

Conformation-networks of two-dimensional lattice homopolymers

Yu-Pin Luo^a, Hung-Yeh Lin^b, Ming-Chang Huang^{a,*}, Tsong-Ming Liaw^c

^a Department of Physics, Chung-Yuan Christian University, Chungli 320, Taiwan

^b Institute of Electrophysics, National Chiao-Tung University, Hsinchu 300, Taiwan

^c Computing Center, Academia Sinica, Nankang 115, Taiwan

Received 6 July 2005; received in revised form 6 June 2006; accepted 8 June 2006

Available online 19 June 2006

Communicated by C.R. Doering

Abstract

The effect of different Monte Carlo move sets on the folding kinetics of lattice polymer chains is studied from the geometry of the conformation-network. The networks have the characteristics of small-world: the local connections are more clustered than that of the corresponding random lattices, and the characteristic path lengths increase logarithmically with the number of nodes. One of the elementary moves, rigid rotation, has drastic effect on the geometric properties of the network. The move increases greatly the connections and reduces significantly the shortest path lengths between conformations. Including rigid rotation to the move set results in the increase of the dimensionality of the conformation space to the value about 4.

© 2006 Elsevier B.V. All rights reserved.

Protein folding is a complex process for which, a sequence of amino acids folds into a unique and stable structure in a relatively short time [1]. The lattice models have been used widely as coarse-grained models for the theoretical study of folding process [2–7]. In the lattice models, protein is viewed as a chain of m monomers, and the conformations are given by all possible self-avoiding walks of the chain on a two or three-dimensional lattice. The energy of a conformation, in general, depends on the number of intrachain contacts, and how to assign the contact energy is model dependent. The kinetics of folding process can then be studied by Monte Carlo simulations for which, a move set is designed for the change of conformations. In principle, different move sets, satisfying the requirement of ergodicity, should reach the same equilibrium canonical distribution after sufficiently long time simulations. However, different move sets may yield different perspectives of folding kinetics. Chan and Dill analyzed the folding kinetics of two different move sets for 2D homo- and hetero-polymers by using the Metropolis transfer matrix method [5,6]. Their

results indicate that a move set adopted for the study affects strongly the kinetic sequence of foldings and the shape of the energy landscape. Same conclusions were also given by Hoang and Cieplak [8] via the comparison between the dynamics of three different move sets. Thus, understanding the nature of a move set is essential for the interpretation of simulation results.

In this Letter, we explore the characteristics of different move sets via the analyses of the corresponding conformation-networks [9,10] for the 2D homopolymers with monomers $m \leq 16$. Though the chain lengths considered are relatively short, the networks can be constructed by exact enumeration. Scala, Amaral, and Barthélemy studied various networks obtained from the mappings of a particular conformation space, and showed that the geometric properties are similar to those of small-world networks [10]. This leads to the question, whether the conformation-networks obtained from different move sets all show the small-world characteristics. There appears two essential characteristics for a small-world network: (i) the local connection is more cliquy than that of random lattices, and (ii) the characteristic path length increases logarithmically with the number of nodes [11,12]. Thus, firstly we analyze the characteristic path lengths and clustering coefficients of the net-

* Corresponding author.

E-mail address: ming@phys.cycu.edu.tw (M.-C. Huang).

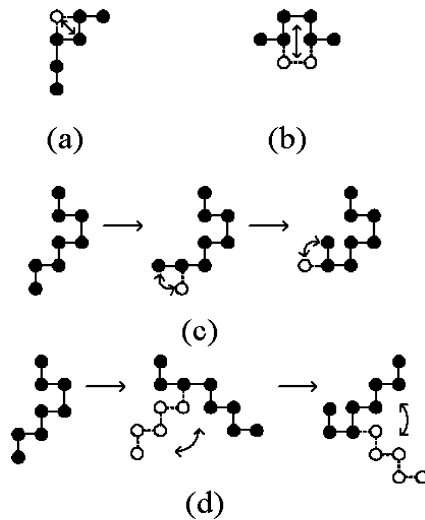


Fig. 1. Examples of typical Monte Carlo moves: (a) end flip, (b) corner shift, (c) crankshaft move, and (d) rigid rotation. The current conformation is shown in thick lines, and possible new conformations are shown in broken lines.

works. To further differentiate the networks, we compute and compare the degree distributions of a node, the correlations between degrees of nearest-neighbors, and the distributions of the distances between two nodes. Finally, we also discuss the stability of the networks.

For the dynamical simulations of lattice polymers, the typical elementary moves include the end flip (ef), corner shift (cs), crankshaft (cr) and the rigid rotation (rr), as shown in Fig. 1. Here, some specially designed moves, such as snake move [8], are excluded from the consideration. We focus the study on the move sets S_1 , S_2 , and S_3 , defined as follows. The conventional move set S_1 consists of ef, cs and cr [7,13,14], based on locality. However, the ergodicity cannot be generally satisfied for S_1 [5,6,8]. In two dimensions, it prohibits the reaching of one conformation from the others for 16 monomers, and the number of such conformations increases rapidly for more monomers and/or dimensions. The problem can be remedied by involving moves of rr type which have been realized in some simple diffusive motions for groups of monomers [15]. While ef itself can be viewed as short-scale rigid rotation, an ergodic move set, say S_2 , can be achieved by simply combining ef with rr. Finally, the ergodic move set S_3 contains all the moves of four types.

For the construction of the network associated with a move set, firstly we identify all possible self-avoiding conformations of the chain of m monomers as the nodes of the network associated with a move set. The node-number is denoted as N_m for which, the degeneracy caused by the rotation and the mirror symmetry has been excluded. Two nodes are then connected by an edge if a move of the given move set can transfer one to the other. Thus, different move sets yield different edge distributions between the nodes and hence different networks. We refer the edge-number as E_m . The values of N_m and E_m for three networks with various numbers of monomers m are listed in Table 1. Note that because all edges are undirected and have the same weight, the networks can be viewed as the folding networks in high temperature limit.

Table 1

Various geometric quantities of the conformation-networks S_1 , S_2 , and S_3 with different number of monomers m : the numbers of nodes N , the numbers of edges E , the average edge number per node $\langle k \rangle$, the characteristic path length $\langle l \rangle$, and the average of clustering coefficients \bar{C}

m	10	12	14	16
N	2034	15037	110188	802075
E_{S_1}	6966	57451	464687	3702485
E_{S_2}	13194	117839	1005304	8314161
E_{S_3}	16397	147673	1268544	10554679
$\langle k \rangle_{S_1}$	6.8496	7.6413	8.4344	9.2323
$\langle k \rangle_{S_2}$	12.9735	15.6732	18.2471	20.7316
$\langle k \rangle_{S_3}$	16.1229	19.6413	23.0251	26.3184
$\langle l \rangle_{S_1}$	7.6369	11.0731	15.0046	19.4403
$\langle l \rangle_{S_2}$	4.5953	5.8286	7.0726	8.3236
$\langle l \rangle_{S_3}$	3.9555	4.9611	5.9723	6.9869
\bar{C}_{S_1}	0.1092	0.0861	0.0684	0.0554
\bar{C}_{S_2}	0.0699	0.0523	0.0434	0.0369
\bar{C}_{S_3}	0.0666	0.0471	0.0366	0.0229

The edge-number associated with a node is also referred as the degree of the node, and the degree distribution $P(k)$ is defined as the probability for a node to have degree k . Then, the mean degree of a network is

$$\langle k \rangle = \sum_k k P(k), \quad (1)$$

which is equal to $2E_m/N_m$. The $\langle k \rangle$ values of different networks with various numbers of monomers m are given in Table 1. The scaling of $\langle k \rangle$ with N_m behaves as $\langle k \rangle = a + b \log(N_m)$ with $(a, b) = (3.79, 0.92)$ for S_1 , $(3.07, 2.99)$ for S_2 , and $(2.77, 4.01)$ for S_3 , as shown in the insets of Fig. 2. Thus, the mean degree of the move set S_2 (S_3) is about two (two and half) times the value of S_1 . A larger value of the mean degree of a network should give more throughway accessibility to the native conformation and reduce the chance of being trapped in local minimum in the folding process [5,6]. The results of $P(k)$ vs. $\Delta k = k - \langle k \rangle$ are shown in Fig. 2 for S_1 , S_2 , and S_3 , respectively, with $m = 10, 12, 14$, and 16. Scala et. al. studied the sub-networks of S_1 for which, a sub-network is specified by a given end-to-end distance and generated by the moves, corner shift and crankshaft move [10]. Their results showed that the form of $P(k)$ is Gaussian. Then, we employ the Gaussian function,

$$P(k) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(k - \langle k \rangle)^2}{2\sigma^2}\right], \quad (2)$$

to fit the data of S_1 , S_2 , and S_3 , and the best fittings are given as the solid lines in Fig. 2. For S_1 , the distribution agrees with the above Gaussian form in which, the variances of different m are $\sigma = (0.5748)\sqrt{N_m}$. Comparing with the result of S_1 , the distribution for S_2 , shown in Fig. 2(b), does not fit so well, and the result of S_3 , shown in Fig. 2(c), exhibits significant deviation, but obviously the distributions are not scale-free [16–18].

The deviation from the Gaussian form for the degree distributions reflects in the asymmetry between the distributions of low and high degrees, and the asymmetry can be clarified further by measuring the degree–degree correlations. The corre-

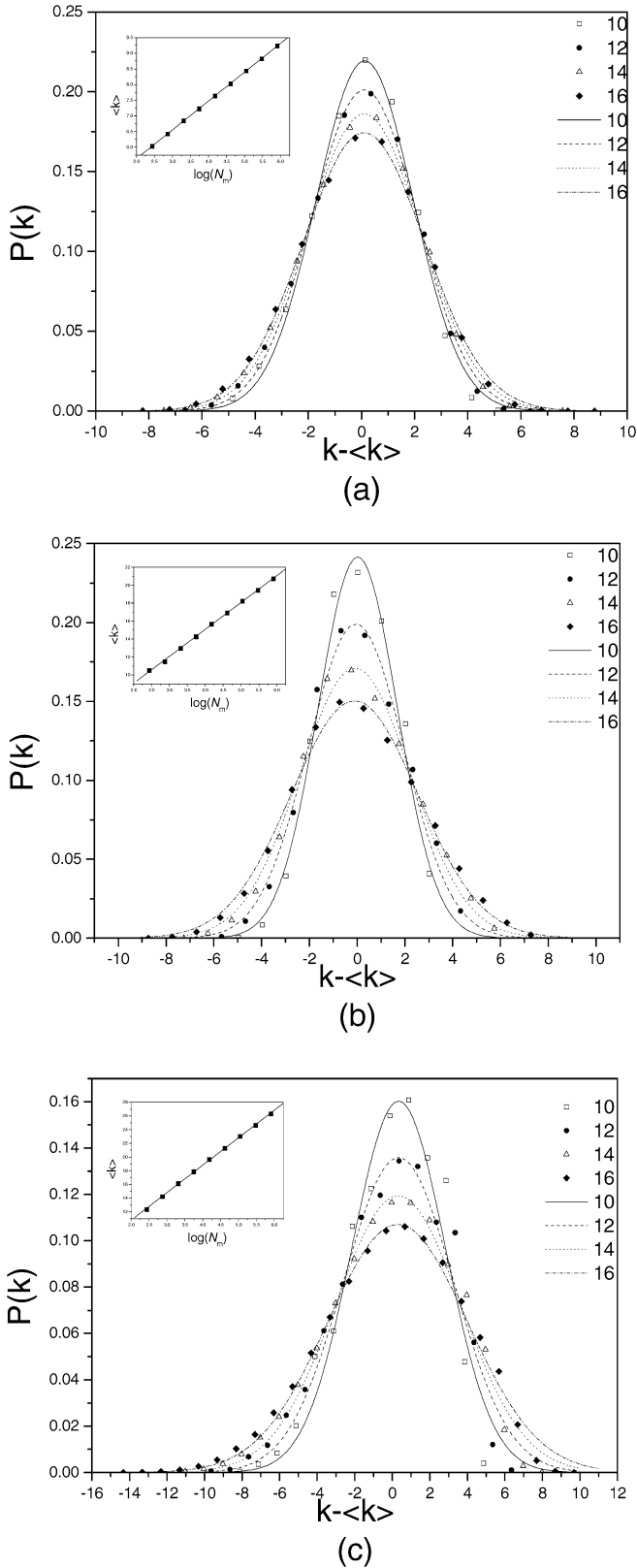


Fig. 2. The degree distribution, $P(k)$, versus $\Delta k = k - \langle k \rangle$ for the networks associated with different move sets: (a) S_1 , (b) S_2 , and (c) S_3 . Here, $\langle k \rangle$ is the average edge number per node, and the solid lines are the best fittings of the Gaussian function given in the text. For each network, the plot of $\langle k \rangle$ vs. $\log(N_m)$ for the node number N_m with the monomer number m ranged from 8 to 16 is shown in the inset, and the straight solid line corresponds to the relation $\langle k \rangle = a + b \log(N_m)$ with the values of a and b given in the text.

lations are characterized by a joint probability $P(k_1, k_2)$, which is defined as the probability of a node with the degree k_1 connected by an edge to another node with degree k_2 . Explicitly, we write

$$P(k_1, k_2) = \frac{1}{2E_m} \sum_{i,j=1}^{N_m} \delta(k_i - k_1) a_{ij} \delta(k_j - k_2), \quad (3)$$

where k_i (k_j) is the degree of node i (j); and a_{ij} is 1 if nodes i and j are connected by an edge, and 0 otherwise. Here, the normalization condition,

$$\sum_{k_1, k_2} P(k_1, k_2) = 1, \quad (4)$$

is imposed. For the absence of correlations, such as the classical random graphs and many other equilibrium networks, the joint probability factorizes, $P^*(k_1, k_2) \propto k_1 P(k_1) k_2 P(k_2)$ [19,20]. After a proper normalization as Eq. (4), we obtain

$$P^*(k_1, k_2) = \frac{1}{A} \{k_1 P(k_1) k_2 P(k_2) [1 + \delta(k_1 - k_2)]\}, \quad (5)$$

with the normalization constant $A = \langle k \rangle^2 + \sum_k k^2 P^2(k)$. For the conformation-networks, elementary moves are independent of one another, and we expect both Eqs. (3) and (4) give about the same result. Owing to the poor statistics for the joint probability, Pastor-Satorras, Vazquez, and Vespignani [19] introduced the average degree of the connected neighbors of a node as a function of the degree of this node, denoted by $\bar{k}(k)$ and defined as

$$\bar{k}(k) = \sum_{k_1} k_1 [P(k_1, k) + P(k, k_1)]. \quad (6)$$

Then, the corresponding $\bar{k}^*(k)$ for the absence of correlations is

$$\bar{k}^*(k) = \frac{1}{A} [2\langle k^2 \rangle k P(k)] \quad (7)$$

with

$$\langle k^2 \rangle = \sum_k k^2 P(k). \quad (8)$$

We then compare $\bar{k}(k)$ with $\bar{k}^*(k)$ for S_1 , S_2 , and S_3 with $m = 16$, and the plots of $\bar{k}(k) - \bar{k}^*(k)$ vs. $k - \langle k \rangle$ are shown in Fig. 3. The properties revealed from the results are as follows. All networks, as expected, have little correlations for which, the values are positive (negative) for large (small) k values. The enhance (positive correlation) and the depletion (negative correlation) in two regions appear almost in a symmetric way for S_1 , but quite asymmetric for S_3 . The asymmetry results from the fact that the additional edges caused by the elementary move rr increases drastically for only few nodes which have high degree in S_1 and the additional number is below the average for most nodes which have low or medium degree. This explains the significant deviation of the degree distribution from the Gaussian form.

The degree of local connections of the networks can be measured by the clustering coefficients. The clustering coefficient

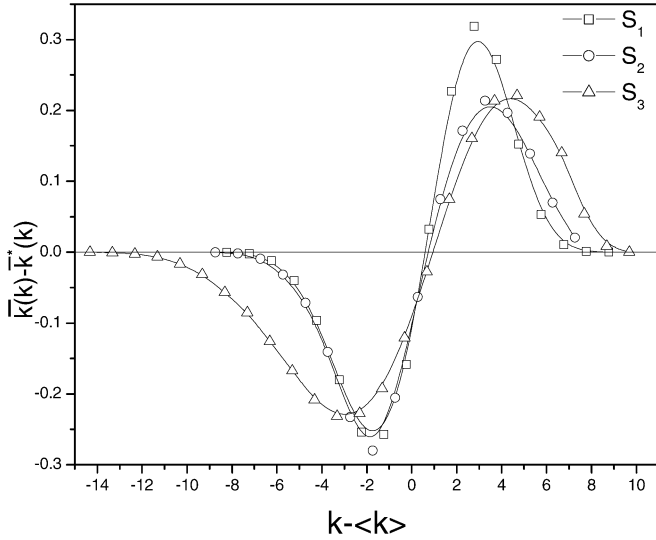


Fig. 3. The plot of $\bar{k}(k) - \bar{k}^*(k)$ vs. $k - \langle k \rangle$ for the networks S_1 , S_2 , and S_3 with $m = 16$. Here, $\bar{k}(k)$ is the average degree of the connected neighbors of a node with degree k , and $\bar{k}^*(k)$ is the result of $\bar{k}(k)$ for the absence of correlations.

of the node i is defined as

$$C_i = \frac{2 \sum k_i}{k_i(k_i + 1)}, \quad (9)$$

where k_i is the degree and $\sum k_i$ is the existent edge-number among the k_i nearest neighbors of the node i . Then, the degree of local connections of a network can be characterized by the average of the clustering coefficients of the nodes, denoted by \bar{C} . The \bar{C} values for S_1 , S_2 , and S_3 with different m values are listed in Table 1. The data shows $\bar{C}_{S_1} > \bar{C}_{S_2} > \bar{C}_{S_3}$. For the network with the node-number N and the average edge-number $\langle k \rangle$, the corresponding random network has the average clustering coefficient $\bar{C}_{\text{ran}} \approx \langle k \rangle / N$. The results of the ratios $\bar{C} / \bar{C}_{\text{ran}}$ vs. the node-number N_m for S_1 , S_2 , and S_3 are shown in Fig. 4 with logarithmic scales. Our results indicate that the average clustering coefficients of the conformation-networks are much larger than that of random networks. In particular, the network of S_1 is less random than that of S_2 and S_3 . Thus, the kinematics based on S_1 have more chances to be trapped in some cliquy conformations than those based on S_2 and S_3 .

We may define the minimum number of elementary moves required for transferring one node to the other as the distance between the two [5,6]. Thus, the distance l between pairs of nodes is the minimum number of edges required to connect the two nodes. The distribution $P(l)$ gives the probability of distance l between two randomly chosen nodes. The characteristic length of the network can be defined as the average of the distances of all node-pairs,

$$\langle l \rangle = \sum_l l P(l). \quad (10)$$

The values of $\langle l \rangle$ for S_1 , S_2 , and S_3 with different m values are listed in Table 1. The characteristic length of S_2 is about half of the length of S_1 . For the distributions $P(l)$, the scaled plots of $P_{\text{scaled}}(l) = \sqrt{2\pi} \sigma P(l)$ vs. $\Delta l_{\text{scaled}} = (l - \langle l \rangle) / \sqrt{2} \sigma$ are shown in Fig. 5, where the solid lines are obtained by setting

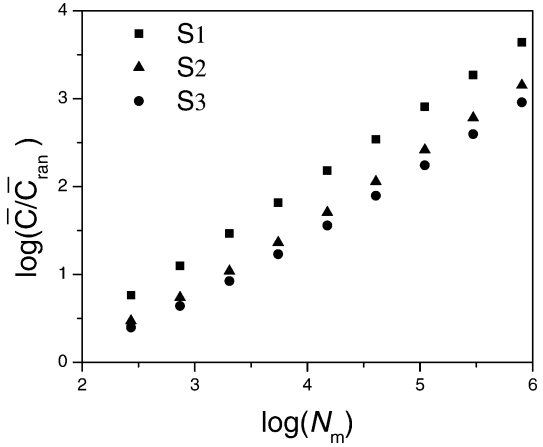


Fig. 4. The ratios of the average clustering coefficients, \bar{C} , of the networks S_1 , S_2 , and S_3 to the average clustering coefficients of the corresponding random networks \bar{C}_{ran} versus $\log(N_m)$ with the node number N_m and the monomer number m ranged from 8 to 16.

the variances σ as $\sigma_{S_1} = 0.0489(m)^{1.7}$, $\sigma_{S_2} = 0.3057(m)^{0.6}$, and $\sigma_{S_3} = 0.5057(m)^{0.6}$, for m monomers. The above variances are determined by first finding the least square fit to Eq. (2) to obtain $\sigma(m)$, and then taking the average over $\sigma(m)$ of different m . The distributions all agree with the Gaussian form of Eq. (2). The variance of $P(l)$ for S_2 is much smaller than that for S_1 , and this implies that the distance between two nodes does not vary much for the networks based on S_2 and S_3 .

For the small-world networks, there exists a cross-over size $N^* \sim p^{-1}$ such that the characteristic lengths $\langle l \rangle$ obey the finite-size scaling law [21–23],

$$\langle l \rangle = (N^*)^{1/d} f\left(\frac{N}{N^*}\right), \quad (11)$$

where d is the dimensionality of the underlying regular lattice, and $f(x)$ is a scaling function with the limits, $f(x) \sim x^{1/d}$ for $x \ll 1$ and $f(x) \sim \ln x$ for $x \gg 1$. By taking the hypothesis that the conformation-network may be a small-world network, we use the scaling form of Eq. (11) to fit the data, and the results are shown in Fig. 6. Note that we do not take the data from $m \leq 4$ for which the node-number N is less than 5, and the statistics for the region of $N/N^* \ll 1$ is very poor in our results. However, the fittings indicate that (i) the values of $\langle l \rangle$ increase logarithmically with the node-number N for large N ; (ii) we estimate $1/d$ from the fittings of small N as 0.3427, 0.2377, and 0.2155 for the networks S_1 , S_2 , and S_3 , and then the dimensions d are about 3, 4, and 4.5, respectively; and (iii) the cross-over region $N^*(m)$ is around $m = 9$ – 11 ($p \sim 10^{-3}$ – 10^{-4}) for S_1 and 8 ($p \sim 10^{-3}$) for S_2 and S_3 . But, based on the above results, we may conclude that the dimensionality of the conformation-space is $d \geq 3$, and the cross-over region become narrower when the dimensionality gets larger. The $\langle l \rangle$ value may be viewed as the diameter of a network [17]. Then, while as the dimension increases for the networks from S_1 to S_3 sequentially, the diameter of the network decreases. We notice that the dimension obtained by Scala, Amaral, and Barthélemy is 2 [10], and our result is about 3 for the network associated with S_1 . The difference is

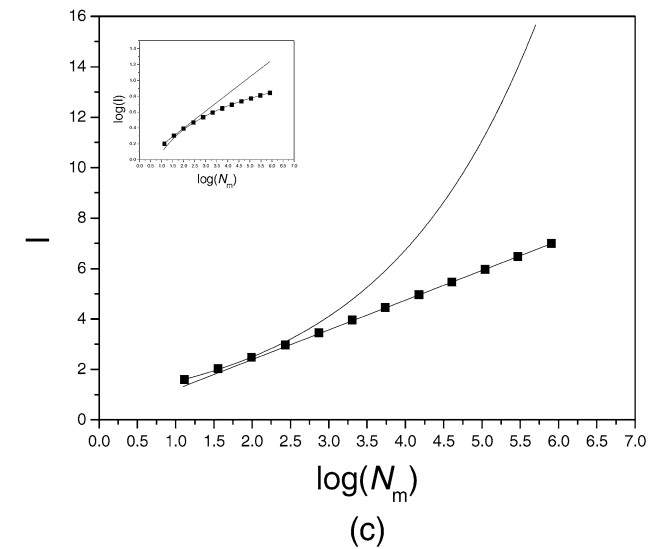
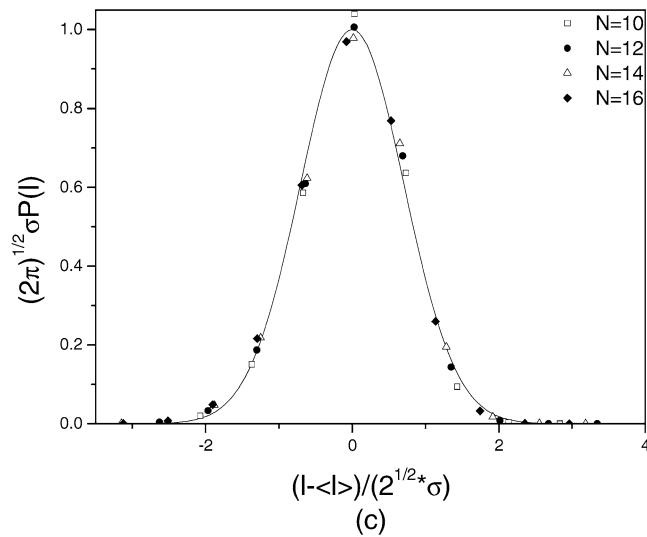
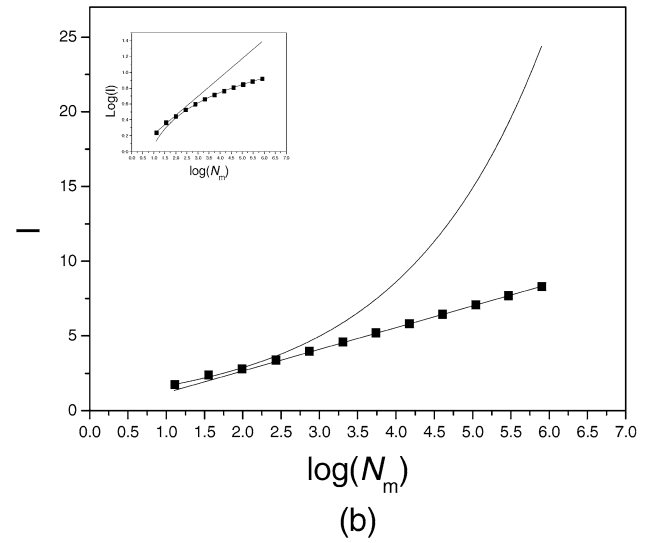
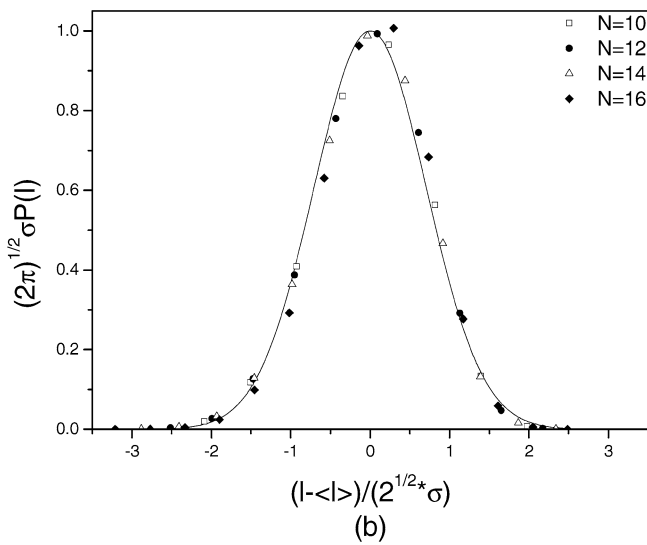
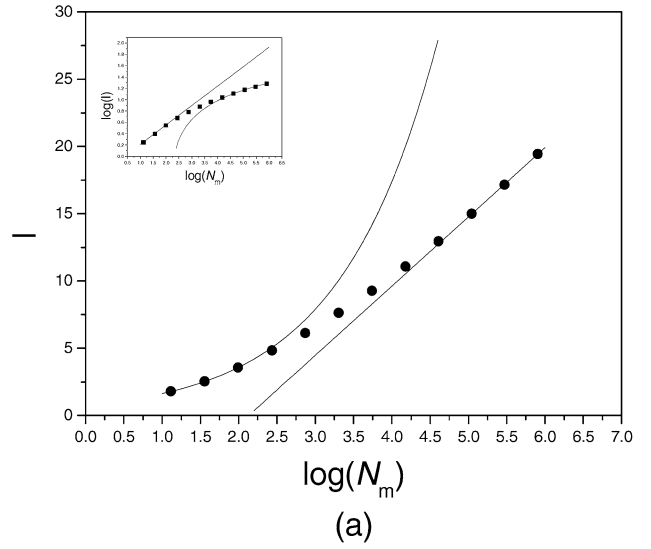
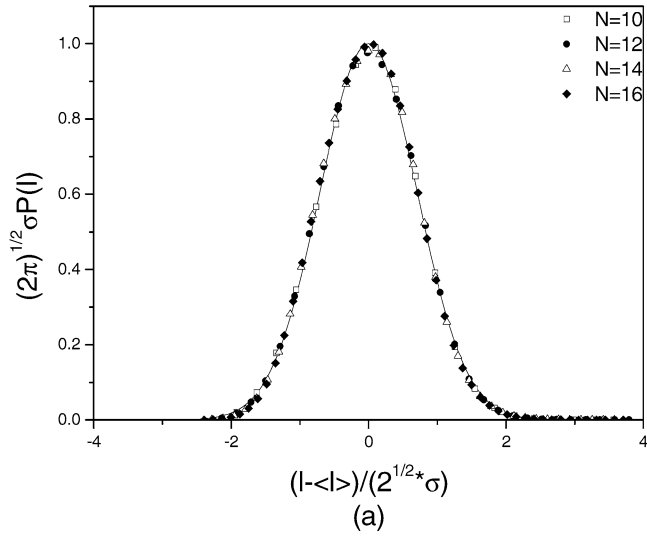


Fig. 5. The scaled result of the distribution function of the shortest path lengths, $P_{\text{scaled}}(l) = \sqrt{2\pi} \sigma P(l)$, versus $\Delta l_{\text{scaled}} = (l - \langle l \rangle) / \sqrt{2} \sigma$ for (a) S_1 , (b) S_2 , and (c) S_3 with $m = 10, 12, 14$, and 16 . The averages of the shortest path lengths for all node-pairs, $\langle l \rangle$, are given in Table 1, and the variances σ are given in the text. The solid lines are the results of the Gaussian form.

Fig. 6. The plots of the characteristic path length $\langle l \rangle$ versus the logarithm of the node-number $\log(N_m)$, for the networks associated with different move sets, (a) S_1 , (b) S_2 , and (c) S_3 , where the monomer number m ranges from 5 to 16. The insets are the plots of $\log(\langle l \rangle)$ versus $\log(N_m)$ for the same data. The solid lines are the results of the limiting scaling forms given in the text.

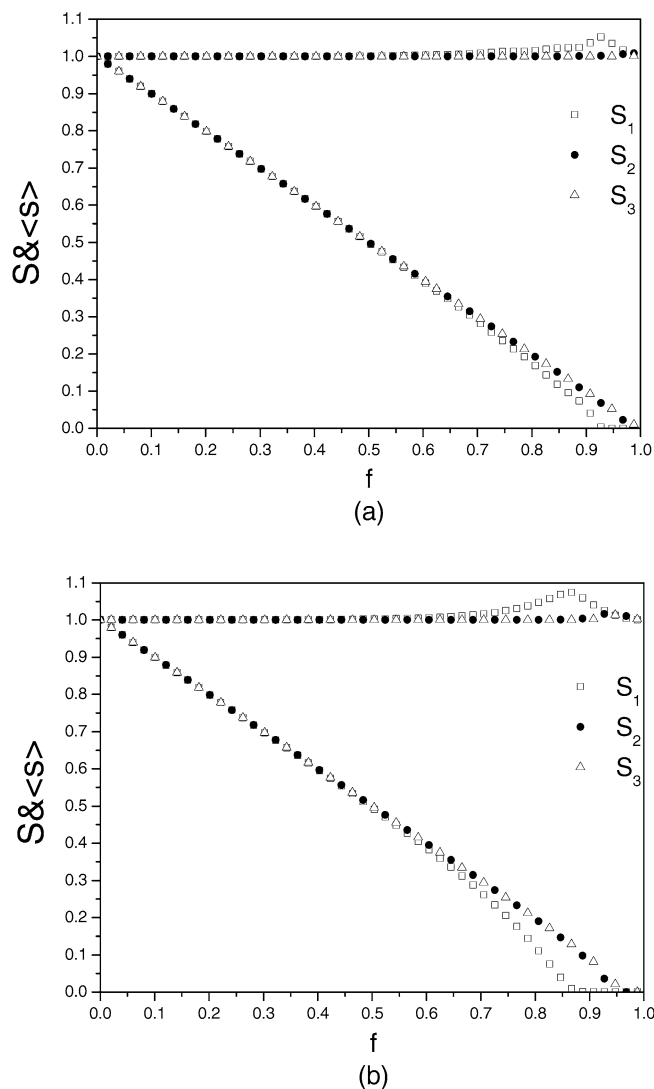


Fig. 7. The fraction of nodes contained in the largest cluster, S , and the average node number, $\langle s \rangle$, contained in the fragmentary clusters excluding the largest one versus the fraction f of the nodes removed for (a) attack and (b) error tolerance of the networks S_1 , S_2 , and S_3 with $m = 16$.

due to the fact that the previous results are based on the networks with fixed end-to-end distances of the chain, such networks exclude the elementary move ef and are sub-networks of the networks of S_1 .

Finally, we analyze the ability of attack and error tolerance of the network by studying the fragmentation caused by node-removal [24]. The nodes with higher degrees of connections are removed preferentially for the analysis of attack tolerance; and the nodes are removed randomly for the error tolerance. By removing a fraction f of the nodes, we measure the fraction of nodes contained in the largest cluster, S , and the average node number, $\langle s \rangle$, contained in the fragmentary clusters excluding the largest one. If only the removed nodes were missing from without further breaking the largest cluster, the S value decreases from 1 down to 0 along the diagonal line as f increases from 0 up to 1; and the $\langle s \rangle$ value remains to be one for $0 < f \leq 1$ if the removed nodes were isolated from each other. For most networks, we may expect that while as the S values

start to decrease more rapidly than the diagonal line at some fraction f_m , and drop to zero at the critical fraction f_c ; the $\langle s \rangle$ value start to increase more rapidly from $\langle s \rangle = 1$ at f_m , and reach the maximum at f_c . The results of S and $\langle s \rangle$ as function of f are shown in Fig. 7 for the networks S_1 , S_2 , and S_3 with $m = 16$. Our results show that the f_c value is very closed to 1, and the stability of the networks is very analogous to random networks.

In summary, we divide the frequently used Monte Carlo moves into three different move sets, and construct the corresponding conformation-networks. The networks all have the characteristics of small-world: (i) the local neighborhood is more cliquy than that of random networks, and (ii) the characteristic path length increases logarithmically with the number of nodes. The dimensionalities of the conformation-spaces are $d \geq 3$. Our analyses also indicate that the networks are as robust as random graphs. Among different elementary moves, the rigid rotation has drastic effect on the geometric properties of the network: (i) it renders the connection distribution to be non-Gaussian, (ii) it reduces greatly the characteristic path length, and (iii) it drives the network more closer to random networks. Thus, the rigid rotation may change the folding kinetics significantly from that of the local moves, corner shift and crankshaft move.

Acknowledgements

This work was partially supported by the National Science Council of Republic of China (Taiwan) under the Grant No. NSC 93-2212-M-033-005. We thank the National Center for High-performance Computing and the Computing Centre of Academia Sinica for providing the computation facilities.

References

- [1] T.E. Creighton (Ed.), Protein Folding, Freeman, New York, 1992.
- [2] J. Skolnick, A. Kolinski, Annu. Rev. Phys. Chem. 40 (1989) 207.
- [3] E. Shakhnovich, G. Farztdinov, A.M. Gutin, M. Karplus, Phys. Rev. Lett. 67 (1991) 1665.
- [4] R. Miller, C.A. Danko, J. Fasolka, A.C. Balazs, H.S. Chan, K.A. Dill, J. Chem. Phys. 96 (1992) 768.
- [5] H.S. Chan, K.A. Dill, J. Chem. Phys. 99 (1993) 2116.
- [6] H.S. Chan, K.A. Dill, J. Chem. Phys. 100 (1994) 9238.
- [7] A. Sali, E. Shakhnovich, M. Karplus, J. Mol. Biol. 235 (1994) 1614.
- [8] T.X. Hoang, M. Cieplak, J. Chem. Phys. 109 (1998) 9192.
- [9] L.A.N. Amaral, A. Scala, M. Barthélémy, H.E. Stanley, Proc. Natl. Acad. Sci. USA 97 (2000) 11149.
- [10] A. Scala, L.A.N. Amaral, M. Barthélémy, Europhys. Lett. 55 (2001) 594.
- [11] D.J. Watts, D.H. Strogatz, Nature (London) 393 (1998) 440.
- [12] D.J. Watts, Small Words: The Dynamics of Networks Between Order and Randomness, Princeton Univ. Press, Princeton, NJ, 1999.
- [13] N.D. Socci, J.N. Onuchic, J. Chem. Phys. 101 (1994) 1519; N.D. Socci, J.N. Onuchic, J. Chem. Phys. 103 (1995) 4732.
- [14] R. Melin, H. Li, N.S. Wingreen, C. Tang, J. Chem. Phys. 110 (1999) 1252.
- [15] J. Skolnick, A. Kolinski, J. Mol. Biol. 235 (1994) 1614.
- [16] A.-L. Barabási, R. Albert, Science 286 (1999) 509.
- [17] R. Albert, H. Jeong, A.-L. Barabási, Nature (London) 401 (1999) 130.
- [18] H. Jeong, B. Tombor, R. Albert, Z.N. Oltavi, A.-L. Barabási, Nature (London) 407 (2000) 651.
- [19] R. Pastor-Satorras, A. Vazquez, A. Vespignani, Phys. Rev. Lett. 87 (2001) 258701.

- [20] S.N. Dorogovtsev, J.F.F. Mendes, *Evolution of Networks: From Biological Nets to The Internet and WWW*, Oxford Univ. Press, New York, 2003.
- [21] M. Barthélémy, L.A.N. Amaral, *Phys. Rev. Lett.* 82 (1999) 3180;
M. Barthélémy, L.A.N. Amaral, *Phys. Rev. Lett.* 82 (1999) 5180, Erratum.
- [22] A. Barrat, *cond-mat/9903323*.
- [23] M.E.J. Newman, D.J. Watts, *Phys. Lett. A* 263 (1999) 341.
- [24] R. Albert, H. Jeong, A.-L. Barabási, *Nature (London)* 406 (2000) 378;
R. Albert, H. Jeong, A.-L. Barabási, *Nature (London)* 409 (2001) 542, Erratum.