Theoretical study of protein folding: outlining folding nuclei and estimation of protein folding rates

2005 J. Phys.: Condens. Matter 17 S1539

(http://iopscience.iop.org/0953-8984/17/18/011)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **17** (2005) S1539–S1551 doi:10.1088/0953-8984/17/18/011

Theoretical study of protein folding: outlining folding nuclei and estimation of protein folding rates

O V Galzitskaya, S O Garbuzynskiy and A V Finkelstein

Institute of Protein Research, Russian Academy of Sciences, 142290, Pushchino, Moscow Region, Russian Federation

E-mail: ogalzit@vega.protres.ru

Received 28 September 2004, in final form 3 November 2004
Published 22 April 2005
Online at stacks.iop.org/JPhysCM/17/S1539

Abstract

Our theoretical approach for prediction of folding/unfolding nuclei in three-dimensional protein structures is based on a search for free energy saddle points on networks of protein unfolding pathways. Under some approximations, this search is rapidly performed by dynamic programming and results in prediction of Φ values, which can be compared with those found experimentally. We show that the presented theoretical approach can be used to outline a folding nucleus in proteins' 3D structure. We demonstrate that incorporation of such 'details' as hydrogen atoms (in addition to the heavy ones) improves prediction of the folding nuclei. The model provides good predictions of folding nuclei for proteins whose 3D structures have been determined by x-ray, and is less successful for proteins whose structures have been determined by NMR. Besides, the same dynamic programming-based calculation yields the transition state free energy, and thus allows one to estimate the protein folding rate. A more direct estimate of the folding rate can be obtained from Monte Carlo simulation of refolding of known 3D protein structure, which is also described in this work. The refolding times obtained from dynamic programming and Monte Carlo simulations correlate reasonably well with logarithms of experimentally measured folding rates at mid-transition.

1. Introduction

1.1. Folding nucleus from experiment and theory

The understanding of the nucleation mechanism has a long, contradictory and still unfinished story. The story started with pioneer experimental studies of protein folding pathways and, specifically, transition states (TSs) on these pathways; this has been done using site-directed

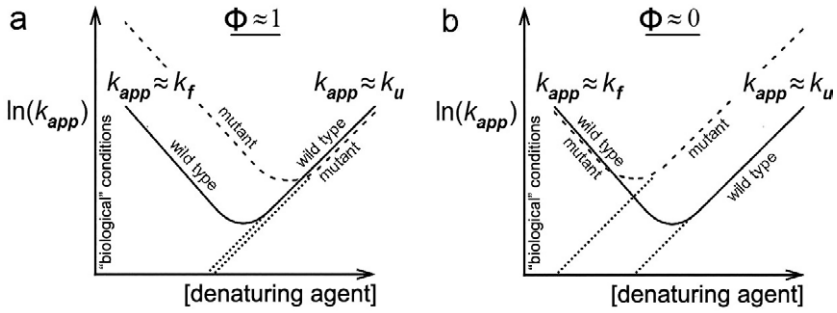

Figure 1. Folding nucleus identification using site-directed mutations (a scheme). (a) Mutation of a residue having its native environment and conformation (i.e., its native interactions) already in the transition state TS changes the mutant's folding rate rather than its unfolding rate. (b) Mutation of a residue which remains denatured in the TS has the opposite effect. 'Wild type' means non-mutated protein. $k_{app} = k_f + k_u$, where k_f is the folding rate and k_u is the unfolding rate: thus, $k_{app} \approx k_f$ in the folding zone (where $k_f \gg k_u$), $k_{app} \approx k_u$ in the unfolding zone (where $k_f \ll k_u$) and $k_f \approx k_u \approx k_{app}/2$ at the mid-transition [2]. Extrapolations which are necessary for Φ_f-value analysis are drawn by dashed lines to the zero denaturant concentration.

mutations [1–3]. The 'folding nucleus', the folded part of the transition state, plays a key role in protein folding: its instability determines the folding and unfolding rates.

It should be stressed that the folding nucleus is not the molten globule, although some of their characteristics may be similar [1]: the nucleus corresponds to the free energy *maximum*, while the molten globule corresponds to the free energy *minimum* [4]. It has been shown that the nucleus looks like some part of the 3D structure of the native protein [1, 2] which is often surrounded by some unstructured, probably molten-globule-like drop.

So far, there is only one (unfortunately, very laborious) experimental method to identify folding nuclei in proteins: to find residues whose mutations affect the folding rate by changing the TS stability as strongly as that of the native protein (figure 1). For the basics of this method and pioneer works see [1, 3, 5–7].

The participation of a residue in the folding nucleus is expressed by the residue's Φ_f value. For a given residue, its Φ_f is defined as

$$\Phi_f = \Delta \ln k_f / \Delta \ln K, \tag{1}$$

where k_f is the folding rate constant, $K = k_f/k_u$ is the folding–unfolding equilibrium constant, and Δ means the shift of the corresponding value induced by mutation of this residue. According to the model of a native-like folding nucleus [1, 2], $\Phi_f = 1$ means that the residue has its native conformation and environment already in the transition state (i.e., that this residue is in the folding nucleus), while $\Phi_f = 0$ means that the residue remains unfolded in the TS. The values $\Phi_f \approx 0.5$ are ambiguous: either the residue is at the surface of the nucleus, or it is in one of the alternative nuclei, belonging to different folding pathways. It is noteworthy that the values $\Phi_f < 0$ and $\Phi_f > 1$ (which would be inconsistent with the model of a native-like folding nucleus) are extremely rare and never concern a residue with a reliable measured $\Delta \ln K$.

To estimate Φ_f, the rates k_f and k_u have to be measured at (or extrapolated to) the same conditions. Usually, being interested in the 'biologically relevant' nucleus, one extrapolates them to the zero denaturant concentration. However, it should be noted that the nucleus corresponding to the protein's mid-transition is outlined more reliably: here the extrapolation is shorter and therefore more robust, especially when the branches of the chevron are curved; the latter suggests a change of the nucleus with the folding conditions [8].

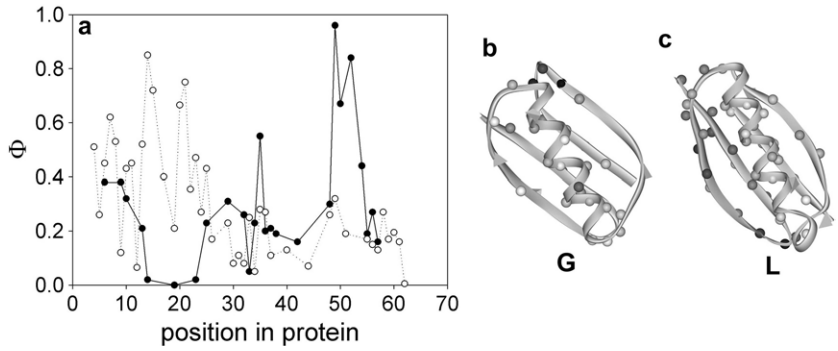

Figure 2. (a) Profiles of experimental Φ_f values obtained for B1 domains of protein G (filled circles) and of protein L (open circles). (b), (c) Schemes of three-dimensional structures of these domains coloured according to the Φ_f values of the amino-acid residues, from white ($\Phi_f = 0$) to black ($\Phi_f = 1$). The experimentally studied residues are shown as beads against the background of the native chain fold. Φ_f values are given for these only. Adapted from [63]. Although sequence identity of B1 domains of G and L proteins is as low as 15% [64], the RMSD between C_α atoms of these two structures after their superposition is 1.35 Å, indicating that the 3D structures of these domains are similar. Nevertheless, their folding nuclei have different locations.

The major assumptions underlying the Φ_f analysis of the folding nucleus by point mutations [1] are that the mutations do not change substantially either the folding pathway, or the nucleus, or the structure of the folded state, or the unfolded state ensemble. Experimentally, this is proved to be usually correct when the mutated residue is not larger than the initial one, and when the mutation is not connected with introduction of charges inside the globule; the proof is made by double mutations [3]. However, some strong mutations can significantly affect the distribution of structures in the TS ensemble [9].

Several other observations have been made.

(1) The TS-stabilizing contacts are very diverse. In some proteins the nucleus is stabilized by hydrophobic interactions [10–12]; in some it includes hydrogen bonds and salt bridges [13, 14].
(2) The position of the nucleus relatively to the whole protein structure is very diverse. In some it is situated in the centre, in the hydrophobic core [10, 12, 15]; in some it is on the boundary of the globule [12–14].
(3) The accessible surface areas of the nuclei are also rather different [16].

Proteins with different amino acid sequences but with similar three-dimensional structures have similar folding nuclei as a rule [17–19]. However, there are several examples which show that this is not always so (figure 2).

Summing up the experimental data, Grantcharova *et al* [20] conclude that mutations, both artificial and natural, can radically change folding pathways (create and destroy folding intermediates, transforming two- into multi-state folding proteins and *vice versa*, shift the folding nuclei to the opposite side of the molecule, etc)—without any considerable variation of three-dimensional structures of native proteins [20]. This means that the native structure is a subject of much more severe natural selection than that of the folding nucleus and than of folding pathways—at least when we speak about relatively small proteins, which in any case usually fold much faster than they are synthesized by a ribosome.

As regards the theoretical search for folding/unfolding nuclei in proteins, several different approaches have been suggested.

The idea of specific nuclei, reinforced by the lattice simulations of protein folding [21], generated an evolutionary approach to prediction of the nuclei. It is based on the search for a set of highly conserved residues having no obvious functional role [22–25]. It should be mentioned that this approach, at best, can give only the common part of the nuclei existing in homologous proteins. Moreover, some recent observations show that the residue conservatism across the homologous proteins correlates with deep immersion into the hydrophobic core of a protein [25] rather than into the folding nucleus [26]. It should be noted that there is some correlation between the nuclei (the regions of high Φ_f values) and the hydrophobic cores and secondary structures [27–30], but it is rather low on average [26, 28].

The most direct approach to the theoretical search for the nucleus is to generate a plausible transition state for unfolding (which must coincide with that for folding closely to mid-transition) using the all-atom molecular dynamic simulations of protein unfolding [31–33]. According to these simulations, held for very few small proteins at highly denaturing conditions (otherwise, the calculation takes too long), the unfolding is hierarchic [34–36] (at least when it occurs far from the equilibrium): tertiary interactions break early, whereas secondary structures remain longer. The repeated trajectories show a statistical distribution around the experimentally found transition states and demonstrate a broad ensemble of the TS structures. However, these simulations usually need extremely denaturing conditions (500 K, etc) to be completed. Therefore, the transition states found for such an extreme *un*folding can be, in principle, rather different from those existing for folding [37]. Recently, however, some molecular dynamic simulations of unfolding of very small proteins [38–40] have been performed at more realistic, although also highly denaturating conditions. They have been performed at temperatures accessible for 'wet' experiments (350 K), as well as for simulations on current supercomputers. They gave TS structures which are consistent with experiment [40]; however, these simulations take enormous time and can be performed for very small proteins only.

Further progress is due to the analysis of multidimensional networks of the protein folding–unfolding trajectories performed by various algorithms [41, 42]. All these approaches [41–43] use different approximations and algorithms, consider only the attractive native interactions (the 'Gō model' [44]) to reduce the energy frustrations and heterogeneity of interactions, and model the trade-off between the formation of attractive interactions and the loss of conformational entropy during protein folding. These works also simulate unfolding of known 3D protein structures rather than their folding, but the unfolding is considered close to the mid-transition point, where folding and unfolding pathways coincide according to the detailed balance principle. Under these 'near-equilibrium' conditions, all single-domain proteins demonstrate two-state (i.e., 'all-or-none') transitions both in thermodynamics [45] and kinetics [46, 47]. This means that at the mid-transition all semi-folded and misfolded globules are unstable relative to both native and unfolded states of protein chain, and this allows us to take into account only the pathways going from the native to the unfolded state and to neglect those leading to misfolded globules, stabilized by non-native interactions.

These works allowed the authors to outline the folding nuclei. Despite the relative simplicity of these models, they give a promising ($\sim$50%) correlation with experimental Φ values [48–51]. This suggests that the chain's folding pattern and the size of the protein, taken into account by these models, play more important roles in folding than the high resolution details of protein structure [42, 52, 53].

In this paper we briefly describe our approach for the prediction of folding nuclei and estimation of protein folding rates.

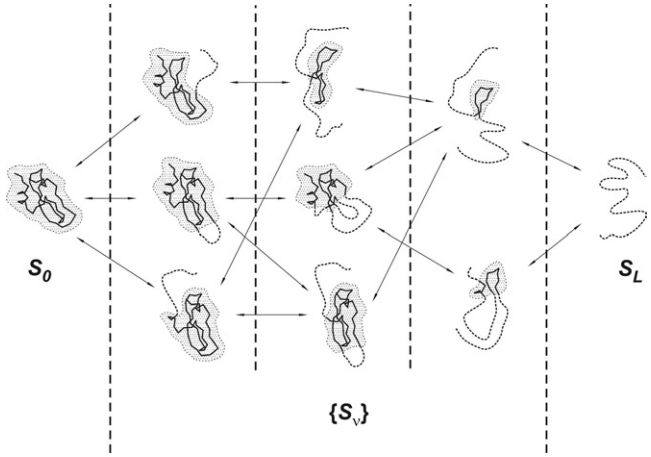

Figure 3. A sketch of the network of pathways of sequential unfolding (and folding) of native 3D protein structure (S_0). S_L is the coil where all L links of the protein chain are disordered. In each of the many intermediates of the type S_ν, ν chain links (shown in the dashed line) are unfolded, while the other L–ν links keep their native positions and conformations (they are shown as the solid line against the background of a dotted cloud denoting the globular part of the intermediate). The central structure in the lower line exemplifies a microstate with ν unfolded links forming one closed unfolded loop and one unfolded tail; the central structure in the central line exemplifies a microstate where ν unfolded links form two closed unfolded loops. The networks used in computations are much larger than the one shown in the sketch: they include millions of semi-folded microstates.

2. Outlining folding nuclei

2.1. Approximations used in the model and estimation of free energy

We consider a network of simplified stepwise unfolding pathways (see figure 3), each step of which is the removal of one 'chain link' (which includes several residues) from the native protein 3D structure. The removed chain fragments are assumed to form a random coil; they lose all their non-bonded interactions and gain the coil entropy. The next is the assumption that the chain residues remaining in the globule keep their native positions and conformations and that the unfolded regions do not fold into another, non-native globule. Thus, we actually neglect non-native interactions. The main simplification is that we concentrate on the TS and its free energy, rather than on a detailed description of the chain motions.

To use dynamic programming in searching for the TS for a network of folding–unfolding pathways, we have, for computational reasons, to restrict this network by no more than $\sim 10^6$ intermediate microstates. Therefore, we divide an N-residue protein chain into $L \sim 20$–30 chain links. For the same computational reasons, we consider only the intermediates with no more than two closed disordered loops in the middle of the chain plus the N- and the C-terminal disordered tails.

Thus, our model considers the native structure S_0, the unfolded state S_L and an ensemble of intermediate microstates S_ν consisting of a native-like part and of ν unfolded chain links ($\nu = 0$ for S_0, $\nu = L$ for S_L, L being the total number of the chain links, and $\nu = 1, \ldots, L-1$ for the semi-folded intermediates with ν disordered links). The model uses a simple free energy estimate [51]:

$$F(S) = \varepsilon \times n_S^{nb} - T\left[\nu_S \times \sigma + \sum_{\text{loops}\in S} S_{\text{loop}}\right]. \tag{2}$$

Here S is a microstate; n_S^{nb} is the number of native atom–atom contacts in the native-like part of S (n_S^{nb} does not include contacts of neighbour residues, also existing in the coil); ε is the energy of one contact; ν_S is the number of residues in the unfolded part of S; T is the temperature; and σ is the entropy difference between the coil and the native state of a residue (we take $\sigma = 2.3R$ according to Privalov [45], R being the gas constant). The sum Σ is taken over all closed unfolded loops (see the legend to figure 3) protruding from the native-like part of S.

At the point of equilibrium between the native state S_0 and the coil S_L, we have $F(S_0) = F(S_L)$; i.e., the average contact energy ε (which is influenced by the solvent and the temperature) is

$$\varepsilon = -TN\sigma/n_0^{nb}. \tag{3}$$

Here at the mid-transition n_0^{nb} is the number of contacts in the native structure and N is the total number of the protein chain residues. It follows from equations (2) and (3) that the $F(S)/T$ values (which only determine the transition state, see equation (6) below) do not depend on temperature, provided that the solvent composition corresponds to the mid-transition at this temperature.

The entropy spent to close a disordered loop between the still fixed residues k and l is estimated [52] as

$$S_{\text{loop}} = -\tfrac{5}{2}R\ln|k-l| - \tfrac{3}{2}R(r_{kl}^2 - a^2)/(2Aa|k-l|); \tag{4}$$

here r_{kl} is the distance between the C_α atoms of residues k and l, $a = 3.8$ Å is the distance between the neighbour C_α atoms in the chain, and A is the persistence length for a polypeptide (according to Flory [54], we take $A = 20$ Å). The term $-\frac{5}{2}R\ln|k-l|$ is the main term in equation (3); the coefficient $-\frac{5}{2}$ (rather than Flory's value $-\frac{3}{2}$) follows from the condition that a loop cannot penetrate inside the globule [52].

2.2. Transition states at the protein unfolding pathways

Let us consider some unfolding pathway $w = (S_0 \to S_1 \to \cdots \to S_L)$; then $F_w^{\#} = \max\{F(S_0), F(S_1), \ldots, F(S_L)\}$ is the free energy of the TS ('free-energy barrier') for pathway w. The most efficient kinetic pathway has the minimal (over all the pathways) TS free energy, $F_{\min}^{\#} = \min_{\text{possible } w}\{F_w^{\#}\}$: this pathway passes from S_0 (the native state) to S_L (the coil) via the lowest free energy barrier. Let $S_{\nu-1} \in \{S_{\nu-1} \to S_\nu\}$ mean that $S_{\nu-1}$ can be transformed into S_ν in an elementary step (i.e., by removal of one link from the globular part of $S_{\nu-1}$). At every pathway $S_0 \to S_1 \to \cdots \to S_{L-1} \to S_L$, all intermediates satisfy conditions $S_1 \in \{S_1 \to S_2\}, \ldots, S_{L-2} \in \{S_{L-2} \to S_{L-1}\}$ (while the condition $S_{L-1} \in \{S_{L-1} \to S_L\}$ is satisfied automatically). Thus,

$$\begin{aligned} &F_{\min}^{\#} = \min\{\max\{F(S_0), F(S_1), \ldots, F(S_L)\}\}, \\ &S_1, \ldots, S_{L-1} \\ &S_1 \in \{S_1 \to S_2\} \\ &\ldots \\ &S_{L-2} \in \{S_{L-2} \to S_{L-1}\} \end{aligned} \tag{5}$$

where the maximum is taken over the microstates' free energies along every pathway $S_0 \to S_1 \to \cdots \to S_{L-1} \to S_L$, and the minimum is taken across all the pathways. Despite a huge number of possible pathways, the $F_{\min}^{\#}$-value can be calculated by dynamic programming [55, 56].

The intermediates S with $F^{\#}(S) = F_{\min}^{\#}$ give a narrow ensemble of 'the best' transition microstates $\{S_{\min}^{\#}\}$ with the minimal free energy, while the intermediates with $F^{\#}(S) = F(S)$

give a more broad ensemble $\{S^{\#}\}$ of all the possible passes over the free energy barrier separating S_0 from S_L. Although the ensemble $\{S^{\#}\}$ may be somewhat redundant (since a pathway to the TS high in free energy may pass via some TS of the lower free energy), it has been shown [41] that this ensemble of all the possible passes describes the TS better than the ensemble $\{S^{\#}_{\min}\}$ of 'the best' TSs only. Further, the ensemble $\{S^{\#}\}$ is used only.

To outline the nucleus, we investigate the ensembles $\{S^{\#}\}$ of all possible transition states. The value of the Boltzmann probability of microstate $S^{\#}$ in the ensemble $\{S^{\#}\}$ is

$$P(S^{\#}) = \exp(-F(S^{\#})/RT)/\exp(-F^{\#}/RT), \tag{6}$$

where

$$\exp(-F^{\#}/RT) = \Sigma_{S\#}\exp(-F(S^{\#})/RT) \tag{7}$$

is the partition function of the totality of transition states, and $F^{\#}$ is their total free energy. The sum is taken over the whole ensemble $\{S^{\#}\}$. The lower the free energy $F(S^{\#})$, the higher the weight $P(S^{\#})$, the more rapid the pathway via this $S^{\#}$ (according to the conventional exponential dependence of reaction rate on the transition state free energy [57]), and therefore the more the chains use this pass $S^{\#}$ at folding and unfolding.

2.3. Computation of Φ *values*

The theory estimates the Φ values as follows. According to equations (1) and (2), the value $\Delta \ln K = \Delta_r[F(S_0) - F(S_L)]$ is equal to $\varepsilon \times \Delta_r(n_0^{nb})$, where ε is the contact energy, and $\Delta_r(n_0^{nb})$ is the residue r mutation-induced change in the number of contacts in the native state S_0 (since all native contacts are assumed to be equal, and no contacts are assumed to be present in the unfolded state S_L).

Correspondingly, $\Delta \ln k_{\mathrm{f}} = \Delta_r[F(TS) - F(S_L)] = \varepsilon \times \langle\Delta_r(n_S^{nb})\rangle_{S\#}$, where $\langle\Delta_r(n_S^{nb})\rangle_{S\#}$ is the same residue r mutation-induced change in the number of native contacts in the transition state, averaged over the TS ensemble $\{S^{\#}\}$. This change can be calculated as

$$\langle\Delta_r(n_S^{nb})\rangle_{S\#} = \Sigma_{S\#}P(S^{\#})\Delta_r(n_{S\#}^{nb}), \tag{8}$$

where $P(S^{\#})$ is the Boltzmann probability of microstate $S^{\#}$ in the TS ensemble (see equation (9)), and $\Delta_r(n_{S\#}^{nb})$ is the residue r mutation-induced change in the number of native contacts in microstate $S^{\#}$.

The values $\Delta_r(n_S^{nb})$ can be calculated for each microstate S from atomic coordinates of non-mutated protein when we know what atoms are deleted or substituted in the mutant. However, this calculation assumes that the protein structure is not disturbed by mutation. Therefore, we have to consider only those mutations which do not insert new atomic groups.

The computed values

$$\Phi = \langle\Delta_r(n_S^{nb})\rangle_{S\#}/\Delta_r(n_0^{nb}) \tag{9}$$

are to be compared with the experimental Φ_{f} values to estimate the correlation between the theory and experiment.

2.4. Calculated Φ *values in comparison with experiment*

The calculations are performed for all 17 proteins, whose experimental Φ_{f} values are known for many residues, and the 3D structure is known as well [51]. Transition states for the folding–unfolding pathways are found using the DP technique and a simple free energy estimate given by equation (2). All the computations concern the point of thermodynamic equilibrium between the native and the coil state of each separate wild-type protein. This point is defined by equation (3).

The averaged (over all 17 proteins) coefficient of correlation between computed and experimental Φ values is 0.40 ± 0.37 when the calculations are based on the heavy atoms only. This is close to the correlations obtained earlier by us [41] and by other researchers, who used considerable sets of proteins to verify results of their calculations [42, 43, 50]. In particular, this coefficient is 0.35 ± 0.34 for all four methods considered by Alm *et al* [50].

However, when we take into account hydrogen atoms in addition to the heavy ones, the averaged correlation coefficient increases to 0.54 ± 0.27 and the results become more significant. The positive effect caused by inclusion of hydrogen atoms seems to be due to increased (and thus more reliable) statistics of contacts.

The Φ-value plots (see figure 3 in [51]) show that theoretical and experimental data are usually in a reasonable agreement. But, although the correlation with experiment is rather high (0.6–0.9, see figures 4(a), (b)) for half of the proteins used in this study, for some other proteins the correlation is below 0.35 (and even negative, see figure 4(c)). The latter mostly concerns proteins with the NMR-resolved 3D structures.

The prediction of Φ values for the β hairpin of the B1 domain of protein G has been made in the work of Chang *et al* [58]. Our predicted Φ values for this β hairpin in the whole protein are in agreement with published experimental data (the correlation coefficient for the whole protein is 0.76; see figure 4(b)) and have the same behaviour as in the work of Chang *et al* [58].

We observe that the nuclei outlined in x-ray structures have much better agreement with experimental data than the ones outlined in NMR structures. That is, the averaged correlation coefficient for 11 x-ray-resolved protein structures is 0.65 ± 0.18 when all (heavy and hydrogen) atoms are taken into account. In a contrast, for six NMR-resolved proteins the averaged correlation coefficient is only 0.34 ± 0.32, even when all atoms are taken into account. Thus, the NMR structures, which are less accurate than the x-ray ones [59], are less suitable for calculation of Φ values.

3. Estimation of protein folding rates

3.1. Estimation of protein folding rate from calculation of transition state free energy

Since the folding rate k_{f} should be proportional to $\exp(-F^{\#}/RT)$, and since our model allows us to calculate the transition state free energy $F^{\#}$ (see equation (7)), we can estimate a correlation of computed $F^{\#}/RT$ values with experimentally obtained folding rates k_{f} (or rather, $\ln(k_{\mathrm{f}})$). The computed (for mid-transition, see above) $-F^{\#}/RT$ values are in a good correlation with $\ln(k_{\mathrm{f\ at\ mid-transition}})$: the correlation coefficient is 0.73 ± 0.11 (figure 5). However, these $-F^{\#}/RT$ values virtually do not correlate with experimental folding rates in water, far from the mid-transition (or rather, with $\ln(k_{\mathrm{f\ in-water}})$): here the correlation coefficient is equal to 0.18 ± 0.23. The absence of the latter correlation is not a surprise, since the computed $F^{\#}$ values correspond to the mid-transition rather than to the in-water ('biological') conditions.

3.2. Advantage and disadvantage of dynamic programming and Monte Carlo methods

To use dynamic programming in searching for the TS for a network of folding–unfolding pathways, we have, for computational reasons, to restrict this network to no more than $\sim 10^6$ intermediates. Therefore, we divide an N-residue protein chain into $L \sim 20$–30 chain links and use a limited number of loops. Using a Monte Carlo (MC) simulation of folding, we can, in principle, avoid these simplifications.

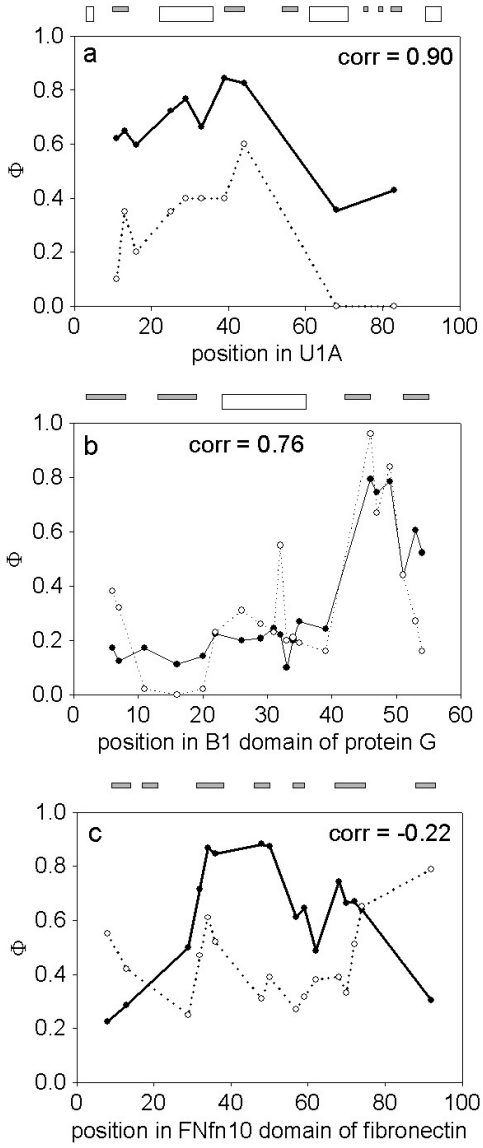

Figure 4. Profiles of Φ values for proteins with high correlation with experiment ((a), (b)) and with the worst correlation with experiment (c). Open circles connected with a dotted line are experimental Φ_f values; filled circles connected with a continuous line are theoretical Φ values for the same residues. The Φ-value calculations are done with hydrogen atoms and with contact distance between heavy atoms $r_{cont} = 6$ Å, contact distance between hydrogen and heavy atoms $r_{H-heavy} = 5$ Å, contact distance between hydrogen atoms $r_{HH} = 4$ Å; the links include $l = 5$ residues. The rectangles at the top of each plot show the native positions in the chain of α and 3_{10} helices (broad rectangles), and of β strands (narrow shaded rectangles).

3.3. Estimation of protein folding rate from Monte Carlo simulations

We investigated folding of the known native structures, starting from the unfolded chain. This was done by Monte Carlo (MC), using the Metropolis scheme [60] at the point of mid-transition.

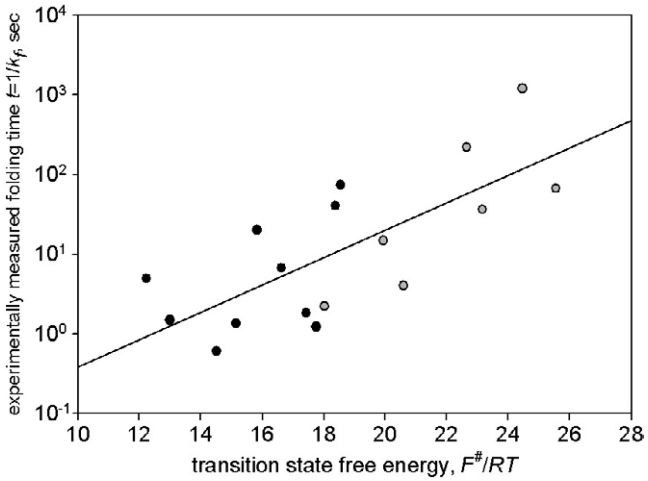

Figure 5. Correlation between the computed transition state free energy $F^{\#}/RT$ and the mid-transition folding time $t = 1/k_f$ (measured in seconds and represented in a logarithmic scale) for all 17 investigated wild-type proteins (PDB entries are the following: 1bf4, 1btb, 1fkb, 1pgb, 1ris, 1rnb, 1shg, 1srm, 1ten, 1tiu, 1ttf, 1urn, 2ci2, 2ptl, 2vil, 3chy; the 3D structure of the suc1 protein was kindly presented by J Schymkowitz). The correlation coefficient is 0.73 ± 0.11. Ten black circles correspond to the proteins presented in figure 7.

The free energy function is given by equations (2), (3). The kinematic scheme of elementary movements includes, as above, insertion of a residue from the coil to its native position in the known 3D structure or removal of a residue from its native position to the coil [61]. In this way, we did a travel from the unfolded state to the known 3D structure without visiting the misfolded states.

An elementary MC step was performed as follows. We randomly chose a residue. If the chosen residue had been already fixed in the native position we tried to put it to the coil. If the chosen residue was in the coil, we tried to put it in the native position. Then we computed the free energy difference, ΔF, between new and previous structures. According to Metropolis *et al* [60], the MC step leads to the new structure with a probability $w = \exp(-\Delta F/RT)$ if $\Delta F > 0$, and $w = 1$ if $\Delta F \leqslant 0$.

To estimate the characteristic time of coming to the native structure (first passage time, $t_{1/2}$) we performed 50 MC runs for every protein [62], and $t_{1/2}$ was determined as the number of MC steps required to complete 50% of MC runs (25 of 50 runs) (see [62]). Having a limit of 10^8 MC steps, we were only able to calculate $t_{1/2}$ for ten proteins from the 17 investigated here. Figure 6 presents typical MC kinetics for one of the proteins. It should be noted that eight of the ten proteins reached the native state while the remaining two proteins arrived at the stable states which were close to the native state but still had several residues unstructured. It is interesting that the structures of both these proteins are NMR resolved.

The computed (for mid-transition) $t_{1/2}$ values for ten proteins are in a good correlation with those experimentally measured at mid-transition protein folding time: the correlation coefficient is 0.70 ± 0.05 (figure 7). It should be mentioned here that the coefficient of correlation obtained with the Monte Carlo method is better than one obtained from dynamic programming-based calculation for the same ten proteins (the correlation coefficient is 0.48 ± 0.10).

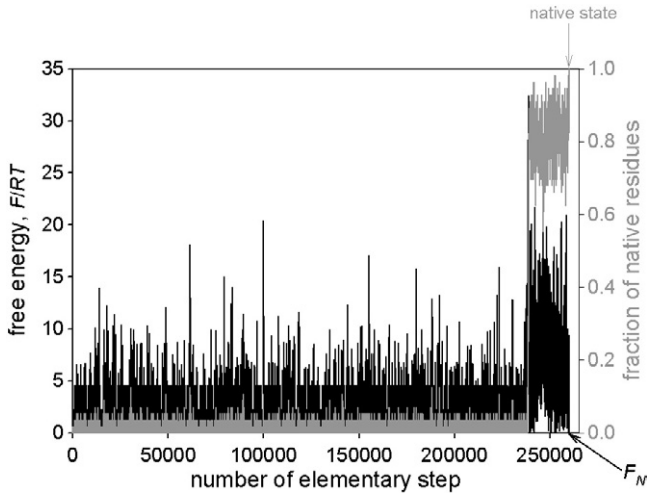

Figure 6. Monte Carlo kinetics for refolding of src SH3 domain. The plots show a dependence of the free energy (black line) and fraction of native residues (grey line) on the number of MC step. The native state here is achieved in 2.5×10^5 steps. Arrows point to the native state and to $F_{\rm N}$, free energy of the native state.

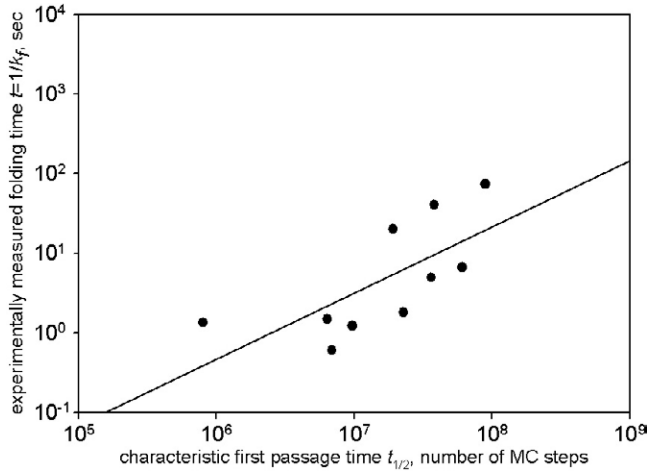

Figure 7. Correlation between the computed characteristic first passage time $t_{1/2}$ (the number of MC steps made until half of the molecules fold) and the experimentally measured folding time (i.e., $t = 1/k_{\rm f}$) at the mid-transition for ten proteins (PDB entries are as follows: 1bf4, 1pgb, 1rnb, 1shg, 1srm, 1ttf, 2ci2, 2ptl, 2vil, 3chy), where the folding simulation has been completed within 10^8 MC steps. The experimental folding time t is measured in seconds; the time of simulation, $t_{1/2}$, is measured in MC steps. Both are represented in a logarithmic scale. The correlation coefficient is 0.70 ± 0.05.

4. Conclusion

This work shows that the presented theoretical approach is able to outline the folding nucleus in proteins' 3D structure. Thus, this approach captures some basic characteristics of protein folding and unfolding, though it neglects many details of inter-residue interactions and chain movements. The model provides good predictions of folding nuclei for proteins whose 3D

structures have been determined by x-rays, and exhibits a more limited success for proteins whose structures have been determined by NMR. Besides, the same dynamic programming-based calculation yields the transition state free energy, and thus allows one to estimate the protein folding rate. A more direct estimate of the folding rate can be obtained from Monte Carlo simulation of refolding of known 3D protein structure, which is also described in this work. The refolding times obtained from dynamic programming and Monte Carlo simulations correlate reasonably well with logarithms of experimentally measured folding rates at mid-transition.

Acknowledgments

This work was supported by the programme MCB RAS, by the Russian Science School programme, by the Russian Foundation for Basic Research, and by an International Research Scholar's Award to AVF from the Howard Hughes Medical Institute.

References

[1] Matouschek J T, Kellis Jr, Serrano L and Fersht A R 1989 Mapping the transition state and pathway of protein folding by protein engineering *Nature* **340** 122–6
[2] Matouschek J T, Kellis Jr, Serrano L, Bycroft M and Fersht A R 1990 *Nature* **346** 440–5
[3] Fersht A R, Matouschek A and Serrano L 1992 *J. Mol. Biol.* **224** 771–82
[4] Ptitsyn O B 1995 *Adv. Protein Chem.* **47** 83–229
[5] Leffler J E and Grunwald E 1963 *Rates and Equilibria of Organic Chemistry* (New York: Dover)
[6] Matthews C R 1987 *Methods Enzymol.* **154** 127–32
[7] Goldenberg D P, Frieden R W, Haack J A and Morrison T B 1989 *Nature* **338** 498–511
[8] Otzen D E, Kristensen O, Proctor M and Oliveberg M 1999 *Biochemistry* **38** 6499–511
[9] Burton R E, Huang G S, Daugherty M A, Calderoni T L and Oas T G 1997 *Nat. Struct. Biol.* **4** 305–10
[10] Itzhaki L S, Otzen D T and Fersht A R 1995 *J. Mol. Biol.* **254** 260–88
[11] Fulton K, Main E, Daggett V and Jackson S E 1999 *J. Mol. Biol.* **291** 445–61
[12] Kragelund B B, Osmark P, Neergaard T B, Schiodt J, Kristiansen K, Knudsen J and Poulsen F M 1999 *Nat. Struct. Biol.* **6** 594–601
[13] Lopez-Hernandez E and Serrano L 1996 *Fold. Des.* **1** 43–55
[14] Grantcharova V P, Riddle D S, Santiago J V and Baker D 1998 *Nat. Struct. Biol.* **5** 714–20
[15] Chiti F, Taddei N, White P, Bucciantini M, Magherini F, Stefani M and Dobson C 1999 *Nat. Struct. Biol.* **6** 1005–9
[16] Jackson S E 1998 *Fold. Des.* **3** R81–91
[17] Martinez J C and Serrano L 1999 *Nat. Struct. Biol.* **6** 1010–6
[18] Riddle D S, Grantcharova V P, Santiago J V, Alm E, Ruczinskiy I and Baker D 1999 *Nat. Struct. Biol.* **6** 1016–24
[19] Perl D, Welker Ch, Schindler T, Schroder K, Marahiel M A, Jaenicke R and Schmid F X 1998 *Nat. Struct. Biol.* **5** 229–35
[20] Grantcharova V, Alm E J, Baker D and Horwich A L 2001 *Curr. Opin. Struct. Biol.* **11** 70–82
[21] Abkevich V I, Gutin A M and Shakhnovich E I 1994 *J. Chem. Phys.* **101** 6052–62
[22] Shakhnovich E, Abkevich V and Ptitsyn O 1996 *Nature* **379** 96–8
[23] Ptitsyn O B 1998 *J. Mol. Biol.* **278** 655–66
[24] Ptitsyn O B and Ting K-L 1999 *J. Mol. Biol.* **291** 671–82
[25] Mirny L A and Shakhnovich E I 1999 *J. Mol. Biol.* **291** 177–96
[26] Plaxco K W, Larson S, Ruczinski I, Riddle D S, Thayer E C, Buchwitz B, Davidson A R and Baker D 2000 *J. Mol. Biol.* **298** 303–12
[27] Nölting B and Andret K 2000 *Proteins Struct. Funct. Genet.* **41** 288–98
[28] Galzitskaya O V, Skoogarev A V, Ivankov D N and Finkelstein A V 1999 Folding nuclei in 3D protein structures *Proc. Pacific Symp. on Biocomputong'2000* ed R B Altman, A K Dunker, L Hunter, K Lauderdale and T E Klein (Singapore: World Scientific) pp 131–42
[29] Cota E, Steward A, Fowler S B and Clarke J 2001 *J. Mol. Biol.* **305** 1185–94
[30] Mirny L and Shakhnovich E 2001 *J. Mol. Biol.* **308** 123–9
[31] Li A and Daggett V 1996 *J. Mol. Biol.* **257** 412–29

[32] Caflisch A and Karplus M 1995 *J. Mol. Biol.* **252** 672–708
[33] Brooks C L III, Gruebele M, Onuchic J N and Wolynes P G 1998 *Proc. Natl Acad. Sci. USA* **95** 11037–8
[34] Lazaridis T and Karplus M 1997 *Science* **278** 1928–31
[35] Tsai J, Levitt M and Baker D 1999 *J. Mol. Biol.* **291** 215–25
[36] Daggett V and Fersht A R 2003 *Nat. Rev. Mol. Cell Biol.* **4** 497–502
[37] Finkelstein A V 1997 *Protein Eng.* **10** 843–5
[38] Mayor U, Johnson C M, Daggett V and Fersht A R 2000 *Proc. Natl Acad. Sci. USA* **97** 13518–22
[39] Ferguson N, Pires J R, Toepert F, Johnson C M, Pan Y P, Volkmer-Engert R, Schneider-Mergener J, Daggett V, Oschkinat H and Fersht A 2001 *Proc. Natl Acad. Sci. USA* **98** 13008–13
[40] Mayor U, Guydosh N R, Johnson C M, Grossman J G, Sato S, Jas G S, Freund S M V, Alonso D O V, Daggett V and Fersht A R 2003 *Nature* **421** 863–7
[41] Galzitskaya O V and Finkelstein A V 1999 *Proc. Natl Acad. Sci. USA* **96** 11299–304
[42] Alm E and Baker D 1999 *Proc. Natl Acad. Sci. USA* **96** 11305–10
[43] Munõz V and Eaton W A 1999 *Proc. Natl Acad. Sci. USA* **96** 11311–6
[44] Taketomi H, Ueda Y and Gō N 1975 *Int. J. Pept. Protein Res.* **7** 445–59
[45] Privalov P L 1979 *Adv. Protein Chem.* **33** 167–241
[46] Fersht A R 1995 *Curr. Opin. Struct. Biol.* **5** 79–84
[47] Fersht A R 1997 *Curr. Opin. Struct. Biol.* **7** 3–9
[48] Baker D 2000 *Nature* **405** 39–42
[49] Takada S 1999 *Proc. Natl Acad. Sci. USA* **96** 11698–700
[50] Alm E, Morozov A V, Kortemme T and Baker D 2002 *J. Mol. Biol.* **322** 463–76
[51] Garbuzynskiy S O, Finkelstein A V and Galzitskaya O V 2004 *J. Mol. Biol.* **336** 509–25
[52] Finkelstein A V and Badretdinov A Ya 1997 *Fold. Des.* **2** 115–21
[53] Clementi C, Jennings P A and Onuchic J N 2000 *Proc. Natl Acad. Sci. USA* **97** 5871–6
[54] Flory P J 1969 *Statistical Mechanics of Chain Molecules* (New York: Interscience)
[55] Aho A, Hopcroft J and Ullman J 1976 *The Design and Analysis of Computer Algorithms* (Reading, MA: Addison-Wesley)
[56] Finkelstein A V and Roytberg M A 1993 *Biosystems* **30** 1–19
[57] Moore J W and Pearson R G 1981 *Kinetics and Mechanism* (New York: Wiley)
[58] Chang I, Cieplak M, Banavar J R and Maritan A 2004 *Protein Sci.* **13** 2446–57
[59] Bastolla U, Farwer J, Knapp E W and Vendruscolo M 2001 *Proteins Struct. Funct. Genet.* **44** 79–96
[60] Metropolis N, Rosenbluth A, Rosenbluth M, Teller A and Teller E 1953 *J. Chem. Phys.* **96** 768–80
[61] Galzitskaya O V and Finkelstein A V 1998 *Fold. Des.* **3** 69–78
[62] Galzitskaya O V and Finkelstein A V 1995 *Protein Eng.* **8** 883–92
[63] Galzitskaya O V 2002 *Mol. Biol.* **36** 386–90
[64] McCallister E L, Alm E and Baker D 2000 *Nat. Struct. Biol.* **7** 669–73