Journal of Biological Physics **26:** 273–294, 2000.

Scaling of Folding Properties in Go Models of Proteins

MAREK CIEPLAK and TRINH XUAN HOANG
Institute of Physics, Polish Academy of Sciences, Al. Lotnikow 32/46, 02-668 Warsaw, Poland

Abstract. Insights about scaling of folding properties of proteins are obtained by studying folding in heteropolymers described by Go-like Hamiltonians. Both lattice and continuum space models are considered. In the latter case, the monomer-monomer interactions correspond to the Lennard-Jones potential. Several statistical ensembles of the two- and three-dimensional target native conformations are considered. Among them are maximally compact conformations which are confined to a lattice and those which are obtained either through quenching or annealing of homopolymers to their compact local energy minima. Characteristic folding times are found to grow as power laws with the system size. The corresponding exponents are not universal. The size related deterioration of foldability is found to be consistent with the scaling behavior of the characteristic temperatures: asymptotically, the folding temperature becomes much lower than the temperature at which glassy kinetics become important. The helical conformations are found to have the lowest overall scaling exponent and the best foldability among the classes of conformations studied. The scaling properties of the Go-like models of the protein conformations stored in the Protein Data Bank suggest that proteins are not optimized kinetically.

Key words: Go model, molecular dynamics, protein folding, scaling properties

1. Introduction

Protein sequences found in nature are roughly between 30 and 5000 monomer long. Figure 1 shows that a randomly sampled distribution of the sequence lengths, N, is peaked around $N = 100$ and then it falls down so that there are few proteins with an N that is larger than 1300. Furthermore, all larger sized proteins are multi-domained. Why are there no proteins of much bigger sizes, such as, for instance, corresponding to typical DNA sequences? An answer to this question is related to a more general issue: how do folding properties of proteins scale with N?

A natural choice of a quantity for studies of scaling is to consider a characteristic folding time, t_{fold}, as discussed by Thirumalai [2]. t_{fold} was determined numerically first by Gutin, Abkevich, and Shakhnovich [3] and then by Zhdanov [4] and Cieplak, Hoang and Li [5] for several lattice models. In this paper, we discuss scaling of t_{fold} in off-lattice models.

Values of t_{fold} depend on temperature, T, at which folding takes place. There are two characteristic temperatures, T_f and T_{min}, that we focus on here. Studying

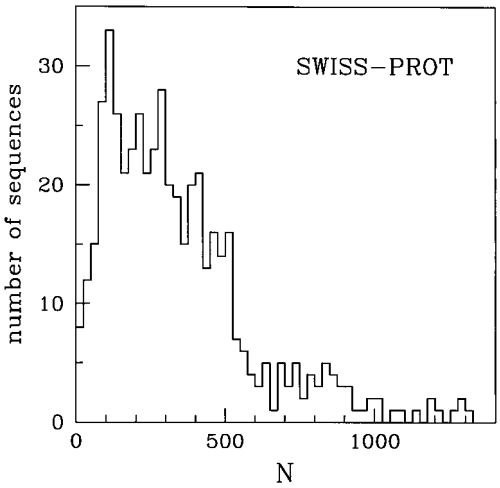

Figure 1. The distribution of polymer lengths among 500 proteins picked randomly from entries of the SWISS-PROT Protein Sequence Database [1].

their scaling behavior has proved to offer clues about the size related limitations of the functionality of proteins [5]. The first of these is known as the folding temperature. It relates to the thermodynamic stability and it may be defined as a temperature at which the probability to occupy the native state crosses $\frac{1}{2}$ [6]. The second temperature corresponds to conditions that make folding optimal [6–8] and it thus relates only to the kinetics. The significance of T_{min} is that on lowering T below T_{min}, the glassy effects become more and more pronounced and at a glass temperature, T_g, t_{fold} starts to exceed a preassigned large threshold value [6]. T_g is defined through this threshold value which necessarily depends on N. Thus T_g is not sufficiently well suited for studies of scaling.

The relevance of T_f and T_{min} taken together, is that they define a kinetic criterion according to which a sequence may act as a good folder: T_f should be greater than T_{min}, which often holds for small N's. Our studies of lattice heteropolymers [5] suggest emergence of scaling to asymptotically bad folding conditions in which the glassy effects prohibit a substantial occupation of the native state. These conditions correspond to T_f being much less than T_{min}. It follows that there is a characteristic value of N, N_c, that marks an onset of a systematic worsening in the physiological functionality of proteins. Our lattice based estimate (with certain Monte Carlo dynamics) was an N_c of order 300 and it was obtained by locating a point at which a monotonically increasing T_{min} was intersecting with an initially growing and then saturating dependenee of T_f on N.

Existence of the saturation effect in T_f has been illustrated by Takada and Wolynes [9] in their droplet approximation of protein folding. The indefinite growth in T_{min} with N seems to be related to the necessity of rearranging larger and larger segments to secure the optimal folding conditions. Such a growth, however, is not

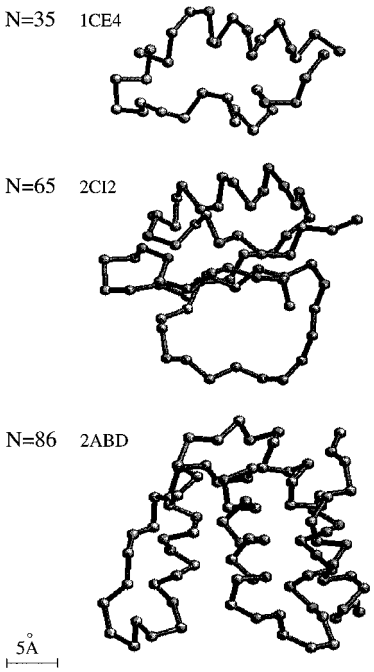

Figure 2. Native conformations of three proteins of the set studied in this paper. The Protein Data Bank [13] code name and the numbers of monomers in a sequence are indicated.

a universal feature. It is not observed in the Ising spin system analogs of proteins [10] where search for a ground state involves flipping spins. It is also not observed in two-dimensional lattice [10] and three-dimensional off-lattice models of helices [11] which suggests absence of packing hindrance in the helices. Notice that the monomer-monomer interactions in helices are only of a local kind (as counted along the sequence) and thus no natural motion of the system may induce an overall restructuring.

Essentially all information about the scaling of folding properties of proteins comes from the studies of lattice models [3, 5] with, necessarily, declared dynamics. In this paper, we turn to continuum space models of proteins and focus specifically on the Go models [12] in which only the native interactions are assumed to play a primary role in the dynamics of folding. We consider Hamiltonians, described in details in Section 3, in which these interactions are of the Lennard-Jones form and the time evolution is obtained by solving Newton's equations with a Langevin noise corresponding to a given temperature. For a comparison, some discussion will refer also to the lattice models in which case the interactions reside in contacts that are formed by nonconsecutive beads one lattice constant apart and the dynamics will consist of Monte Carlo single- and double monomer moves. The input to this modelling is the native conformation in which each aminoacid is coarse-grained

and becomes a bead located at its α-carbon position. Some of such conformations are shown in Figure 2.

We have calculated t_{fold}, T_f, and T_{min} for 21 single domain Go-modelled Protein Data Bank structures [13] with N ranging between 29 and 98. A short account of a part of the results on the scaling of t_{fold} has been given in Reference [14]. Nine of the selected structures belong to a set of proteins considered by Plaxco *et al.* [15] or are their close homologies. These are: the SH3 domain of 1efn (57), 2ptl (63), 2ci2 ($83 - 18 = 65$; 18 are not resolved), 1csp (67), 1ubq (76), 1hdn (85), 2abd (86), 1ten (90), and 1aps (98), where the numbers in brackets indicate the corresponding values of N. The additional 12 structures are: 1cti (29), 1cmr (31), 1ce4 (35), 1bba (36), 1erc (40), 1crn (46), 7rxn (52), 5pti (58), 1tap (60), 1aho (64), 1ptx (64), 1erg (70). All of these native conformations were picked from the low-N end of the size distribution to allow for a thorough equilibrium and kinetic characterization. Our studies of these structures indicate well defined overall trends in t_{fold} which we shall discuss in Section 4. We demonstrate that inclusion of additional steric constraints in the Hamiltonian [11] does not alter the trends in any significant manner. We shall use the notation:

PDB – for models of proteins without the steric constraints,

PDBS – for models of proteins with the steric constraints.

An interpretation of these trends, however, requires making comparison to properties calculated for various classes of decoy conformations defined in continuum space. These classes form statistical ensembles in which a given value of N has multiple realizations of the native conformation within a similar class of geometry. Specifically, we consider 5 classes of decoys and all of them are studied only without an implementation of the steric constraints since this is the case which involves much less computer time. We describe these structures in Section 2.

The classes of decoys studied here differ in the way they fill space and in their packing arrangement. Our main finding is that

$$t_{fold} \sim N^{\lambda}, \tag{1}$$

and that exponent λ depends on the class of the structures. The values of λ, when calculated at T_{min}, are listed in Table I. In 3 dimensions they range between 1.7 and 3.2. For the PDB and PDBS structures the exponent is about 2.5 ± 0.2 and 2.7 ± 0.1 respectively which indicaties that the proteins are not optimized kinetically [14] even though they might be optimized functionally or geometrically [16]. A sensitivity of λ to the class of decoys is consistent with the overall lack of precise universality found in the lattice models [5, 3].

In Section 5 we focus on the scaling of the characteristic temperatures and demonstrate the general deterioration-of-folding scenario in the classes of the decoy conformations sets in already at quite small values of N. The largest value of N_c based on a rough estimate corresponds to the helical conformations which

Table I. The exponents λ for the classes of conformations studied. Two values of λ for HB correspond to the two different slopes shown in Figure 10

Structure	λ
HC	2.2 ± 0.1
HA	1.7 ± 0.1
HB	$0.9 \pm 0.1, 3.2 \pm 0.1$
HQ	2.7 ± 0.2
CL	2.6 ± 0.2
CL'	2.1 ± 0.1
PDB	2.5 ± 0.2
PDBS	2.7 ± 0.1

indicates that they have the best foldability among the classes of conformations considered. A borderline behavior is found for the PDB. This seems to indicate a marginal character of behavior that is present already at small N with N_c even of order 40. This might reflect on the inadequacies of the Lennard-Jones based Go model but another possibility is that the borderline behavior is, in fact, real.

2. Classes of the decoy conformations

The classes of decoy conformations that we consider here are as follows.

CL: compact native conformations generated on a grid as a self-avoiding random walk within a compact box of lattice constant 3.8 Å (the length of a peptide bond) and then stabilized by appropriate Lennard-Jones interactions. Three dimensional ($3D$) examples of such conformations are shown in Figure 3. The analogous $2D$ conformations will be denoted by CL'.

HC: conformations obtained through slow homopolymer cooling. The procedure involves generating a self-avoiding random walk, assigning identical strengths to all inter-bead interactions, and then annealing to a compact conformation. Figure 4 shows a sample of such a conformation.

HQ: similar to HC but with a rapid quenching instead of annealing. Examples of the resulting HQ conformations are shown in Figure 5. These structures are significantly spread out in space.

HA: similar to HC but the α-helices of various lengths are first built into the initial states and then kept through the annealing process by assigning much stronger couplings to the helical secondary structures. Examples of these conformations are shown in Figure 6.

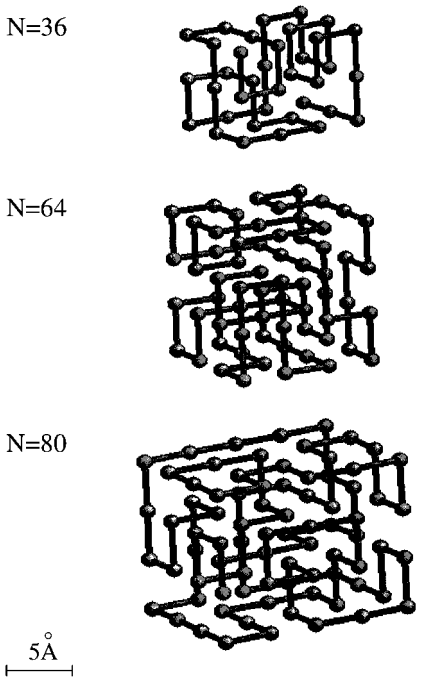

Figure 3. Examples of decoy native conformations generated through the compact walks (CL) procedure. The conformations with $N = 36$, 64, and 80 are constructed on $3 \times 3 \times 4$, $4 \times 4 \times 4$ and $4 \times 4 \times 5$ grids respectively.

HB: similar to HA but the helical segments are replaced by β-sheet conformations, as shown in Figure 7. The number of beads in each strand is fixed to be equal to 8.

We have considered 11 realizations of conformations CL, HQ, HC, and HB, and 5 realizations of HA for each N.

3. The model

Scaling studies of any complicated system have a chance of success if the model under consideration becomes sufficiently simplified. The smaller the number of parameters to vary, the smaller the number of statistical ensemble of representative systems that are needed to identify any scaling trends. The Go heteropolymers [12] appear to be the desired minimal coarse grained models of proteins since much of their properties are defined just by the shape of the native conformation – with disregard to any unsettled details of the true interactions between the aminoacids. Despite these simplifications, the Go models may actually behave in a more realistic way than the models that are more correct atomically [17].

There are several variants of the Go models that are set in the continuum space. For instance, Zhou and Karplus [18] and Dokholyan *et al.* [19] have considered

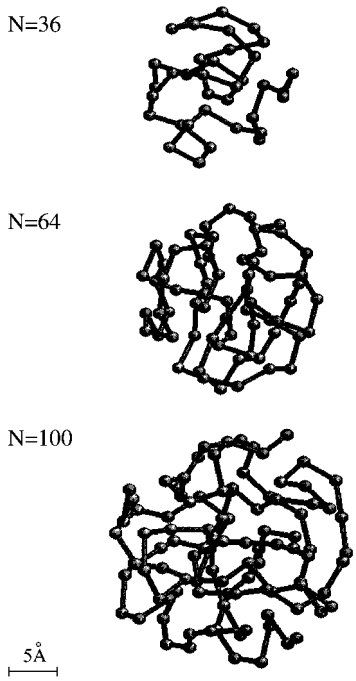

Figure 4. Examples of decoy native conformations generated through the homopolymer cooling (HC) procedure.

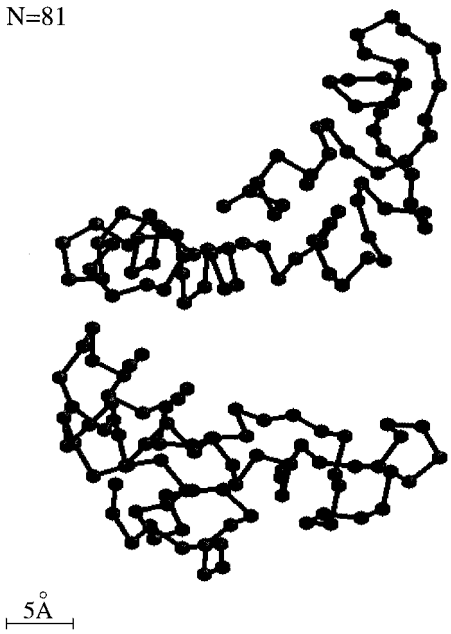

Figure 5. Examples of decoy native conformations generated through the homopolymer quenching for $N = 81$. Note the lack of any compactness.

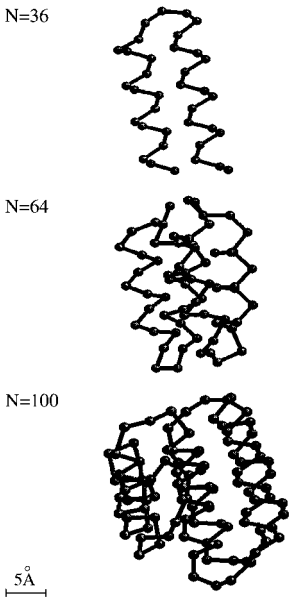

Figure 6. Examples of decoy native conformations generated through the homopolymer cooling with α-helices (HA) procedure. The top conformation is a two-helix bundle with 17 monomers in each helical branch. The middle conformation corresponds to a four-helix bundle where the helices are of 15 monomer's length. In the bottom conformation there are 6 helices where the lengths are 15, 14, 15, 15, 15, and 15.

models with a square well potential which allow for a simplified discretized time evolution. Wolynes *et al.* [20] have implemented an associative memory Hamiltonian in which the contact potentials are assumed to be the Gaussian functions. Clementi *et al.* [21], on the other hand, have studied the 12–10 power law potentials. Here, we focus on the Go-like approach based on the Lennard-Jones potentials [22, 23]. A thorough presentation of the methodology and of a list of results on folding of secondary structures have been given in our previous studies [11]. An outline of this is as follows.

The target native conformation is represented by beads on a chain. If this is not a decoy conformation, the coordinates of the beads are taken as positions of the α-carbons from the Protein Data Bank. The potential energy of the system has the following form:

$$E_p(\{\mathbf{r}_i\}) = V^{BB} + V^{NAT} + V^{NON} + V^{ST}. \tag{2}$$

The first term represents rigidity of the backbone potential, the second term corresponds to interactions in the native contacts and the third term to those in the non-native contacts. Finally, the last term corresponds to the steric constraints if these are taken into account. Two monomers are assumed to be in a native contact if their distance in the native conformation is smaller than some value d_{nat}. For PDB, HC, HA and HB we use $d_{nat} = 7.5$ Å, whereas for CL $d_{nat} = 6$ Å. The

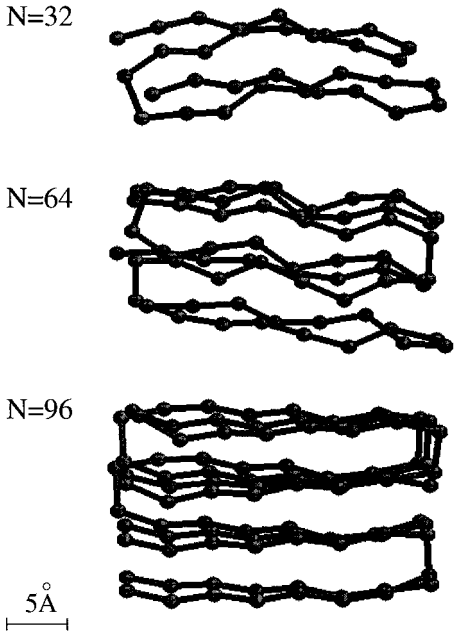

Figure 7. Examples of decoy native conformations generated through the homopolymer cooling with β-sheets (HB) procedure. The lengths of the β-strands are fixed to be of 8 monomers. The numbers of strands in the conformations (from top to bottom) are 4, 8 and 12 respectively.

smaller d_{nat} used for the structures CL is due to the fact that they are much more compact than the structures in the remaining classes.

The backbone potential takes the form of the sum over harmonic [24] and anharmonic [25] interactions

$$V^{BB} = \sum_{i=1}^{N-1} [k_1(r_{i,i+1} - d_0)^2 + k_2(r_{i,i+1} - d_0)^4], \tag{3}$$

where $r_{i,i+1} = |\mathbf{r}_i - \mathbf{r}_{i+1}|$ is the distance between two consecutive beads; $d_0 = 3.8$ Å, $k_1 = \epsilon$ and $k_2 = 100\epsilon$, where ϵ is the Lennard-Jones energy parameter corresponding to a native contact.

The interaction between residues that form a native contact in the target conformation is taken to be of the Lennard-Jones form:

$$V^{NAT} = \sum_{i<j}^{NAT} 4\epsilon \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right], \tag{4}$$

where the sum is over all pairs of residues i and j (but those which are immediate neighbors along the chain) which form the native contacts in the given target conformation. $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ is the monomer to monomer distance. The parameters σ_{ij} are chosen in a way that each contact in the native conformation is stabilized

at the minimum of the potential. Essentially, $\sigma_{ij} = 2^{-1/6} \cdot d_{ij}$, where d_{ij} is the corresponding native contact's length.

Residues that do not form the native contacts interact via a repulsive soft core potential. Our potential for non-native contacts, given below, differs from the model of Iori *et al.* in that it falls to 0 after some cut-off distance, d_{cut}, which improves foldability. Restricting the number of interactions has been shown to reduce frustration and lead to an improvement of the folding properties [22]. The purpose of introducing the cut-off distance is to make sure that the target conformation is in fact the ground state of the system.

$$V^{NON} = \sum_{i<j}^{NON} V_{ij}^{NON}, \tag{5}$$

$$V_{ij}^{NON} = \begin{cases} 4\epsilon \left[\left(\frac{\sigma_0}{r_{ij}} \right)^{12} - \left(\frac{\sigma_0}{r_{ij}} \right)^{6} \right] + \epsilon & , \; r_{ij} < d_{cut} \\ 0 & , \; r_{ij} \geq d_{cut}. \end{cases} \tag{6}$$

Here, $\sigma_0 = 2^{-1/6} \cdot d_{cut}$. For distances shorter than d_{cut} the potential is purely repulsive. We chose $d_{cut} = \langle d_{ij} \rangle$ which is a mean value of the lengths of the contacts.

The steric constraints can be represented by

$$V^{ST} = V^{BA} + V^{DA}, \tag{7}$$

where the terms correspond to the bond angle and dihedral angle potentials respectively.

Following Reference [26], we use the following potentials for the bond and the dihedral angles

$$V^{BA} = \sum_{i=1}^{N-2} \frac{k_\theta}{2} (\theta_i - \theta_{0i})^2 \tag{8}$$

$$V^{DA} = \sum_{i=1}^{N-3} [A(1 + \cos \phi_i) + B(1 + \cos 3\phi_i)], \tag{9}$$

where $k_\theta = 20\epsilon/(rad)^2$, $A = 0\epsilon$ and $B = 0.2\epsilon$. Our angle dependent potentials differ from those used in Reference [26], since in our case we take θ_{0i} to be, in general, site-dependent and equal to the bond angles that appear in the native targets. Introduction of the steric constraints to the model described by Equation (1) shifts the native state away from the target conformation because the target need not correspond to a minimum of the dihedral potentials. However, we have found that for our choice of the parameters A and B the shift is insignificant. The true native states are found by a multiple zero-temperature quench procedure from low energy

conformations generated by MD trajectories that start in the target conformation. The conformational distances as defined in Reference [11] from the native states to the targets never exceeded 0.1 Å.

The dynamics of the system are captured by the Langevin equation

$$m\ddot{\mathbf{r}} = -\gamma\dot{\mathbf{r}} + F_c + \Gamma, \tag{10}$$

where r is a generalized coordinate of a bead, m is the monomer's mass, $F_c = -\nabla_r E_p$ is the conformation force, γ is a friction coefficient and Γ is the random force which is introduced to balance the energy dissipation caused by friction. Both the friction and the random force represent the effects of the solvent and they control the temperature [27]. Γ is assumed to be drawn from the Gaussian distribution with the standard variance related to temperature by

$$\langle\Gamma(0)\Gamma(t)\rangle = 2\gamma k_B T\delta(t), \tag{11}$$

where k_B is the Boltzmann constant, T is temperature, t is time and $\delta(t)$ is the Dirac delta function.

The Langevin equations are integrated using the fifth order predictor-corrector scheme [28]. The friction and random force terms are included in the form of a noise perturbing the Newtonian motion at each integration step. In the case of the model with the steric constraints, the forces associated with the angle-dependent potentials are calculated through a numerical determination of the derivatives of the potential.

In the following, the temperature is measured in the reduced units of ϵ/k_B. The integration time step is taken to be $\Delta t = 0.005\tau$, where τ is a characteristic time unit. At low values of friction, τ coincides with the period of oscillations, τ_m near the Lennard-Jones minimum and is equal to $\sqrt{ma^2/\epsilon}$, where a is a Van der Waals radius of the amino acid residues. The value of a is chosen to be equal to 5 Å, and this value is roughly equal to $\langle\sigma_{ij}\rangle$. Our simulations are performed with $\gamma = 2m\tau^{-1}$ which is a standard choice in molecular dynamics studies of liquids. Going into still higher values of γ, however, has been argued to be more realistic [26]. In a previous study [11] we have checked that changing γ leads to a change in the scale of the folding times but does not effect the value of T_{min}.

When studying the homopolymer coolings we chose the Lennard-Jones σ of 5 Å which is a standard value of the Van der Waals radius of the amino acid residues. Once a chain is generated, it is first warmed up to $T = 5$ and then slowly cooled, in 50 temperature steps, to $T = O$. The duration of a run at each T varies from 20τ to 50τ depending on the system size. The Lennard-Jones couplings are assigned to all pairs of beads and, in the process of the target formation, there is no cut-off in the interactions. When generating conformations with the secondary structures, HA and HB, we increase the amplitude of the Lennard-Jones potentials by about 10 times in those contacts, which correspond to the hydrogen bonds within the α-helices and the β-sheets. This makes these contacts stable

during annealing. Once a conformation is constructed, we switch over to the Go Hamiltonian.

4. The folding times

The bigger the N, the longer time is needed to find the native state when starting from a random open conformation. Using arguments from polymer theory Thirumalai [2] (see also [29]) has argued that a power law scaling for t_{fold} should characterize proteins which fold through direct pathways with a nucleation mechanism. The exponent λ for the two-state folders has been estimated to be between 3.8 and 4.2, and the folding time has been also proposed to depend linearly on the viscosity of the solvent. For indirect pathways, the folding time is determined primarily by activation process with barriers which were argued to scale as $N^{1/2}$. There have been a number of other theories for the dependence of t_{fold} on N that were based on arguments regarding some power law dependence of the barrier heights on N [9, 30, 31] and thus would lead to an exponential law for t_{fold}.

A numerical evidence for the power law, was first provided by Gutin, Abkevich and Shakhnovich [3] who studied three dimensional ($3D$) lattice sequences with N up to 175. For each N, they considered 5 sequences and selected one that folded the fastest under its optimal conditions. The corresponding folding time, t_{01}, was the quantity that was used in studies of scaling. They discovered that t_{01} grows as a power law with the system size and the corresponding exponent λ depended on the kind of distribution of the contact energies B_{ij} in the Hamiltonian

$$H = \sum_{i<j} B_{ij}\Delta_{ij}, \tag{12}$$

which pointed to the existence of a variety of kinds of the energy landscapes [32]. In Equation (2), Δ_{ij} is either 1 or 0 depending on whether the monomers i and j make a contact or not. For random and designed sequences, with the B_{ij}'s generated from the data base of Reference [33], $\lambda \approx 6$ and ≈ 4, respectively [3]. Finally, for the Go model, in which $B_{ij} = -1$ for native contacts and 0 for non-native contacts, $\lambda \approx 2.7$. The power law scaling has been confirmed by our study [5] which was dedicated to the two- and three-D lattice Go models and involved larger statistics.

The lattice models studies suggest that the power law dependence of t_{fold} on N is likely to be found also in continuum space heteropolymers and this is indeed what we find. Since the precise values of the exponents have been found to depend on temperature [5], we start by identifying T_{min} for each sequence under study. It should be noted that on increasing the system size the temperature window of kinetic optimality becomes narrower and narrower – both for the lattice and off-lattice models. This is illustrated in Figure 8 which compares the T-dependenee of t_{fold} for two values of N. The lattice models come with the Monte Carlo dynamics which consists of the single and two-monomer moves whereas the off-lattice mod-

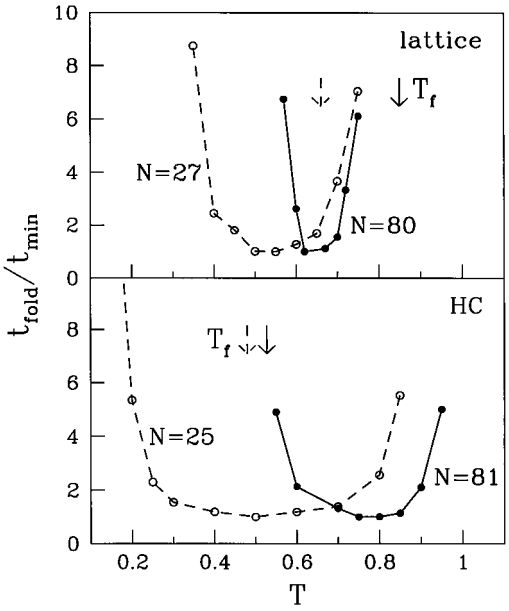

Figure 8. The temperature dependence of the median folding time for selected individual Go sequences in the lattice model with $N = 27$ and 80 (top panel) and for the HC sequences in the off-lattice model with $N = 25$ and 81 (bottom panel). The median has been determined based on 200 starting conformations. The arrows indicate the corresponding values of T_f and, together with T_{min}, illustrate the worsening of the foldability with N.

els evolve according to Newton's laws. The folding times shown in Figure 9 are divided there by

$$t_{min} = t_{fold}(T_{min}), \tag{13}$$

i.e. by the value of t_{fold} at the bottom of the overall U-shaped dependence on T. This allows one to focus on the width of the U as a function of N and, furthermore, enables to make a more meaningful comparison of the Monte Carlo and molecular dynamics time scales.

An issue that arises when studying folding in continuum space systems is what are the criteria to determine whether a particular conformation may be considered as being sufficiently close to the native state to declare folding. In this paper, we declare the system to be in its native state if all of its native contacts are established. A native contact is said to be established if the distance between the two monomers, say, i-th and j-th, is shorter than $1.5\sigma_{ij}$, where σ_{ij} is the corresponding Lennard-Jones length parameter. The use of such a criterion has been tested to provide a correct geometry of the native conformation.

Figures 10 and 11 show the scaling of the folding time at T_{min} for the off-lattice classes of the decoy conformations: HC, HA, HB, HQ, CL and CL'. The power law is observed in all cases, but it becomes of a crossover type in the case of HB.

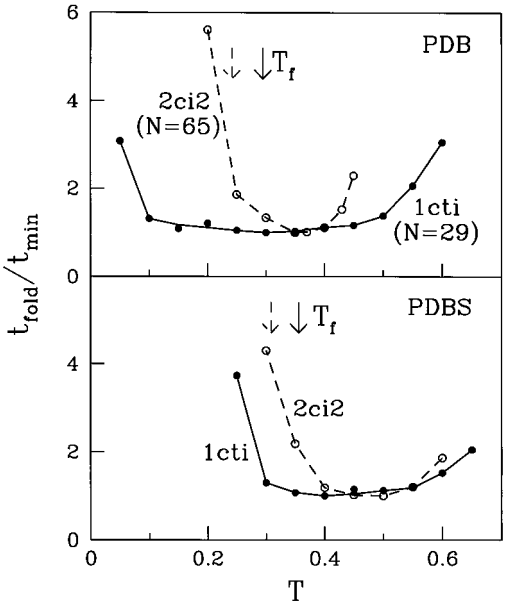

Figure 9. The temperature dependenec of the median folding time of a 29-residue trypsin inhibitor (1cti) and a 65-residue chymotrypsin inhibitor 2 (2ci2) for the Go model without the steric constraints (top panel) and with the steric constraints (bottom panel). The median has been determined based on 200 folding trajectories for each temperature. The arrows indicate the corresponding values of T_f and T_{min}.

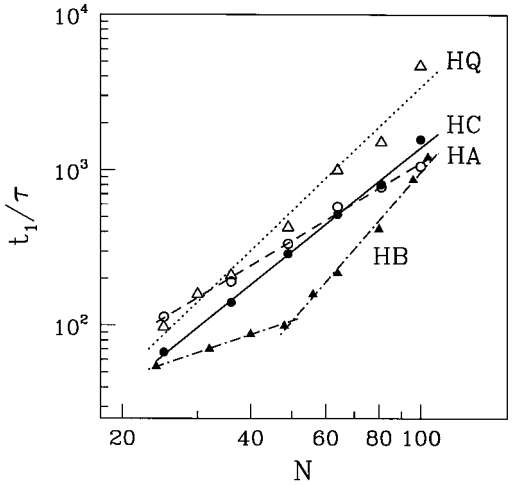

Figure 10. Power law dependence of the median folding times at T_{min} on N. 11 sequences have been considered for each N and T_{min} was determined separately for each sequence. t_1 is the the median value of the corresponding t_{min}'s. The corresponding exponents are shown in Table I. The points for HB are fitted by two different slopes for $N < 64$ and $N \geq 64$.

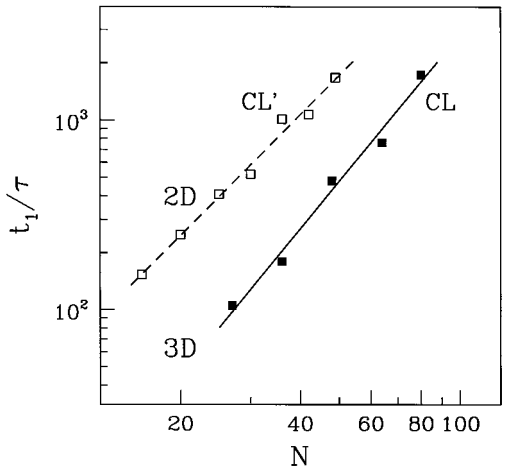

Figure 11. Power law dependence of the median folding times at T_{min} on N for the maximally compact targets on 3 and $2D$ lattice. Notice a weak dependence on dimensionality.

The exponents are summarized in Table I. With the exception of HB, the exponents vary from 1.7 to 2.7. Notice that they are all smaller than the lattice model value of around 3 (the value of 3 has been also derived in a heuristic lattice model proposed by Camacho [34]). The most lattice-like conformations CL are characterized by a λ which is also the closest to the lattice estimates.

The crossover behavior found for HB may have to do with the fact that all of our β-sheet conformations have a fixed strand's length of 8 beads and thus too short chains might not yet be in the proper scaling β-sheet regime. In fact, the short β-sheet conformations have a much lower compactness level than the long chains. Notice that an earlier study by Zhdanov [4] has reported a power law scaling for the β-sheets within an HP-like model on the lattice with N not exceeding 40.

An interesting issue is that for the continuum space models the exponent seems to be only weekly sensitive to the dimensionality of the target conformation. This is illustrated in Figure 11. The Go model for the maximally compact $2D$ conformations, CL', yields the exponent λ which is close to that of CL provided the dynamics are still defined in the $3D$ space. The lattice Monte Carlo simulations [5], on the other hand, show a strong dependence on D.

Note, however, that the Monte Carlo dynamics are only a poor approximation of the time evolution of the Newtonian dynamics. In order to illustrate the arbitrary nature of the standard Monte Carlo dynamics [6], we consider another model of the Monte Carlo dynamics in which only the single monomer moves are allowed, i.e. in which the double monomer moves are prohibited [35]. Figure 12 shows that the change in the allowed moves of the Monte Carlo dynamics results in a substantial change in the value of λ. There is thus no reason for which the continuum space exponent should agree with any of the lattice based estimates.

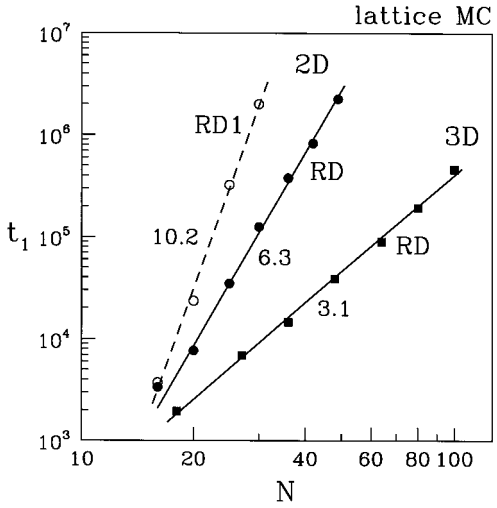

Figure 12. The power law dependence of the median folding times on N at T_{min} in the Go lattice systems. The exponents are 3.1 ± 0.1 in $2D$ and 6.3 ± 0.2 in $3D$ for RD and 10.2 ± 0.5 in $2D$ for RD1.

As shown in Table I, the exponent λ sensitively depends on the geometrical class of the native conformations. The smallest exponents correspond to the HA conformations whereas the largest – to the HQ and long HB conformations. Conformations generated by homopolymer annealing (HC) yield an intermediate value of λ. We conclude that the native geometry plays an important role in determining the speed of folding. Native conformations HA and short HB that were obtained through annealing with the built in secondary structures, have the scaling exponent which is substantially smaller than the one obtained for the more random conformations – CL, HC and HQ. HA and short HB are more optimal kinetically than CL, HC and HQ.

The next issue is how does the scaling behavior of proteins, as modelled by the PDB and PDBS sequences, relate to that of the decoys. Figure 12 shows that the folding time at T_{min} scales in an essentially same fashion for both kinds of sequences. There is, however, one PDB sequence (1aps), for which the folding time jumps away from the typical trend. This sequence probably suffers from a large topological frustration and it appears not to be a good folder. Experiments also have reported that 1aps has a very low folding rate [15]. For the rest of PDB's we found the exponent λ to be about 2.5. For PDBS (with fewer sequences) the exponent is about 2.7, which agrees with that for PDB within the error bars which indicates small role of the steric constraints in kinetics. The exponents λ for PDB i PDBS are found to be comparable to that of CL and HQ. They are noticeably larger than that obtained for the other decoys, HC, HA and short HB, even if one takes into account the larger error bars in the PDB data (for each N we consider just one PDB conformations). Thus, the realistic protein structures taken from PDB

appear to be less optimized than those which are artificially generated such as HC, HA and short HB. Geometrical frustration in PDB might be due to various kind of loops connecting the secondary elements. The loops, for instance, are optimized in HA. Note that λ for HA is smaller than 2 – the scaling exponent suggested by de Gennes [36] in his analysis of the time scale for the coil to globule transition of a homopolymer. This indicates the optimality of HA and is consistent with the result [37] that alpha-helices are optimal folders. The full proteins, however, appear not to be optimal.

The HQ structures a priori might seem to be models of random sequences of aminoacids and thus be bad folders as suggested by their spread out shapes. However, the HQ structures involve primarily local contacts which leads, overall, to a PDB-like scaling exponent.

The grid CL conformations do not look similar to the PDB ones and yet their exponents λ are almost identical. The CL conformations have a low level of kinetic optimization of the geometry relative to HA and short HB.

5. Scaling properties of T_F and T_{min}

We now discuss the scaling of characteristic temperatures. T_{min} is determined from the kinetic data as discussed in the previous section. Instead of characterizing the kinetics by T_{min}, one might, following Socci and Onuchic [6], consider the better known glass transition temperature, T_g, at which t_{fold} starts exceeding a certain threshold value. However, there are two problems with any studies of T_g. The first one has been already mentioned: the threshold value of t_{fold} must itself depend on N. The second problem is that the very notion of T_g presupposes existence of an equilibrium glassy phase in a finite system. It should be noted that an equilibrium glassy phase of heteropolymers arises in mean field theories [32, 38] but it is not expected to arise in systems with a finite N [7].

Even though T_{min} is conceptually well defined, when it comes to scaling, this quantity is not without its own problems. One can see from Figures 8 and 9 that the U-shaped dependence of t_{fold} vs. T is usually broad at the bottom, especially for low values of N. Thus temperatures in a broader neighborhood of T_{min} are essentially still optimal for folding and a more sharply defined characteristic temperature would be of more relevance. In order to eliminate the adverse aspects of the definitions of both T_g and T_{min}, we introduce T_{g2}: this a temperature below T_{min} at which t_{fold} becomes twice as large than t_1. Thus T_{g2} delineates the low T bound of the temperatures that are optimal kinetically and it is already in the region of a well identifiable raise in t_{fold}. In the following, we show the scaling properties of T_{min}, T_{g2}, and T_f.

The folding temperature is calculated by performing long runs that determine the equilibrium probability of the system staying in the native state. The probabilities are determined as a function of T and T_f is obtained by an interpolation to where the value of $1/2$ is crossed. The system is said to be in the native state if

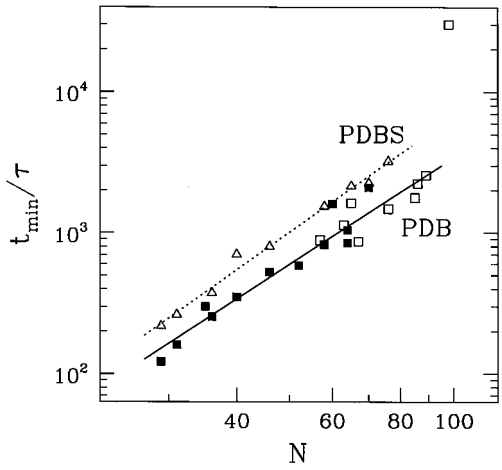

Figure 13. The scaling of the folding time at T_{min} for the PDB and PDBS conformations.

all the native contacts are present. For each T we consider 10 to 15 trajectories that start from the native state. The lengths of the trajectories vary from 15000 to 30000τ depending on the system size. The first 1000 to 2000τ are spent for an equilibration of the chain and are not taken into account when counting the probabilities.

We start our discussion of the dependence of T_f, T_{min} and T_{g2} on N in continuum systems by considering the decoy conformations: CL, HC and HA. Figure 13 shows that for all of these cases T_{min} and also T_{g2} grows with the chain length whereas T_f rapidly achieves saturation. The qualitative picture is then similar to the lattice results [5]. It confirms our previous suggestion that the folding properties of proteins deteriorate with N due to a larger and larger difference between T_{min} and T_f. For all cases, except for HC at small N, we observe that T_f is smaller than T_{min}, which suggests that the models are not very good folders. However, as the border between good and bad folders is rather broad one may define N_c – the system size beyond which the sequences are no longer the good folders – as corresponds to the crossing point where T_{g2} starts to become bigger than T_f. Using this definition we estimate that N_c is roughly equal to equal to 36, 70 and 90 for CL, HC and HA respectively. The largest N_c corresponds to the HC conformations which indicates that this kind of structures has the best foldability. The poorest foldability belongs to CL the conformations that are least similar to those of proteins. Again, the impact of the geometry of the native conformation on the folding properties is evident in our study.

Figure 14 shows the values of T_{min} and T_f for the selected sequences PDB and PDBS. For PDB the values of T_{g2} are also shown. For most cases one observe that T_f is somewhat smaller than T_{min} whereas T_f and T_{g2} show a borderline behavior. Since the fluctuations in the temperature data for the PDB structures are large,

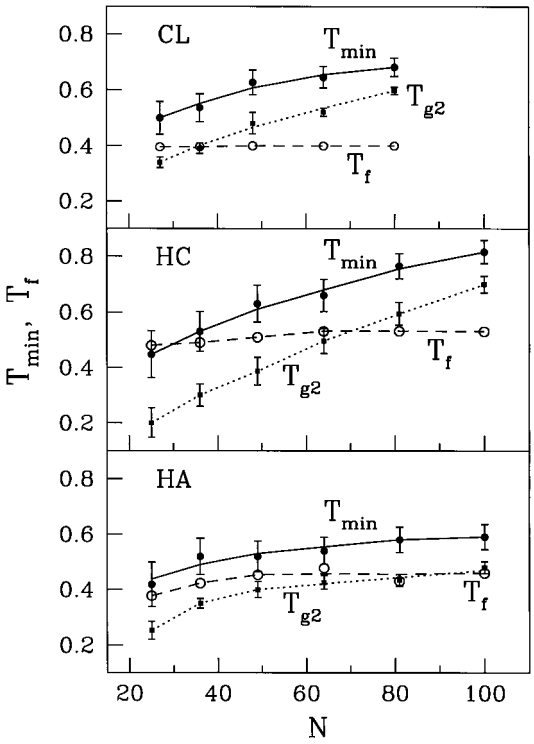

Figure 14. Scaling of T_f, T_{min} and T_{g2} for the decoy conformations CL, HC, and HA as indicated.

establishing trends, without larger statistics, is difficult. We propose an interesting hypothesis that the marginal behavior we observe in our model systems may actually characterize classes of proteins, especially those which have a short lifetime in a living cell. It would be interesting to extend these studies with the use of more realistic potentials.

Acknowledgements

This work was supported by KBN (Grant No. 2P03B-025-13). Discussions with J.R. Banavar, H. Nymeyer and S. Plotkin are appreciated.

References

1. Bairoch, A. and Apweiler, R.: The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, *Nucleic Acids Res.* **28** (2000), 45–48.
2. Thirumalai, D.: From minimal models to real proteins: timescales for protein folding, *J. Phys. I (France)* **5** (1995), 1457–1467.

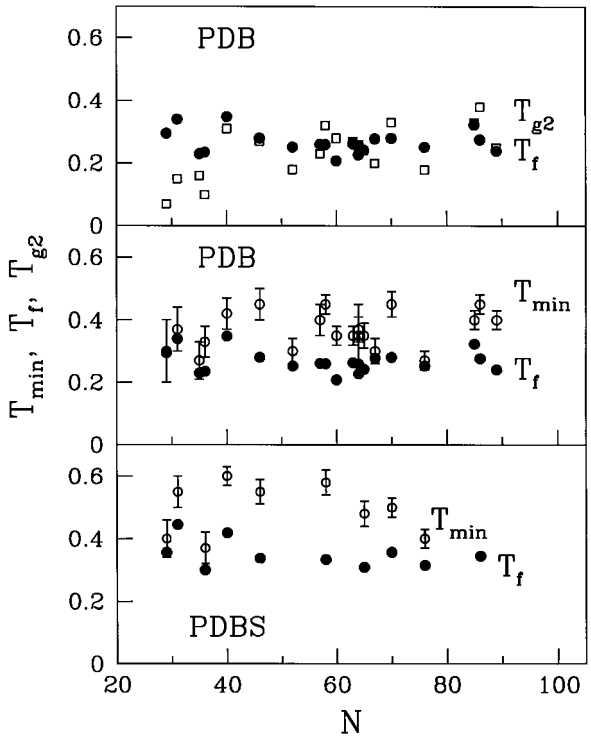

Figure 15. Scaling of T_f, T_{min} and T_{g2} for the PDB sequences (top and middle panels) and of T_f and T_{min} for the PDBS sequences (bottom panel).

3. Gutin, A.M., Abkevich, V.I. and Shakhnovich, E.I.: Chain length scaling of protein folding time, *Phys. Rev. Lett.* **77** (1996), 5433–5436.
4. Zhdanov, V.P.: Folding time of ideal β sheets vs. chain length, *Europhys. Lett.* **42** (1998), 577–581.
5. Cieplak, M., Hoang, T.X. and Li, M.S.: Scaling of folding properties in simple models of proteins, *Phys. Rev. Lett.* **83** (1999), 1684–1687.
6. Socci, N.D. and Onuchie, J.N.: Folding kinetics of protein-like heteropolymers, *J. Chem. Phys.* **101** (1994), 1519–1528; see also Cieplak, M. and Banavar, J.R.: Cell dynamics of folding in two dimensional model proteins, *Fold. Des.* **2** (1997), 235–245.
7. Cieplak, M., Henkel, M., Karbowski, J. and Banavar, J.R.: Master equation approach to protein folding and kinetic traps, *Phys. Rev. Lett.* **80** (1998), 3654–3657.
8. Cieplak, M., Henkel, M. and Banavar, J.R.: Master equation approach to protein folding, *J. Cond. Mat.* **2** (1999), 369–378.
9. Takada, S. and Wolynes, P.G.: Microscopic theory of critical folding nuelei and reconfiguration activation barriers in folding proteins, *J. Chem. Phys.* **107** (1997), 9585–9598.
10. Hoang, T.X., Sushko, N., Li, M.S. and Cieplak, M.: Spin analogs of proteins: scaling of 'folding' properties, *J. Phys. A* **33** (2000), 3977–3988.
11. Hoang, T.X. and Cieplak, M.: Molecular dynamics of folding of secondary structures in Go-like models of proteins, *J. Chem. Phys.* **15** (2000), 6851–6862; Hoang, T.X. and Cieplak, M.: Sequencing of folding events in Go-like proteins, *J. Chem. Phys.*, **113** (2000), 8319–8328.

12. Abe, H. and Go, N.: Noninteracting local-structure model of folding and unfolding transition in globular proteins. II. Application to two-dimensional lattice proteins, *Biopolymers* **20** (1981), 1013–1031.
13. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M.: The Protein Data Bank: a computer-based archival file for macromolecular structures, *J. Mol. Biol.* **112** (1977) 535–542.
14. Cieplak, M. and Hoang, T.X.: Kinetic non-optimality and vibrational stability of proteins, submitted for publication.
15. Plaxco, K.W., Simons, K.T. and Baker, D.: Contact order, transition state placement and the refolding rates of single domain proteins, *J. Mol. Biol.* **277** (1998), 985–994.
16. Micheletti, C., Banavar, J.R., Maritan, A. and Seno, F.: Protein Structures and optimal folding from a geometrical variational principle, *Phys. Rev. Lett.* **82** (1999), 3372–3375.
17. Takada, S.: Go-ing for the prediction of protein folding mechanism, *Proc. Natl. Acad. Sci.* **96** (1999), 11698–11700.
18. Zhou, Y. and Karplus, M.: Interpreting the folding kinetics of helical proteins, *Nature* **401** (1999), 400–403.
19. Dokholyan, N.V., Buldyrev, S.V., Stanley, H.E., Shakhnovich, E.I.: Discrete molecular dynamics studies of the folding of a protein-like model, *Folding Des.* **3** (1998), 577–587; Dokholyan, N.V., Buldyrev, S.V., Stanley, H.E., Shakhnovich, E.I.: Identifying the protein folding nucleus using molecular dynamies, cond-mat 9812284 (1998).
20. Hardin, C., Luthey-Schulten, Z. and Wolynes, P.G.: Backbone dynamics, fast folding, and secondary structure formation in helical proteins and peptides, *Proteins: Struct. Funct. Genet.* **34** (1999), 281–294.
21. Clementi, C., Nymeyer, H. and Onuchic, J.N.: Topological and energetic factors: what determine the structural details of the transition state ensemble and 'n-route' intermediates for protein folding? An investigation for small globular proteins, *J. Mol. Biol.* **298** (2000), 937–953.
22. Li, M.S. and Cieplak, M.: Folding in two-dimensional off-lattice models of proteins, *Phys. Rev. E* **59** (1999), 970–976.
23. Klimov, D.K. and Thirumalai, D.: Mechanisms and kinetics of β-hairpin formation, *Proc. Natl. Acad. Sci. USA* **97** (2000), 2544–2549.
24. Iori, G., Marinari, E. and Parisi, G.: Random self-interacting chains: a mechanism for protein folding, *J. Phys. A* **24** (1991), 5349–3562.
25. Clementi, C., Maritan, A. and Banavar, J.R.: Folding, design and determination of interaction potentials using off-lattice dynamics of model heteropolymers, *Phys. Rev. Lett.* **81** (1998), 3287–3290.
26. Veitshans, T., Klimov, D., Thirumalai, D.: Protein folding kinetics: time scales, pathways and energy landscapes in terms of sequence-dependent properties, *Folding Des.* **2** (1997), 1–22.
27. Grest, G.S. and Kremer, K.: Molecular dynamics simulation for polymers in the presence of a heat bath, *Phys. Rev. A* **33** (1986), 3628–3631.
28. Allen, M.P. and Tildesley, D.J.: *Computer simulation of liquids*, Oxford University Press, New York, 1987.
29. Thirumalai, D. and Klimov, D.K.: Deciphering the timescales and mechanisms of protein folding using minimal off-lattice models, *Curr. Opin. Struct. Biol.* **9** (1999), 197–207.
30. Finkelstein, A.V. and Badredtinov, A.Y.: Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold, *Fold. Des.* **2** (1997), 115–121.
31. Wolynes, P.G.: Folding funnels and energy landscapes of larger proteins within the capillarity approximation, *Proc. Natl. Acad. Sci. USA* **94** (1997), 6170–6175.
32. Bryngelson, J.D. and Wolynes, P.G.: Spin glasses and the statistical mechanics of protein folding, *Proc. Natl. Acad. Sci. USA* **84** (1987), 7524–7528.

33. Miyazawa, S. and Jernigan, R.L.: Estimation of effective interresidue contact energy from protein crystal structures: quasi-chemical approximation, *Macromolecules* **18** (1985), 534–552.
34. Camacho, C.J.: Entropic barriers, frustration, and order: basic ingredients in protein folding, *Phys. Rev. Lett.* **77** (1996), 2324–2327; Camacho, C.J., Gutin, A., Abkevich, V., Shakhnovich, E.: Comment on 'Chain length scaling of protein folding time' and reply, *Phys. Rev. Lett.* **80** (1998), 207–208.
35. Hoang, T.X. and Cieplak, M.: Protein folding and models of dynamics on the lattice, *J. Chem. Phys.* **109** (1998), 9192–9196.
36. de Gennes, P.G.: Kinetics of collapse for a flexible coil, *J. Phys. Lett.* **46** (1985), L639–L642.
37. Maritan, A., Micheletti, C. and Banavar, J.R.: Role of secondary motifs in fast folding polymers: a dynamical variational principle, *Phys. Rev. Lett.* **84** (2000), 3009–3012.
38. Derrida, B.: Random-energy model: limit of a family of disordered models, *Phys. Rev. Lett.* **45** (1980), 79–82.