

# Topological complexity, contact order, and protein folding rates

P. F. N. Faisca<sup>a)</sup>

*Department of Physics, University of Warwick, Coventry CV4 7AL, United Kingdom and CFTC, Av. Prof. Gama Pinto 2 1649-003 Lisboa Codex, Portugal*

R. C. Ball

*Department of Physics, University of Warwick, Coventry CV4 7AL, United Kingdom*

(Received 4 June 2002; accepted 12 August 2002)

Monte Carlo simulations of protein folding show the emergence of a strong correlation between the relative contact order parameter,  $CO$ , and the folding time,  $t$ , of two-state folding proteins for longer chains with number of amino acids  $N \geq 54$ , and higher contact order,  $CO > 0.17$ . The correlation is particularly strong for  $N = 80$  corresponding to slow and more complex folding kinetics. These results are qualitatively compatible with experimental data where a general trend towards increasing  $t$  with  $CO$  is indeed observed in a set of proteins with chain length ranging from 41 to 154 amino acids. © 2002 American Institute of Physics. [DOI: 10.1063/1.1511509]

## I. INTRODUCTION

The search for correlations between the protein folding kinetics and the native state equilibrium properties (i.e., chain length and stability) presents a major challenge for those working in the field of protein folding, both in theory and experiments. Progress has been significantly hindered by difficulty analyzing the folding of protein molecules larger than about 100 amino acids, whose kinetics is widely believed to be based on some multiexponential mechanism.<sup>1</sup> By contrast, for smaller proteins whose folding kinetics is close to single exponential, there seems to be some consensus as to the dependence of the folding time,  $t$ , on native state stability.<sup>2-5</sup>

Experiment and theory appear to be at odds with each other over the dependence of the folding time on the number  $N$  of amino acids in the folding unit. Recent Monte Carlo (MC) simulations<sup>5</sup> of a simple lattice model have proposed that, for two-state proteins, a scaling law of the type  $t \approx N^\lambda$ ,  $\lambda \approx 5$ , appropriately describes the dependence of the folding time on the chain length,  $N$ ; a weaker dependence ( $\lambda \approx 4$ ) has been previously reported in Ref. 6 for the same model Hamiltonian and distribution of contact energies, and in Ref. 7 for a two-letter alphabet model that, apart from the commonly used isotropic contact interactions, also considers orientation-dependent interactions. However, available experimental data shows no correlation between  $t$  and  $N$ .<sup>2,3,8</sup>

We examine here the influence of the native state geometric properties on the protein folding kinetics in the context of MC simulations. One simple parameter of the geometry which has already attracted attention is contact order, measuring the average length of the backbone loops connecting contacting pairs of residues in the structure.<sup>2</sup> Formally, the relative contact order,  $CO$ , is defined as

$$CO = \frac{1}{LN} \sum_{ij}^N \Delta_{ij} |i - j|, \quad (1)$$

where  $N$  is the total number of amino acid residues in the protein,  $L$  is the total number of contacts, and  $\Delta_{ij} = 1$  if residues  $i$  and  $j$  are in contact and is 0 otherwise,  $|i - j|$  is the backbone separation between residues  $i$  and  $j$ . High values of  $CO$  are associated with protein structures where amino acid residues interact on average with others that are far away in sequence (long-range interactions), while those displaying predominantly local interactions are of low contact order.

A high correlation was found between the  $CO$  parameter, and the folding rates for the protein set considered in Ref. 3: Proteins with “low” contact order tend to fold faster than proteins with “high” contact order. This finding strongly supports the view that native geometry strongly influences the kinetics of the rate-limiting step in the two-state mechanism of small protein molecules ( $N < 100$ ) determining their folding rates.

The connection between  $CO$  and the dominant range of residue interactions brings back the controversial issue of the importance of local (and nonlocal) contacts in the dynamics of protein folding. An argument against local contacts is that they might increase the “roughness” of the energy landscape, and therefore the stability of the unfolded state.<sup>9</sup> On the other hand, an argument supporting local interactions is based on the idea that they might provide the ideal substrate for the development of nucleation or initiation sites, small local sequence substructures forming in an early stage of folding and driving the subsequent pathway.<sup>10</sup> Moreover, the formation of nonlocal contacts in early folding is entropically costly as it restricts the number of conformations available to the folding unit.<sup>11</sup>

The limited amount of experimental information available<sup>2,3,12</sup> suggests that investigating this problem within the scope of theoretical models could give more insight.

## II. MODEL AND METHODS

To achieve this goal we consider a simple three dimensional lattice model of a protein molecule whose Hamiltonian is given by the contact approximation,

<sup>a)</sup>Electronic mail: patnev@alf1.cii.fc.ul.pt

TABLE I. The target fraction found for several consecutive intervals of the relative contact order parameter; for each studied chain length,  $N$ , a sample of 400 targets was considered. For  $CO \geq 0.24$ , a maximum target fraction of  $0.015 \pm 0.083$  was found for chain length  $N=36$ .

$N$	$CO$ range					
	(0.12,0.14)	(0.14,0.16)	(0.16,0.18)	(0.18,0.20)	(0.20,0.22)	(0.22,0.24)
36	...	$0.0475 \pm 0.0258$	$0.2150 \pm 0.1058$	$0.3825 \pm 0.1669$	$0.2625 \pm 0.0801$	$0.0775 \pm 0.0413$
48	$0.0025 \pm 0.0019$	$0.1575 \pm 0.0602$	$0.3025 \pm 0.1053$	$0.3350 \pm 0.1138$	$0.1350 \pm 0.0523$	$0.0600 \pm 0.0242$
54	$0.0500 \pm 0.0180$	$0.2325 \pm 0.0754$	$0.3425 \pm 0.1028$	$0.2450 \pm 0.0788$	$0.1075 \pm 0.0376$	$0.0225 \pm 0.0082$
64	$0.0250 \pm 0.0077$	$0.1950 \pm 0.0546$	$0.4125 \pm 0.0988$	$0.2325 \pm 0.0636$	$0.1150 \pm 0.0338$	$0.0200 \pm 0.0062$
80	$0.0685 \pm 0.0167$	$0.2567 \pm 0.0559$	$0.3374 \pm 0.0694$	$0.2493 \pm 0.0546$	$0.0759 \pm 0.0184$	$0.0098 \pm 0.0025$

$$H(\{\sigma_i\}, \{\mathbf{r}_i\}) = \sum_{i>j}^N \epsilon(\sigma_i, \sigma_j) \Delta(\mathbf{r}_i - \mathbf{r}_j), \quad (2)$$

where  $\{\sigma_i\}$  stands for an amino acid sequence ( $\sigma_i$  being the chemical identity of bead  $i$ ) while  $\{\mathbf{r}_i\}$  is the set of bead coordinates that define a certain conformation. The contact function,  $\Delta$ , equals 1 if beads  $i$  and  $j$  are in contact but not covalently linked, and is 0 otherwise. We follow many previous studies in taking the interaction parameters  $\epsilon$  from the  $20 \times 20$  Miyazawa–Jernigan matrix derived from the distribution of contacts in native proteins.<sup>13</sup> Our folding simulations follow the standard MC Metropolis algorithm<sup>14</sup> and the kink-jump MC move set (end-move, corner-flip, and crankshaft).<sup>15</sup> Periodic boundary conditions are applied so that, in practice, the lattice is effectively infinite.

### III. NUMERICAL RESULTS

#### A. Contact order and homopolymer kinetics

We have explored the distribution of the relative contact order parameter over a population of 2000 maximally compact target geometries found by homopolymer relaxation.<sup>16</sup> This distribution is shown in Table I for each of the studied chain lengths  $N=36, 48, 54, 64$ , and  $80$ , which are all commensurate with folding to fill a simple cuboid. There is only a slight shift in the modal contact order with chain length,

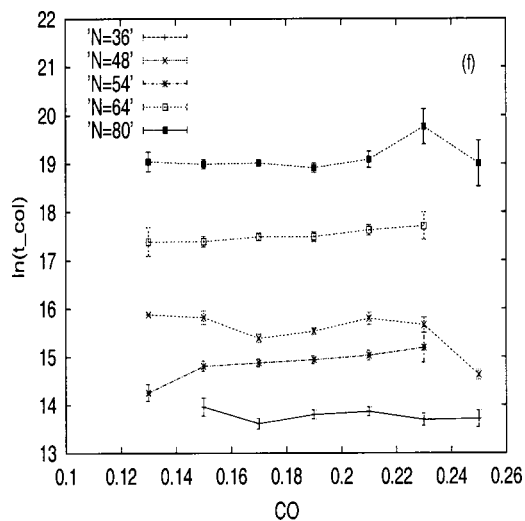


FIG. 1. Dependence of the collapsing time,  $t_{\text{col}}$ , on  $CO$ ,  $t_{\text{col}}$  was computed as the mean first passage time averaged over 400 simulation runs.

form 0.19 for  $N=36, 48$  down to 0.17 for higher  $N$ . However, the target fraction found for  $CO \geq 0.22$  is significantly smaller for  $N \geq 54$  than for shorter chains.

Interestingly, the values of  $CO$  found for these lattice proteins span approximately the same range as those found in real kinetically characterized single domain proteins ( $0.0745 \leq CO \leq 0.2120$ ).<sup>3</sup>

We have also checked the intrinsic kinetic accessibility of the compact configurations obtained, by measuring the time  $t_{\text{col}}$  for these configurations to be reached under homopolymer relaxation. Figure 1 shows there is no evident correlation between  $t_{\text{col}}$  and  $CO$ .

#### B. Finding the optimal folding temperature

To investigate the relationship between protein folding and contact order (at least) 20 target conformations for each chain length were selected so as to sample uniformly across the range of contact order. For each target an ensemble of 100 designed sequences was prepared by using the design method developed by Shakhnovich and Gutin<sup>17</sup> based on random heteropolymer theory, and simulated annealing techniques. The average trained sequence energy,  $\langle E \rangle$ , is shown in Table II along with the standard deviation of the energy distribution,  $\sigma$ . Except for  $N=54$ , the chemical composition of the designed sequences was the same as the one used in Ref. 5. In that study it was shown that the optimal folding temperature,  $T_{\text{fold}}(N)$ , defined as the temperature that minimizes the folding time, is close to a self-averaging parameter.

Since the Shakhnovich and Gutin design scheme preserves the overall sequence chemical composition we can safely use for the 36 and 48 bead long sequences studied here the  $T_{\text{fold}}(36)$  and  $T_{\text{fold}}(48)$  found in Ref. 5.

TABLE II. Average sequence energy under the training scheme of Shakhnovich and Gutin. For each chain length,  $N$ , the average is computed over the total number of sequences ( $\approx 2000$ ) designed for each set of targets.  $\sigma$  is the standard deviation of the mean.

$N$	$\langle E \rangle$	$\sigma$
36	-15.8722	0.0194
48	-23.1407	0.0331
54	-28.8028	0.0158
64	-35.0811	0.0477
80	-46.2858	0.0375

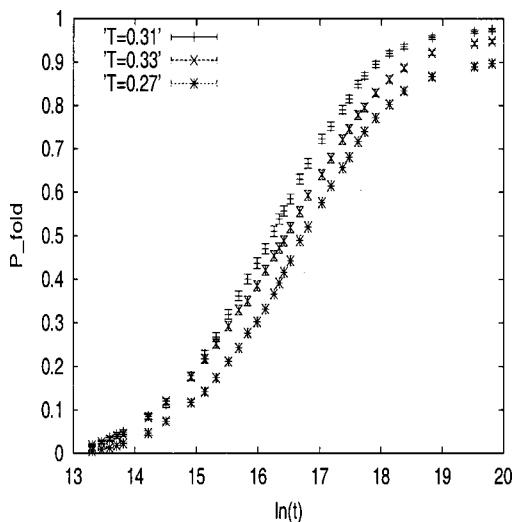


FIG. 2. Dependence of the folding probability,  $P_{\text{fold}}$ , on  $\log(t)$  at three different temperatures.  $P_{\text{fold}}$  was computed as the number of folding simulations which ended up to time  $t$  normalized to the total number of attempted runs. For each curve  $\approx 2000$  simulations were used distributed across the available 48 bead long targets and sequences.

For the longer  $N$ , and most particularly  $N=80$ , foldicity, defined as the fraction of successful folding runs over the total number of attempted runs, was for the vast majority of the targets less than unity. This forced us to define the optimal folding temperature,  $T_{\text{fold}}(N)$ , in such cases as the temperature which optimized foldicity rather than the temperature which minimized the folding time. The case  $N=48$  sits at the margin and provides confirmation that the two approaches to  $T_{\text{fold}}(N)$  are not in conflict, as shown in Fig. 2.

### C. Contact order and folding kinetics

After determining  $T_{\text{fold}}(N)$  we ran a MC folding simulation for every designed sequence. The simulations proceeded until  $\tau_{\text{max}}(N)$  MC steps or until folding was observed. The value of  $\tau_{\text{max}}(N)$  was chosen such that it was much longer than the typical folding time of the studied sequences.

Figure 3 shows the dependence of the folding time,  $t$ , on the contact order parameter for chain lengths,  $N=36$  and  $N=48$ . The folding time was computed as the mean first passage time averaged over 100 simulation runs. In either case the points are close to be uniformly distributed suggesting no correlation between the  $CO$ , and the folding time for these chain lengths. Here and elsewhere error bars indicate  $\pm$ one standard error in the mean.

Figures 4(a)–4(c) show the dependence of foldicity on  $CO$  for  $N=54$ , 64, and 80, respectively. The results presented in Figs. 4(d)–4(f) show, for the same chain lengths, the dependence of the estimated folding time on the relative contact order parameter. Two distinct scenarios emerge from the analysis of the graphs:

(1) For  $CO \leq 0.17$  there is no correlation between foldicity (or folding time) and the relative contact order parameter.

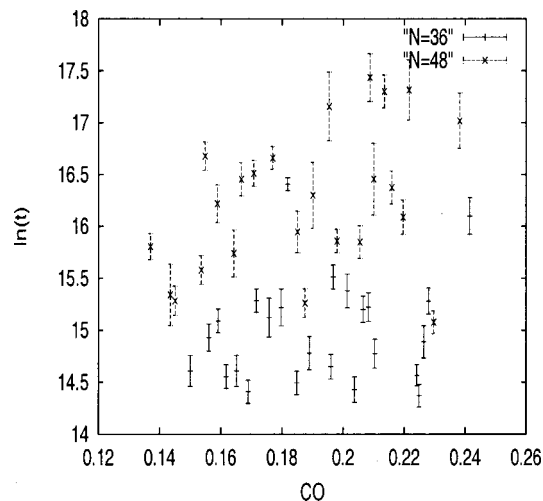


FIG. 3. Dependence of the folding time,  $t$ , on the relative contact order,  $CO$ , for 36 and 48 bead long targets.

(2) For  $CO > 0.17$ , a general trend towards decreasing foldicity with increasing relative contact order can be observed. In this regime, a considerably strong positive correlation of  $r=0.70$ ,  $0.70$ , and  $0.79$ , between  $t$  and  $CO$ , shows up for chain lengths  $N=54$ , 64, and 80, respectively.

## IV. DISCUSSION AND CONCLUSIONS

The “turning point” value of  $CO=0.17$  is actually the peak of the homopolymer relaxation histogram distribution as previously discussed. This means that  $CO$  and folding time are positively correlated only for proteins with predominantly nonlocal contacts. We interpret this result as a consequence of the properties of the move set used to explore the conformational space together with the ruggedness of the energy landscape. As seen in Sec. III A, kink-jump dynamics does not favor the formation of high  $CO$  structures in homopolymers. In proteins, when the native structure is of high  $CO$ , it will be difficult to escape from kinetic traps associated with local energy minima and structures of lower  $CO$ . This confirms and explains our previous findings<sup>5</sup> according to which the folding performance achievable is strongly sensitive to target conformation for chain lengths  $N \geq 80$ .

The comparison of the simulation’s results with the experimental data on a set of 24 two-state proteins, with chain length ranging from 41 to 154 amino acids, reported in Ref. 3 is hindered by the fact that the proteins considered in Ref. 3 fail to exhibit the scaling of folding time with chain length which is typical of lattice model simulations. However, a strong correlation ( $r=0.80$ ) is also found between  $CO$  and the folding times. Moreover, this correlation is considerably improved ( $r=0.97$ ) if only long protein chains ( $N \geq 80$ ) are considered.

As a general conclusion, we might say that results on lattice models encourage the idea that the contact order of the

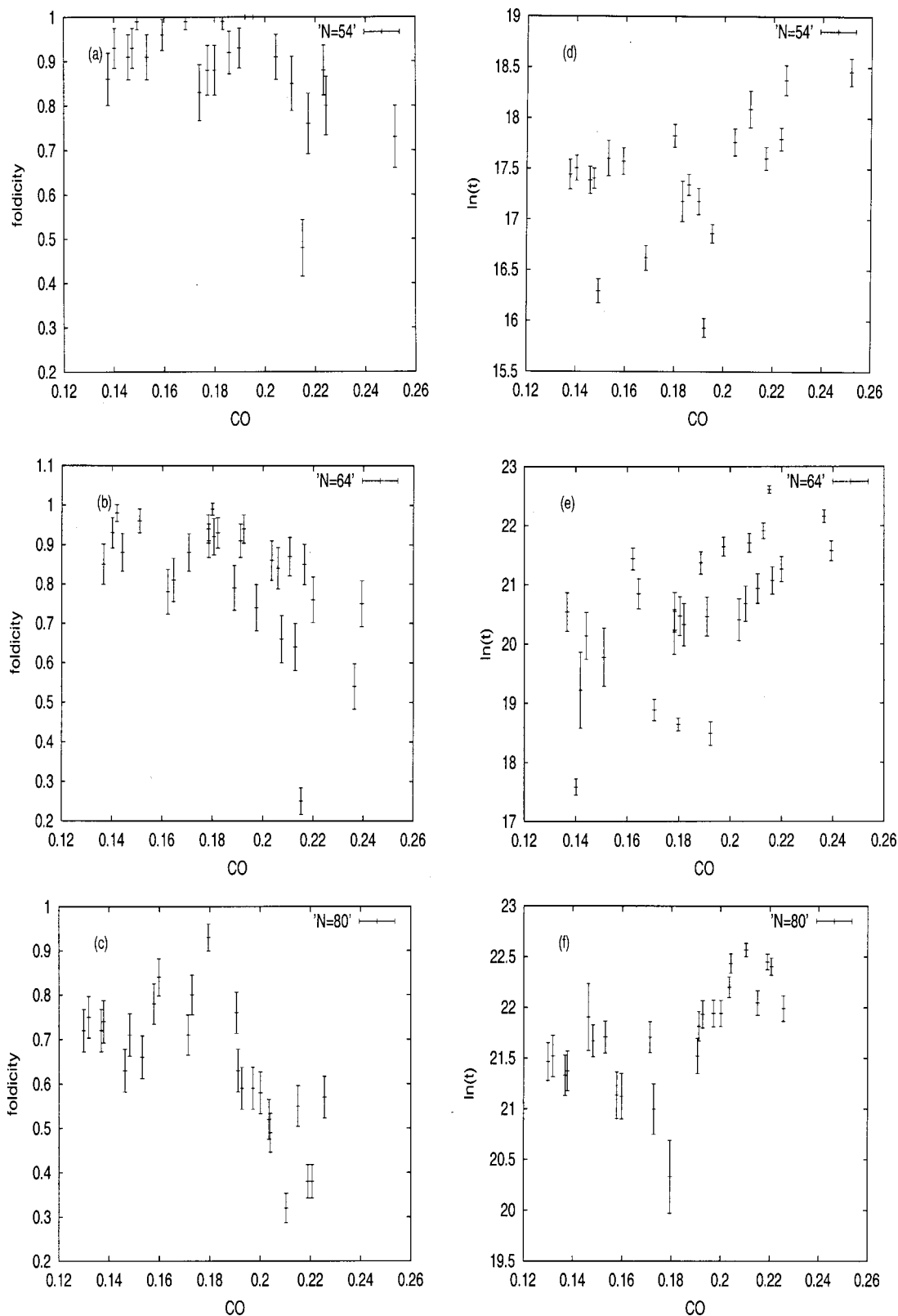


FIG. 4. (a)–(c) Dependence of foldicity on the relative contact order,  $CO$ , for 54, 64, and 80 bead long targets. (d)–(f) Show the dependence of the (estimated) folding time,  $t$ , on  $CO$ . Since we took the number of MC steps to be equal to  $\tau_{\max}(N)$  for every run in which the folded state was not found within  $\tau_{\max}(N)$  steps, the folding times and respective error bars are biased towards smaller values.

native structure plays a significant role in determining the folding rate. The match with the correlation between the  $CO$  and the folding time found from the analysis of experimental data suggests that lattice polymer dynamics

with local moves does capture the key dynamical features of real protein folding.

It would be interesting to know if similar results can be found in the scope of “off-lattice” models, where one

would expect proteins with high helical content to be better folders.

## ACKNOWLEDGMENTS

One of the authors (P.F.N.F.) would like to thank Dr. A. Nunes for helpful suggestions and Programa Praxis XXI for financial support.

<sup>1</sup>A. R. Fersht, Proc. Natl. Acad. Sci. U.S.A. **97**, 1525 (2000).

<sup>2</sup>K. W. Plaxco, K. T. Simons, and D. Baker, J. Mol. Biol. **277**, 985 (1998).

<sup>3</sup>K. W. Plaxco, K. T. Simons, I. Ruczinski, and D. Baker, Biochemistry **39**, 11177 (2000).

<sup>4</sup>T. M. Fink and R. C. Ball, Physica D **107**, 199 (1997).

<sup>5</sup>P. F. N. Faisca and R. C. Ball, J. Chem. Phys. **116**, 7231 (2002).

<sup>6</sup>A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich, Phys. Rev. Lett. **77**, 5433 (1996).

<sup>7</sup>K. Dimitrievski, B. Kasemo, and V. P. Zhdanov, J. Chem. Phys. **113**, 883 (2000).

<sup>8</sup>A. P. Demchenko, Curr. Prot. and Peptide Sci. **2**, 73 (2001).

<sup>9</sup>V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, J. Mol. Biol. **252**, 460 (1995).

<sup>10</sup>D. B. Wetlaufer, Proc. Natl. Acad. Sci. U.S.A. **70**, 697 (1973).

<sup>11</sup>D. Baker, Nature (London) **405**, 39 (2000).

<sup>12</sup>M. O. Lindberg, J. Tangrot, D. E. Otzen, D. A. Dolgikh, A. V. Finkesstein, and M. Oliveberg, J. Mol. Biol. **314**, 891 (2001).

<sup>13</sup>S. Miyazawa and R. Jernigan, Macromolecules **18**, 534 (1985).

<sup>14</sup>N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

<sup>15</sup>D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, Cambridge, 2000).

<sup>16</sup>V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, J. Mol. Biol. **252**, 460 (1995).

<sup>17</sup>E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90**, 7195 (1993).