plots of MC folding steps as a function of NP[*i-j*] distance, |*i-j*|. All five strong folding lattice sequences are plotted individually in (a) and further average is made in (b). Vertical upper and lower bar are standard deviation for the cases, respectively.

restraints show slow folding kinetics compared either to the free lattice chain or to the lattice chain restrained with proximal NP restraints.

These results indicate that if there are sequence of NP events in the folding process, then the sequence of events may be roughly proportional to the NP distance. This rough correlation between NP distance and folding steps shows consistency with the present understanding on protein folding that the native secondary structure (as a result of native packing interaction between proximal residues) forms early and the native tertiary structure forms very late.[2,23-25] These results also imply that if there are more than two NP[*i-j*]s which show almost the same MC folding steps (time), those multiple NP[*i-j*] events can simultaneously occur during folding process.

It should be noted in Figure 4 and 5, however, that the detailed profiles of the correlation between NP distance and folding time are different in five different folding sequences, denoting other expression of protein folding problem that different sequences encode different folding kinetics and different native conformations.

It is clear that this rough correlation between NP distance and folding time does not depend on the number of possible conformations with NP[*i-j*] in the 103346 compact conformations (Figure 6(a))[15] nor on the strength of pairwise NP[*i-j*] interaction energy (Figure 6(b)) of this simple lattice model.[8]

These correlation between NP distance and folding time may be extended to the interpretation of the refolding experimental result from the mutant barnase done by Clarke & Fersht,[18] where the mutant barnase with a proximal NP restraint, obtained by oxidizing [85-102] disulfide bond at the denatured state, folds six times faster than the wild type barnase which does not have any disulfide restraint, while the mutant with a distal NP restraint, [43-80] disulfide bond, folds two times slower than the wild type barnase. Clarke & Fersht interpreted their results as due to the early folding and late folding segments, respectively, based on their previous folding studies of barnase.[26-29]

Folding pathways of NP restrained lattice chains. An additional interesting question is whether there are differences between the folding pathways for the slow and fast folding NP restrained lattice chains. In a statistical thermodynamic description, there are only two stable states in the lattice folding model, non-native state and native state. However, in a microscopic description, there are huge number of possible pathways from random to native states.

Free energy profiles of fast folding NP and slow folding NP restraints are shown in Figure 7. At the denatured states, the free energy of the NP restrained lattice chain is about 2 kT higher than that of the free lattice chain. The transition state lies between NP number of 23 and 25. In the fast fol-

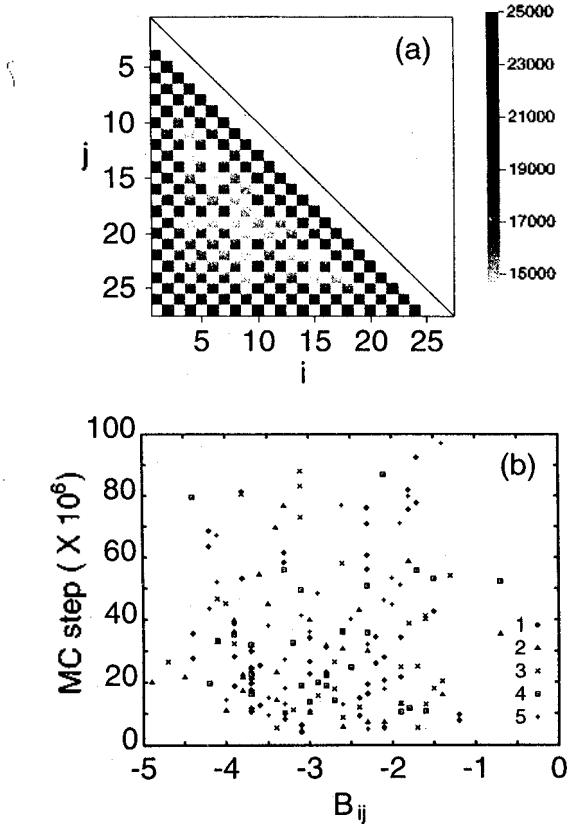

Figure 6. (a) Number of possible compact conformations that have contacts between beads i and j in the 3×3×3 cubic lattice. There are maximum of 36772 different conformations for pair [1-4] and minimum of 14454 conformations for pair [6-13]. (b) Plot of retrained MC folding steps versus pairwise interaction energy for the NP[i-j] restraints, B_{ij} in equation (1).

ding NP restrained lattices, the transition states seem to be the same state with that of the free lattice chain (Figure 7(b)). However, for the distal NP or slow folding NP restraints, the transition states seem to be somewhat different from that of the free lattice chain (Figure 7(c)).

Figure 8 shows the density of states in CN versus N_{NP} space, where CN is number of contacts which include the non-native contacts and N_{NP} is the number of native contacts. The fastest pathway of folding from random into the native conformation may be the diagonal line in this plot. All the states where CN=28 and N_{NP}≠28 mean mis-folded compact conformations. What is clearly seen in this figure is that near the native conformation, lower right corner, there exist significant differences in the conformational population between slow folding and fast folding NP restrained lattice chains. For a slow folding NP restraint, NP[3-26] (Figure 8(c)), the population of the mis-folded conformation where CN=28 and N_{NP}=22 is very high compared with the fast folding NP restrained lattice (Figure 8(b)) or free chain (Figure 8(a)). Therefore, it is supposed from this analysis that the fast folding NP restraint may prevent the lattice chain from getting into the incorrect mis-folded states because it correctly restrain the lattice chain.

Practical implication of fast and slow folding NP restraints. Although many criticisms about the 'reality'

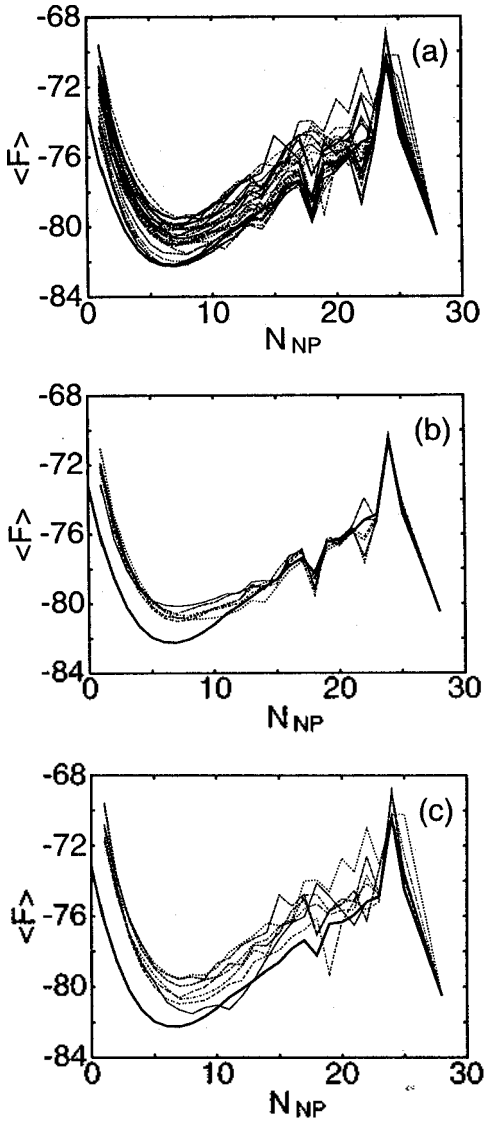

Figure 7. Free energy profiles for the NP[i-j] restrained lattice folding of sequence number 1. (a) Plot of all 28 NP[i-j] restrained lattices. (b) Plot of fast folding NP restraints; [4-29], [9-12], [10-15], [10-19], [11-16], and [11-20]. (c) Plot of slow folding NP restraints; [1-24], [1-26], [3-26], [4-27], [6-27], [8-27], and [10-27]. Bold-faced solid lines in (a), (b), and (c) represent the profile for free lattice chains. Fast and slow folding NP[i-j] restraints, here, are the cases that MC folding steps are less than 1×10^7 steps and greater than 5×10^7 steps, respectively.

of the lattice simulation can be raised, including the co-operativity of the protein folding,[22] the simplicity of the model,[5,30] and the consideration of global energy minimum conformation as the native conformation, the lattice model and the lattice simulation have given us valuable instruction about the protein folding. Some of these are sequence dependence of protein folding[7~9,31] and the co-operativity,[32] the prediction of protein folding and structure,[33~35] and the kinetics of protein folding.[7,8,36]

The instruction of our NP restrained lattice folding study is that NP restraint may be used as a strategy in the study of protein folding mechanism. For an example, introduction

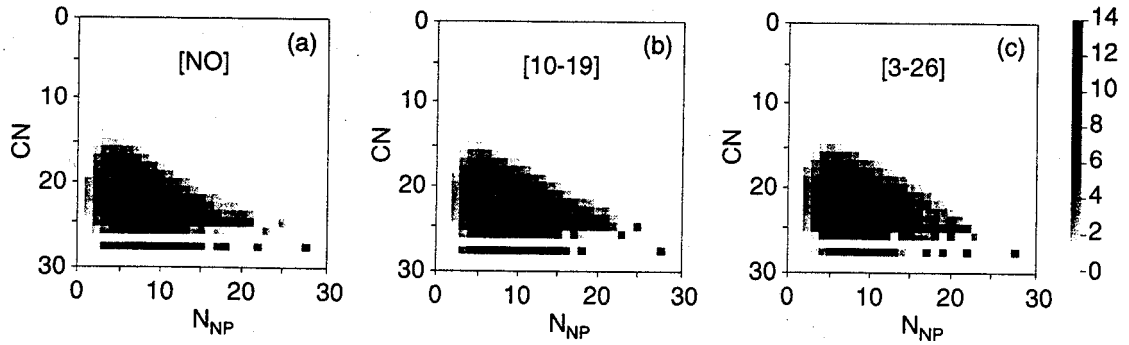

Figure 8. Density of states for free lattice chain, (a); for fast folding NP[10-19], (b); for slow folding NP[3-26], (c). Abbreviations used: NP, native packing; N_{NP}, Number of native packing; CN, number of contacts.

of NP disulfide bond by protein engineering[18] results in either fast folding or slow folding for the protein of known 3D structure. Furthermore, this instruction could be used for the purpose of the high yield of protein refolding by decreasing the level of non-native conformation.

Conclusion

From our MC folding study with a NP restrained lattice model, the followings could be drawn as a conclusion:

(1) There are rough correlation between the size of NP loop and folding kinetics, *i.e.*, lattice chain with a large NP loop shows slow folding kinetics, while lattice chain with optimum or smaller NP restraint (see Figure 5(b)) does fast folding kinetics. From the MC simulation of five strong folding lattices, however, it was also found that the context encoded in primary sequence determines the details of folding kinetics (see Figure 5(a)). Futhermore, the pairwise contact energy of the NP restraint does not affect the folding kinetics (Figure 6(b)).

(2) In denatured state, NP restrained lattice chain is relatively unstable compared to the free lattice chain because of the excluded volume effect (see Figure 7).

(3) This study also suggest that the transition state of the lattice model with slow folding NP restraints may be different from that of fast folding NP restraints or free lattice chain (see Figures 7 and 8).

Acknowledgment. We thank Dr. Hang-Cheol Shin for invaluable discussion on this work, Dr. Sunzong Kang and Dr. Chong-Shul Kim for encouragement, and Sang S. Pak for help in the preparation of the manuscript. This work was supported by Hanil Synthetic Fiber Co.

References

1. Anfinsen, C. B. *Science* **1973**, *181*, 223.
2. Creighton, T. E. *Biochem J.* **1990**, *270*, 1.
3. Baker, D.; Agard, D. A. *Biochemistry* **1994**, *33*, 7505.
4. Kolinski, A.; Skolnick, J. *Proteins* **1994**, *18*, 338.
5. Karplus, M.; Sali, A. *Curr. Opin. Struct. Biol.* **1995**, *5*, 58.
6. Dill, K. A.; Bromberg, S.; Yue, K.; Fiebig, K. M.; Yee, D. P.; Thomas, P. D.; Chan, H. S. *Protein Science* **1995**, *4*, 561.
7. Sali, A.; Shakhnovich, E.; Karplus, M. *Nature* **1994**, *369*, 248.
8. Sali, A.; Shakhnovich, E.; Karplus, M. *J. Mol. Biol.* **1994**, *235*, 1614.
9. Hao, M.-H.; Scheraga, H. A. *J. Phys. Chem.* **1994**, *94*, 9882.
10. Poland, D. C.; Scheraga, H. A. *Biopolymers* **1965**, *3*, 379.
11. Pace, C. N.; Grimsley, G. R.; Thomson, J. A.; Barnett, B. J. *J. Biol. Chem.* **1988**, *263*, 11820.
12. Matsumura, M.; Sognor, G.; Thomson, J. A.; Barnett, B. J. *Proc. Natl. Acad. Sci. USA* **1989**, *86*, 6562.
13. Matsumura, M.; Sognor, G.; Thomson, J. A.; Barnett, B. J. *Nature* **1989**, *341*, 291.
14. Chan, H. S.; Dill, K. A. *J. Chem. Phys.* **1989**, *90*, 492.
15. Chan, H. S.; Dill, K. A. *J. Chem. Phys.* **1990**, *92*, 3118.
16. Tidor, B.; Karplus, M. *Proteins* **1993**, *15*, 71.
17. Matsumura, M.; Matthews, B. W. *Methods Enzymol.* **1991**, *202*, 336.
18. Clarke, J.; Fersht, A. R. *Biochemistry* **1993**, *32*, 4322.
19. Shakhnovich, E. I.; Gutin, A. M. *J. Chem. Phys.* **1990**, *93*, 5967.
20. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087.
21. Kuwajima, K. *Proteins* **1989**, *6*, 87.
22. Abkevich, V. I.; Gutin, A. M.; Shakhnovich, E. I. *Biochemistry* **1994**, *33*, 10026.
23. Kim, P. S.; Baldwin, R. *Annu. Rev. Biochem.* **1982**, *51*, 459.
24. Kim, P. S.; Baldwin, R. *Annu. Rev. Biochem.* **1990**, *59*, 631.
25. Matthews, C. R. *Annu. Rev. Biochem.* **1993**, *62*, 653.
26. Matouschek, A.; Serrano, L.; Fersht, A. R. *J. Mol. Biol.* **1992**, *224*, 819-835.
27. Matouschek, A.; Serrano, L.; Meiering, E.; Bycroft, M.; Fersht, A. R. *J. Mol. Biol.* **1992**, *224*, 837.
28. Serrano, L.; Matouschek, A.; Fersht, A. R. *J. Mol. Biol.* **1992**, *224*, 805.
29. Serrano, L.; Matouschek, A.; Fersht, A. R. *J. Mol. Biol.* **1992**, *224*, 847.
30. Baldwin, R. L. *Nature* **1994**, *369*, 183.
31. Shakhnovich, E. I.; Gutin, A. M. *Nature* **1990**, *346*, 773.
32. Hao, M.-H.; Scheraga, H. A. *J. Phys. Chem.* **1994**, *94*, 4940.
33. Skolnick, J.; Kolinski, A. *Science* **1990**, *250*, 1121.
34. Skolnick, J.; Kolinski, A. *J. Mol. Biol.* **1991**, *221*, 499.
35. Vieth, M.; Kolinski, A.; Brooks, C. L.; Skolnick, J. *J. Mol. Biol.* **1994**, *237*, 361.
36. Leopold, P. E.; Montal, M.; Onuchie *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 8721.