ISSN 0006-2979, Biochemistry (Moscow), 2010, Vol. 75, No. 6, pp. 717-727. © Pleiades Publishing, Ltd., 2010.
Published in Russian in Biokhimiya, 2010, Vol. 75, No. 6, pp. 807-818.
Originally published in Biochemistry (Moscow) On-Line Papers in Press, as Manuscript BM10-023, May 9, 2010.

Is Protein Folding Rate Dependent on Number of Folding Stages? Modeling of Protein Folding with Ferredoxin-Like Fold

O. V. Galzitskaya

Institute of Protein Research, Russian Academy of Sciences, 142290 Pushchino,
Moscow Region, Russia; fax: 8(4967)318-435; E-mail: ogalzit@vega.protres.ru

Received January 27, 2010
Revision received February 24, 2010

Abstract—Statistical analysis of protein folding rates has been done for 84 proteins with available experimental data. A surprising result is that the proteins with multi-state kinetics from the size range of 50-100 amino acid residues (a.a.) fold as fast as proteins with two-state kinetics from the same size range. At the same time, the proteins with two-state kinetics from the size range 101-151 a.a. fold faster than those from the size range 50-100 a.a. Moreover, it turns out unexpectedly that usually in the group of structural homologs from the size range 50-100 a.a., proteins with multi-state kinetics fold faster than those with two-state kinetics. The protein folding for six proteins with a ferredoxin-like fold and with a similar size has been modeled using Monte Carlo simulations and dynamic programming. Good correlation between experimental folding rates, some structural parameters, and the number of Monte Carlo steps has been obtained. It is shown that a protein with multi-state kinetics actually folds three times faster than its structural homologs.

DOI: 10.1134/S0006297910060064

Key words: protein folding, intermediates of folding, two-state transition, folding nucleus, folding rates

The nature of protein folding is an important contemporary problem in biophysics. Besides of its fundamental significance, the understanding of the mechanism of protein folding is of great importance for many practical tasks such as development of drugs and design of *de novo* artificial proteins with given properties [1, 2]. Misfolding of proteins *in vivo* is often concomitant with their aggregation and can be in relevant to many diseases [3, 4].

Today theoretical and experimental investigations of the mechanism of self-organization of globular proteins yield qualitative understanding of protein folding [5-12]. However, the details of the protein folding process are still unknown. The few models of protein folding [13-16] allow us to describe more or less correctly only the folding of small one-domain globular proteins (less than 200 amino acid residues (a.a.)).

The time of folding of different proteins varies over many orders of magnitude, from microseconds to seconds and even hours. Small proteins usually fold quickly and without intermediates of folding (i.e. the process occurs in one stage, and “one-stage” kinetics are observed). Larger proteins fold slower, and during folding metastable intermediates are often observed (i.e. the process of folding occurs in many stages, and “multi-state” kinetics are observed) [11].

One of the first analytical theories of folding of one-domain globular proteins was that of Finkelstein and Badretdinov [7, 10]. In the framework of this theory developed on the base of a nucleation mechanism, it has been shown that the rate of protein folding at the point of the thermodynamic equilibrium depends on the border between two phases in the transition state, and loops protruding from the folded part of the protein create additional surface tension slowing the process of protein folding (Fig. 1). This implies that since the border between two phases depends for a spherical globule on the number of amino acid residues in the protein chain as $L^{2/3}$, then the rate of protein folding should depend on the number of amino acid residues in protein L in the same way:

$$\ln k_{mt} \sim -L^{2/3}, \qquad (1)$$

where $\ln k_{mt}$ is the logarithm of protein folding at the point of thermodynamic equilibrium of two states of the protein, the native and the denatured ones. However, the correlation between $\ln k_{mt}$ and $L^{2/3}$ does not reach unity (see Table 1). This makes us take into account other factors that can additionally affect protein folding rate as well as the length of the protein chain.

Simultaneously, another analytical estimation of the dependence of the protein folding rate on the protein chain size has been suggested on the basis that the native state is more stable than the unfolded one [5]:

$$\ln k_f^w \sim -L^{1/2}, \qquad (2)$$

where $\ln k_f^w$ is the logarithm of the protein folding rate in water. For this the correlation coefficient for a set of 69 proteins was 0.74 [17].

In computer experiments on folding of model protein chains using a cubic lattice [6], it was found that when the stability of the native state is maximal the natural logarithm of rate should depend on the size of the protein chain as:

$$\ln k_f^w \sim -\ln L, \qquad (3)$$

For this case the correlation coefficient was 0.78 for the set of 69 proteins [17].

In a number of computer experiments it has been shown [8] that the rate of folding of model proteins on a lattice with simple potentials depends on the protein size as $\ln k_{mt} \sim -L^{0.61\pm0.18}$. This is in good agreement with dependences found by both Finkelstein and Thirumalai [5, 7, 10].

Despite the different views of dependences of the rate of folding on the length of the protein chain, all analytical theories are similar in that, as it intuitively seems, the time of protein folding should grow with increasing length of protein chain, it growing slower than exp(L).

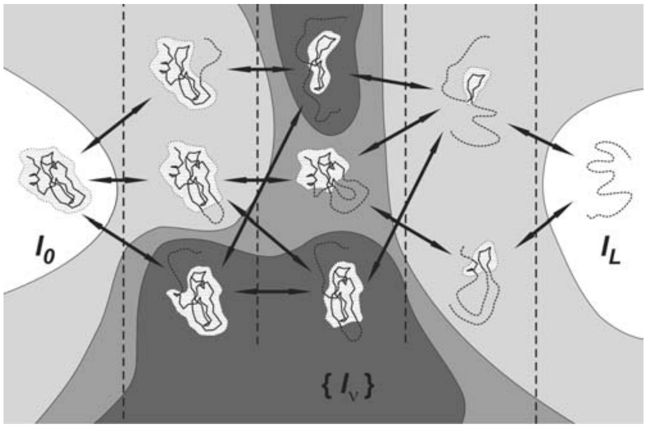

Fig. 1. Sketch of the network of pathways of reversible folding/unfolding of native 3D protein structure (I_0). I_L is coil where all L links of the protein chain are disordered. In each of various intermediates of type I_ν, the ν chain links (shown by dashed lines) are unfolded, while the other $L - \nu$ links keep their native positions and conformations (they are shown as a 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 of pathways used in the calculations are much more complex than this scheme: they include millions of semi-folded microstates.

Table 1. Correlation coefficients between logarithms of folding rates in water and at the point of thermodynamic equilibrium and structural parameters related or not related (*CO*) with size of the protein

Parameter	$\ln k_f$, sec^{-1} 84 proteins	$\ln k_f^{multi}$, sec^{-1} 26 proteins	$\ln k_f^{two}$, sec^{-1} 58 proteins
L	−0.65 ± 0.06	−0.78 ± 0.08	−0.42 ± 0.11
$L^{1/2}$	−0.70 ± 0.06	−0.81 ± 0.07	−0.50 ± 0.10
$L^{2/3}$	−0.69 ± 0.06	−0.80 ± 0.07	−0.47 ± 0.10
$\ln L$	−0.71 ± 0.05	−0.82 ± 0.06	−0.55 ± 0.09
AbsCO	−0.77 ± 0.04	−0.78 ± 0.08	−0.71 ± 0.06
L	−0.65 ± 0.06	−0.78 ± 0.08	−0.42 ± 0.11
CO	−0.01 ± 0.11	0.25 ± 0.18	−0.41 ± 0.11

Parameter	$\ln k_f^{mt}$, sec^{-1} 84 proteins	$\ln k_f^{mt\ multi}$, sec^{-1} 26 proteins	$\ln k_f^{mt\ two}$, sec^{-1} 58 proteins
L	−0.64 ± 0.06	−0.58 ± 0.08	−0.61 ± 0.11
$L^{1/2}$	−0.73 ± 0.05	−0.61 ± 0.12	−0.71 ± 0.06
$L^{2/3}$	−0.69 ± 0.06	−0.60 ± 0.13	−0.67 ± 0.07
$\ln L$	−0.77 ± 0.04	−0.64 ± 0.12	−0.76 ± 0.06
AbsCO	−0.77 ± 0.04	−0.65 ± 0.11	−0.78 ± 0.05
CO	0.00 ± 0.11	−0.01 ± 0.20	−0.17 ± 0.13

In 2003 it was demonstrated that the size of a protein mainly determines the rate of folding of multi-state proteins [18], their correlation coefficient being −0.80 for the dependence $\ln k_f^w \sim -L^P$ when $0 \le P \le 1$; at the same time for one-stage proteins (having chains of about equal size) correlation was not found (the correlation coefficient was −0.07 [18]). It should be noted that the case $P = 0$ corresponds to the dependence $\ln k_f^w \sim -\ln L$, so $L^P = \exp(P\ln L) \approx 1 + P\ln L$ for $P \to 0$.

Thus, both from analytical theories and from computer experiments that resolved Levinthal's paradox it was concluded that the size of protein is one but not the only defining factor in the description of protein folding rate. Such dependence helps to understand the reason for the difference in folding rates of proteins with significantly distinct sizes, but cannot explain why proteins of about equal sizes often fold with significantly different rates.

From the point view of prediction of folding rate, all suggested dependences of folding rates on the number of residues in a protein chain about equally agree with the observed protein folding rates: the correlation coefficient is about 65% (see Table 1).

Besides the size of a protein, there are more factors that can additionally affect the protein folding rate. Thus, from the nucleation mechanism of folding it follows that the topology of a transition state – the path of a protein chain in space – should be the same as that of the native structure [19]. This means that the larger is the number of contacts in a protein between residues which are far along the chain the more probable it is that the protein could not avoid loops protruding from the native-like part of the protein in the transition state (Fig. 1), and the folding will be slower. Just this is observed on comparison with experiment: for proteins of about equal size, the logarithm of folding rate decreases with the growth of contact order (*CO*), which is equal to the average distance (in amino acid residues) along the chain between atoms that are in contact in the native structure normalized by the number of amino acid residues in the protein chain [12]. However, topology itself cannot explain the difference in the folding rates for some proteins with one and the same topology (pathway of protein chain in space) (SH3-domains, cold-shock proteins, domains of fibronectin, and proteins with ferredoxin-like fold) [20-24].

Since *CO* does not depend on the protein size, it is not possible to predict the folding rates of all proteins by using this parameter. Integration of *CO* and the number of amino acid resides in proteins in one equation gives "absolute contact order" $AbsCO = CO \cdot L$, including the influence of topology and size of a protein. This parameter better predicts the protein folding rates than *CO* and $L^{2/3}$ taken separately [25].

There is one more consequence of Finkelstein–Badretdinov theory about the influence of the protein shape on its folding rate, which has been established in [26-28]. Namely, under equal conditions a spherical protein independent of the pathway of folding cannot avoid the large area of border between phases (Fig. 1). But an oblong protein has a possibility to choose such a pathway of folding in which the protein folding goes through the small area of border, and consequently through a lower barrier. Therefore, more spherical proteins should fold more slowly. In [27, 28] it has been shown that proteins of class α/β, on average, are more spherical than proteins with other types of packing of secondary structure. It has also been shown that α/β-proteins, again on average, fold more slowly than proteins from other classes of about the same size. We explained this as follows: under equal conditions for more spherical proteins, which include α/β-proteins, the surface border of phases in the transition state is larger and consequently the folding is slower.

Since the prediction of folding rate is in itself valuable, many parameters for prediction of folding rate have been proposed in recent years [18, 25, 29-36]. Some of them are modifications of *CO* [25, 31, 35], and in others the prediction of folding rate uses secondary structure [30, 32] and number of contacts in the native structure [37]. Besides, dependences have been sought using bioinformatics methods [34, 36, 38]. It has been revealed that there is different dependence of folding rate on the amino acid composition [36] and physicochemical properties of amino acids of proteins [34, 38] for proteins with one-stage and multi-stage kinetics of folding in water. However, the bioinformatics methods cannot give a physical explanation of the results obtained.

The above rather contradictory data demonstrate that the theory of protein folding requires further development and subsequent search for factors affecting the rate of folding. In this work the process of protein folding of separate protein molecules and structural properties connected with the folding have been investigated. The dependence of rate of folding on size of a protein and the number of stages in the process have been statistically analyzed.

METHODS OF INVESTIGATION

Experimental data. Since in this work the kinetics of protein folding is studied, we are interested in the following values: $\ln k_f$ is the logarithm of the rate constant of protein chain folding in water, $\ln k_{mt}$ is the logarithm of rate constant of protein chain folding at the point of thermodynamic equilibrium between two protein states that is reached upon addition of denaturant or increasing temperature. Index "two" near the constant means that the kinetics of protein folding is described by kinetics of two states, and in this case one stage is observed and the protein folding is called "one-stage". Index "multi" means that three or more states of the protein chain are required for describing the kinetics, and in this state more than one stage is observed and the protein folding is called "multi-stage". When we consider the whole set of proteins without division into types of folding, we will omit the index of stages for the logarithm of protein folding rate.

A database of proteins with known experimental folding rates has been used (http://phys.protres.ru/resources/compact.html) [28]. Proteins with disulfide bonds and large ligands are not present in this database. The chosen set includes 25 proteins that fold with accumulation of an intermediate state, 56 proteins that fold by an all-or-none mechanism, and two peptides.

Conditions of modeling: point of thermodynamic equilibrium and network of folding/unfolding pathways. Methods based on the modeling of folding of proteins with known spatial structure are fruitful methods for prediction of protein folding rate. The simplified network of pathways of sequential unfolding of a protein is considered in elsewhere [14, 16, 30] (see Fig. 1). As described in [7, 10], each step on the unfolding pathway consists of transition of one chain link from the native spatial structure of the protein to the unfolded state. This link has lost all the non-valence interactions but has acquired the entropy of coil excluding the entropy spent on closing of disordered loops protruding from the remaining part of the globule. All possible transitions form the network of pathways of protein folding/unfolding. The correlation coefficient of calculated (without any adjustable parameters) and experimentally investigated folding rates at the point of thermodynamic equilibrium for the set of 45 proteins is 0.78 [14], but if two short peptides (α-helix and β-hairpin) are deleted from the set the correlation coefficient decreases to 0.56, and for the set of 17 proteins with well-studied folding nuclei it is 0.73 [16]. Calculated folding rates in water correlate with the experimental data at the level of 0.67 for the set of 37 proteins under analysis of a similar model as reported in [15]. Although the neglect of energy of non-native interactions distinctly coarsens the protein folding/unfolding, the success of the above-mentioned theoretical works suggests that the topology characteristics (and sizes) of proteins play a more important role in the folding process than the details of protein structure and interactions occurring in the protein [10, 39, 40].

We have a possibility to analyze the full network of folding/unfolding pathways though presented in a rather rough resolution (the microstate can include no more than two loops; the link includes several amino acid residues) if the method of dynamic programming is used, and we can analyze separate folding pathways without restrictions peculiar to dynamic programming if the Monte Carlo method is used.

In this work, the process of protein folding/unfolding is modeled at the point of thermodynamic equilibrium between native and denatured protein states (that is under conditions when free energies of native and denatured states of protein molecule are equal; further we will call these conditions the point of thermodynamic equilibrium). At the point of thermodynamic equilibrium small proteins fold by a one-stage mechanism ("all-or-none" transition) because in the thermodynamic [41] and kinetic [42, 43] experiments only two states of protein molecule are observed (native and denatured), but intermediate states are not observed to a significant extent. In other words, under these conditions native and denatured states have (at the point of thermodynamic equilibrium by definition) equal free energies and other (including intermediates) states are destabilized. Therefore, by modeling the protein folding and unfolding process at the point of thermodynamic equilibrium we cannot consider misfolding structures (which under other conditions can be peculiar "dead ends", traps that in principle are able to strongly affect the folding and unfolding of a protein molecule). According to the principle of detailed balance [44], pathways of direct and reverse reactions coincide if both reactions occur under the same conditions. Therefore, we can present the process of protein folding and unfolding at the point of thermodynamic equilibrium as a reversible process.

As a basis, we take the three dimensional structure of the protein in the native state from the database of spatial protein structures (PDB) [45]. The process of protein folding/unfolding is modeled as a process of reversible unfolding of their native spatial structure. We consider the network of unfolding pathways where each pathway is presented as a simplified sequential unfolding of the protein (Fig. 1).

Each step on the unfolding pathway represents the removing of one link from the native spatial structure of the protein (the "link" can consist either of one amino acid residue or several residues along the chain without a break). The removed links are assumed to form an indigested coil, i.e. they loose all the non-bonded interactions and gain coil entropy excluding its relatively small part [7] that is spent for closing the disordered loops protruding from the remaining globule (see semi-unfolded structures in Fig. 1). The next simplification is the assumption that the residues remaining in the globule maintain their native position and that the unfolded regions do not fold to another non-native structure. The last and general assumption is that we concentrate our attention on the transition sates, that is on the stabilities (or strictly, instability) of semi-folded structures rather than on the detailed description of the chain movements.

For simplicity of calculations we restrict the number of closed disordered loops protruding from the structure (no more than two loops) and use "links" consisting of not one but of several amino acid residues.

Thus, we obtain the network of folding/unfolding pathways. Further we calculate the free energy of each state in this network. This is done using the method of dynamic programming. For modeling of the process of protein folding by the Monte Carlo method, we start from a fully unfolded state and for each step we try to put the residue in their native position according to the coordinates from the protein data bank (see below the section "Modeling of folding by the Monte Carlo method").

Estimation of free energy. The process of sequential folding/unfolding of native structure of a protein chain including L links is presented in Fig. 1. This chain has fully folded native state I_0, fully unfolded state I_L, and an ensemble of intermediate partly unfolded structures I_ν consisting of ν unfolded chain links and a native-like globular part of $L - \nu$ links ($\nu = 0$ for native state I_0, $\nu = L$ for fully unfolded state I_L, $\nu = 1, ..., L - 1$ for partly unfolded structures). As mentioned above, structures with non-native-like globular parts are not considered.

The free energy of structure I is presented as follows:

$$F(I) = n_I \times \varepsilon - T[\nu_I \times \sigma + \sum_{\text{loop}\in I} S_{\text{loop}}]\,, \qquad (4)$$

where n_I is the number of atom–atom contacts in the native-like part of I (contacts between neighbors along protein chain amino acid residues are not considered because neighbor residues have contacts in the unfolded state); ε is the energy of one atom–atom contact (all contacts are considered to be equal in energy); T is the temperature; ν_I is the number of amino acid residues in the unfolded part of structure I; σ is the entropy difference between the unfolded and the native state of an amino acid residue (for any residue we take $\sigma = 2.3R$ according to the experimental estimation [41], where R is the gas constant); S_{loop} is described by Eq. (6) (see below), the entropy spent to close a disordered loop protruding from the native-like part of structure I (the sum is taken over all closed loops existing in structure I). Atom–atom contacts are calculated from the three-dimensional structure: two non-hydrogen atoms contact if the distance between their centers is not more than 6 Å. Under modeling with accounting for hydrogen atoms, the limiting contact distance was taken smaller: 4 Å for contacts of hydrogen atoms with each other and 5 Å for contacts between hydrogen and non-hydrogen atoms.

All calculations of free energies in this work correspond to the point of equilibrium between native state I_0 and coil I_L. At this point $F(I_0) = F(I_L)$, that is $n_0 \times \varepsilon = T[N \times \sigma]$, where n_0 is the number of contacts in the native structure and N is the total number of protein chain residues.

So, the energy of one internal protein contact ε and temperature T at the point of thermodynamic equilibrium comply with the relation:

$$\varepsilon = -TN\sigma/n_0 \,. \tag{5}$$

Consequently, we can express all free energies in RT units knowing the native structure of protein and the single experimentally determined parameter – the difference in entropy between unfolded and folded states of amino acid residue (σ).

The entropy spent to close a disordered loop protruding from the globule between the still fixed residues k and l is estimated [39] as:

$$S_{\text{loop}} = -\frac{5}{2} \times R \times \ln|k-l| - \frac{\frac{3}{2} \times R \times (r_{\text{kl}}^2 - a^2)}{2 \times A \times a \times |k-l|}\,, \tag{6}$$

where r_{kl} is the distance between the C^α atoms of residues k and l, $a = 3.8$ Å is the distance between the neighbor C^α atoms in the chain, and A is the persistence length for a polypeptide (according to Flory [46] we take $A = 20$ Å).

We obtain the free-energy landscape by computed free energies for each state in the folding/unfolding pathway (Fig. 1), in which we can find "passes" corresponding to the transition states.

RESULTS AND DISCUSSION

Comparison of folding rates for proteins with simple (one stage) and complex (multi-stage) kinetics of folding. To exclude the influence of length (number of amino acid residues), we analyzed data on folding rate within a given size range where the average size of proteins is nearly the same. Therefore, the set of proteins was divided into two groups where the number of proteins is sufficient for statistical analysis: the ranges are 50-100 and 101-151 a.a. Inside each range of length, the protein folding rate was averaged to compare these values for each group of proteins.

The results of averaging are presented in Table 2. The expected result is that the average folding rate for multistate proteins from 101-151 a.a. size range is lower than for proteins with two-state kinetics from the same size range. The unexpected result is that the proteins with multi-state kinetics from size range 50-100 a.a. fold as fast as proteins with two-state kinetics. On average, it turned out that proteins with two-state kinetics from the size range 101-151 a.a. fold faster than proteins from the same group from the size range 50-100 a.a. Another unexpected observation is that among homologous proteins from size range 50-100 a.a., proteins with multi-state kinetics usually fold faster than proteins with simple kinetics of folding (En-HD (homeodomain) vs. C-Myb, hTRF1, and hRAP1; hypF-N (acylphosphatase-like domain) vs. CT AcP (acylphosphatase) and mAcP (muscle acylphosphatase); see below). This result contradicts the general opinion that multi-state folders should fold slower than proteins with simple folding kinetics, and the occurrence of intermediates of folding should decelerate the folding process of protein chains. This opinion probably arose due to the fact that multi-state proteins are usually larger than proteins with a simple mechanism of folding [11]. In the size range of 101-151 a.a., multi-state proteins indeed

Table 2. Average values of radius of cross section and the logarithms of folding rates in water for proteins with "all-or-none" folding mechanism and with accumulation of an intermediate state

Mechanism of folding	"All-or-none"		With accumulation of intermediate state	
Size of range	50-100 a.a.	101-151 a.a.	50-100 a.a.	101-151 a.a.
Number of proteins	36 proteins	7 proteins	9 proteins	8 proteins
Radius of cross section, V_{ASA}/S_{ASA}, Å	3.54 ± 0.03	3.77 ± 0.05	3.75 ± 0.05	4.14 ± 0.07
$\ln(k_f)$ in water, sec^{-1}	4.98 ± 0.57	7.03 ± 1.50	5.12 ± 0.90	1.58 ± 0.84

fold slower, but this is not due to the differences in their length, so we consider the same-length proteins. Another explanation of the fact is that proteins of the considered class are more compact than proteins with simple folding kinetics. This is revealed in that the surface of the boundary between two phases (folded and unfolded) will be larger for more spherical, compact proteins than that for non-spherical proteins (see Table 2) [26-28].

Modeling of protein folding with ferredoxin-like fold. One of the possibilities to study the influence of details of the sequence on the folding process is to consider two nearest proteins with similar topology from the same family. Thus the transition states for four proteins with a ferredoxin-like fold have been characterized and experimentally studied: AcP (acylphosphatase), Ada2h (human activation domain of procarboxypeptidase A2), U1A (spliceosomal protein U1A), and S6 (ribosomal protein S6) (two helices are packed on β-sheet with five or four strands). These proteins have a symmetrical position of secondary structure elements that is destroyed by connection of these elements in the chain. The transition states of proteins Ada2h (human activation domain of procarboxypeptidase A2) and AcP (acylphosphatase) are similar in structure despite the low sequence similarity (13%) and different length of the secondary structure elements [47, 48]. For both proteins the second α-helix and the inside strands are more structured in the transition state than the rest of the protein structure (see Fig. 2). An alternative nucleus, which includes the first α-helix, has been found for protein U1A and another nucleus with both α-helices – for protein S6 [49]. However, the folding rates of Ada2h and AcP differ by three orders of magnitude. The authors explain such a result by the difference in the relative contact order for these proteins [47]. A strong correlation is observed between the relative contact order and the logarithm of folding rate for some proteins with similar topology (HPr, MerP, and U1A) [50].

Modeling of folding by the Monte Carlo method. The process of protein folding is modeled as "travel" of one protein molecule along the network of folding/unfolding pathways [51]. The start is performed from the fully unfolded state; the finish corresponds to the native structure of the protein (i.e. fully folded state). The state with maximal free energy on the folding pathway is considered as the transition state. Each step is modeled in the following way. One residue from all amino acid residues is chosen randomly. Then, if the residue is native it should "unfold", and if it is unfolded it should "fold". The changing of free energy in the case of such an elementary step is calculated. If free energy decreases such an elementary step is accepted; if the free energy grows such an elementary step is accepted with probability exp[$-\Delta F/RT$]

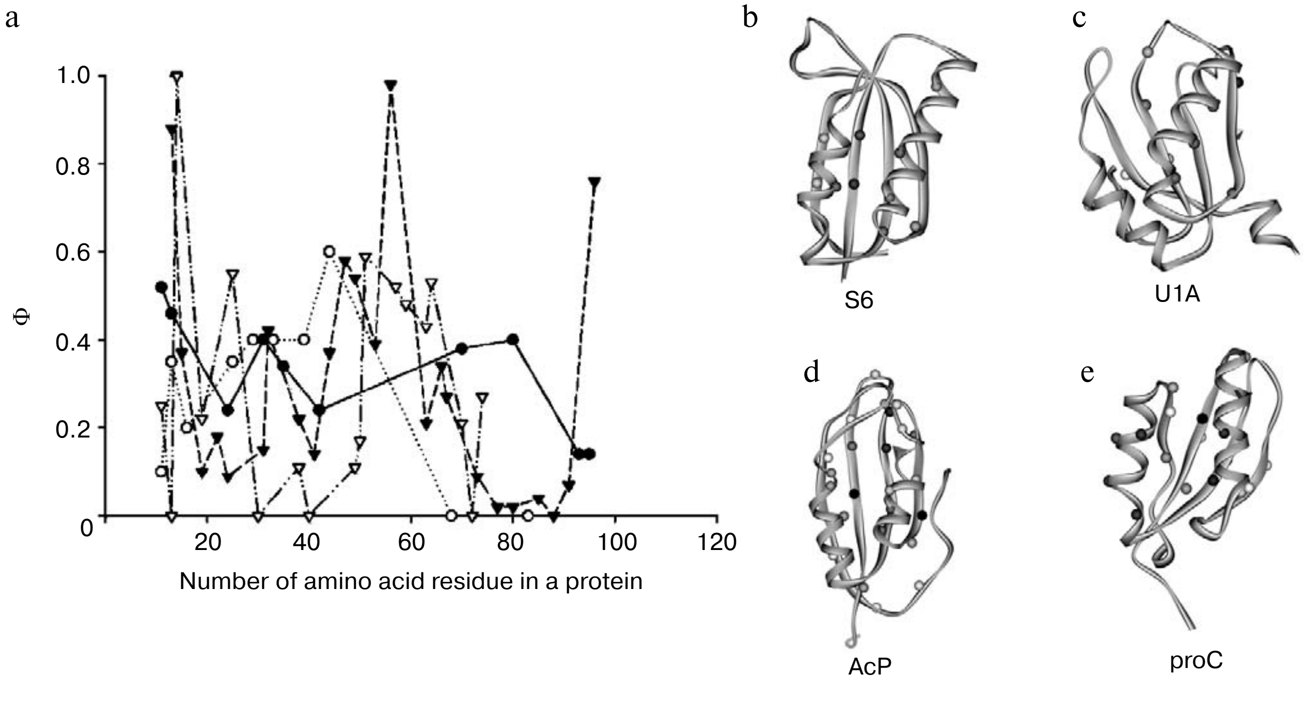

Fig. 2. a) Profiles of Φ_f-values obtained from experiments for proteins with a ferredoxin-like fold. Investigated residues are shown by filled circles for U1A (spliceosomal protein), by open circles for S6 (ribosomal protein), by filled triangles for AcP (acylphosphatase), and open triangles for proC (procarboxypeptidase A2). According to the model of native-like folding nucleus [35], $\Phi_f = 1$ means that the given residue is included in the folding nucleus, and $\Phi_f = 0$ means that it is not involved in the folding nucleus. Interpretation of values $\Phi_f \approx 0.5$ is ambiguous: in this case the residue can be on the surface of the nucleus or there are several folding pathways (that means several nuclei) where the residue is involved in one of the alternative nuclei. b-e) Schemes of three-dimensional structures of these proteins drawn according to their Φ_f values of the amino acid residues from white ($\Phi_f = 0$) to black ($\Phi_f = 1$). Triangles on the structure correspond to residues with $\Phi_f < 0$ and $\Phi_f > 1$.

(standard criterion of Metropolis [52]). For each protein, 50 runs were performed with 10^8 steps.

To understand to what extent the model reflects reality, it is necessary to obtain results based on the model that can be compared with the experimental data. In this work we will make a comparison with the rate of protein molecule folding.

The Monte Carlo method allows us to obtain the folding time of a separate molecule determined in Monte Carlo steps (see Fig. 3). As theoretically calculated time of protein folding we use time ($t_{1/2}$), for which half of the molecules fold (that is 25 of the 50 runs resulted in the native structure formation [53]).

The typical folding kinetics for six proteins with ferredoxin-like fold obtained using the Monte Carlo method is presented in Figs. 3 and 4. The protein molecule starts from the fully unfolded state, and for a rather long time stays near the unfolded state; then it overcomes the free energy barrier and quickly folds completely. One can see for protein HypF-N (1gxt) that until 63,174,273 Monte Carlo steps the molecule stays near the unfolded state, then quickly in 600 steps overcomes the general barrier. After that the molecule stays for some time near the native state (90% of amino acid residues are in the native position) and then comes to the native state. The same concerns the other proteins as well (see Table 3). It should be mentioned that for both proteins near native surroundings (1o6x, 1urn) there are states that are more favorable in energy than the native state. Usually such states correspond to structures in which several residues do not fold, and these residues have a small number of contacts in the native structure.

Calculated (at the point of thermodynamic equilibrium) values $t_{1/2}$ for five proteins (excluding human activation domain of procarboxypeptidase A2, the length of which is 71 a.a., which is shorter than the other five proteins, see Table 3) have a good correlation coefficient with the experimentally measured time of folding at the point of thermodynamic equilibrium: the correlation coefficient between the logarithm of folding rate and the logarithm of number of Monte Carlo steps is –0.98, and between the folding rate and the radius of cross section it is –0.85 (see Table 3). This illustrates that for the given protein family the size of cross section, which determines the border between the ordered and disordered parts of protein and thus the value of general folding barrier, explains the difference in the folding time for the given set of proteins with ferredoxin-like fold.

Estimation of protein folding rate using calculated free energy of transition state. In the case when the analysis of the network of folding/unfolding pathways has been made using dynamic programming, we calculated the effective value of the free-energy barrier:

$$F(TS) = -RT\ln\sum_{I\in TS}\exp[-\frac{F(I)-F(I_L)}{RT}]\,, \qquad (7)$$

where I_L is the fully unfolded protein chain, TS is the ensemble of transition states. The calculated effective values of free energy barrier correlate with the logarithm of experimentally measured folding times at the point of thermodynamic equilibrium: the correlation coefficient is –0.97 (see Table 3).

On the other hand, the values of free energy barriers correlate somewhat worse with the logarithm of experimentally measured time of folding in water, far from the point of thermodynamic equilibrium: in this case the correlation coefficient of the theory and experiment is –0.67.

Table 3. General characteristics of folding (experimentally obtained in water and at the point of thermodynamic equilibrium constants of protein folding rates, calculated free-energy values of barriers on the folding pathway, and number of Monte Carlo steps) for proteins with ferredoxin-like fold

Name of protein, PDB code	Number of amino acid residues in structure	Radius of cross section, V/S, Å	$\ln(k_f)$ in water, sec^{-1}	$\ln k_f^{mt}$, sec^{-1}	$t_{1/2}$, number of Monte Carlo steps	F/RT, free energy value of barrier calculated using dynamic programming
HypF-N, 1gxt	88	3.88	4.4	–0.8	24 874 862	21.07
Common-type acylphosphatase, 2acy	98	3.87	0.8	–4.4	65 114 308	24.83
Muscle acylphosphatase, 1aps	98	3.96	–1.6	–7.3	169 227 106	25.39
Activation domain of human procarboxypeptidase A2, 1o6x	71	3.27	6.8	1.5	384 761	18.37
Spliceosomal protein U1A, 1urn	96	3.84	4.6	–0.4	22 656 530	23.086
Ribosomal protein S6, 1ris	96	3.90	6.1	–3.9	84 634 031	25.53

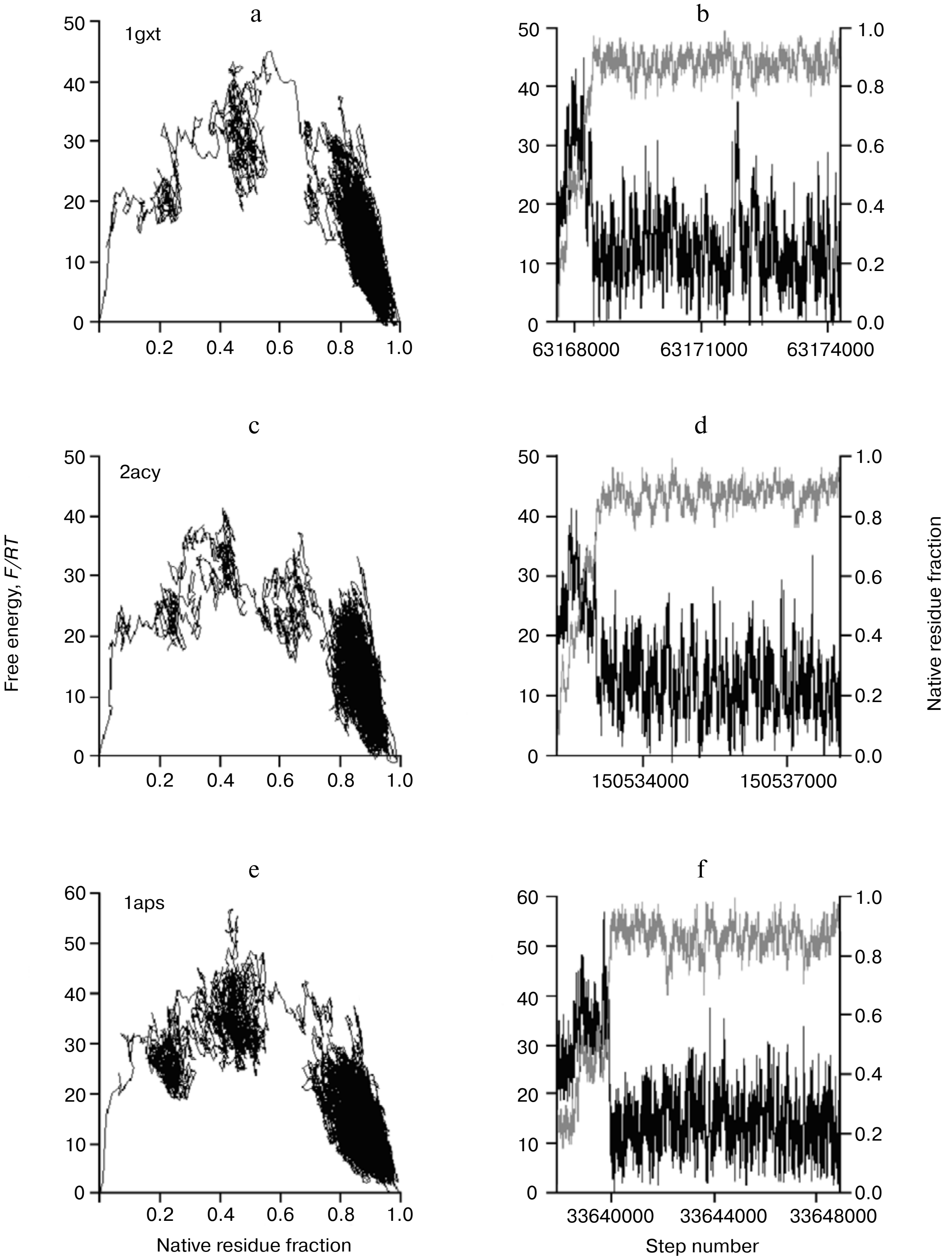

Fig. 3. a, c, e) Dependences of free energy on the fraction of amino acid residues fixed in the native position for one of the chosen folding trajectories for proteins with a ferredoxin-like fold. b, d, f) The fragment of the same folding trajectory. The black curve is the dependence of free energy on the number of the elementary step in the folding process by the Monte Carlo method. The gray curve is the dependence of the number of native amino acid residues on time (number of Monte Carlo steps).

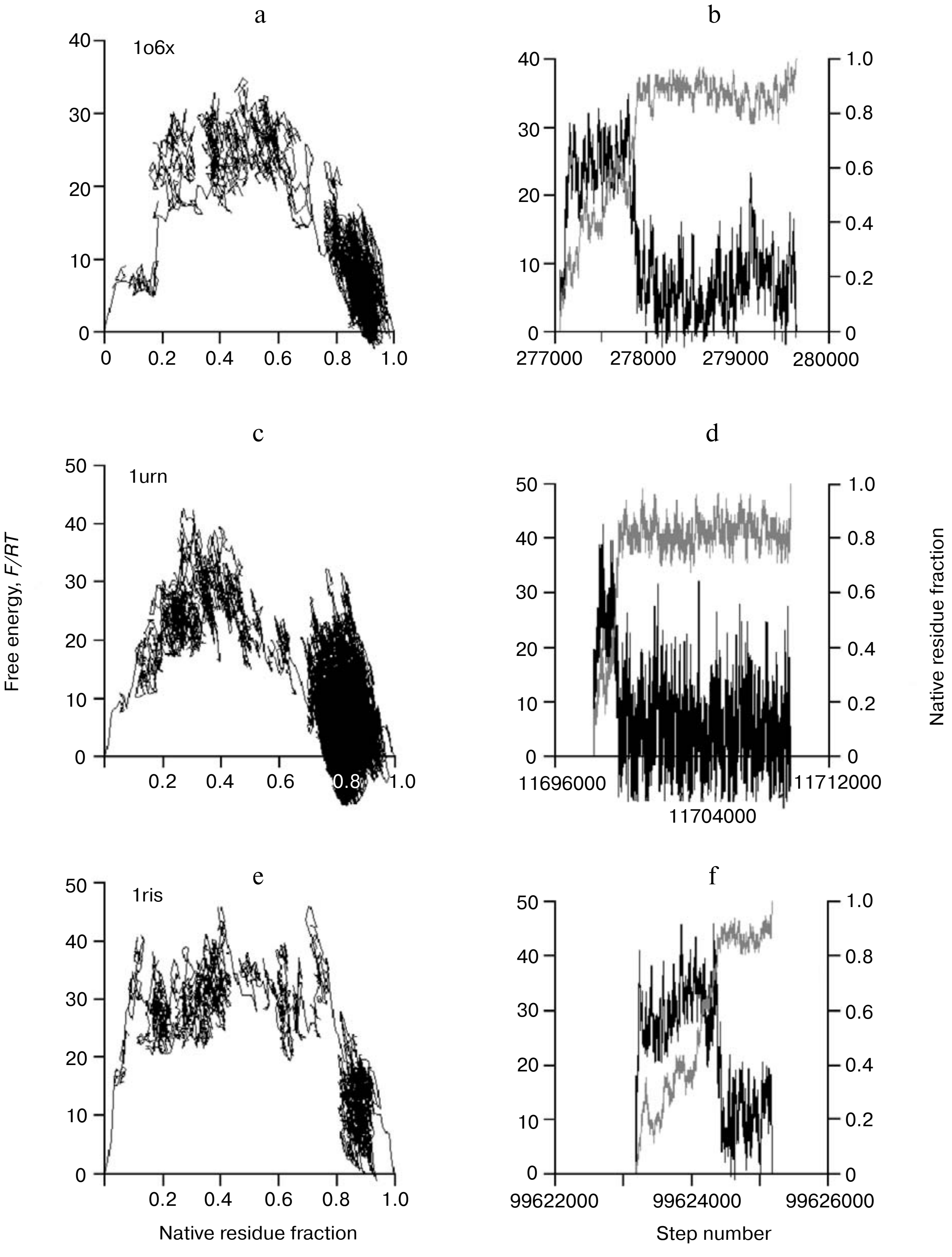

Fig. 4. a, c, e) Dependences of free energy on the fraction of amino acid residues fixed in the native position for one of the chosen folding trajectories for proteins with a ferredoxin-like fold. b, d, f) The fragment of the same folding trajectory. The black curve is the dependence of free energy on the number of elementary step in the folding process by the Monte Carlo method. The gray curve is the dependence of the number of native amino acid residues on time (number of Monte Carlo steps).

The decrease in the correlation in the last case is not surprising since the calculated values of free energy barriers correspond to the point of thermodynamic equilibrium.

In this work, statistical analysis of protein folding rates with known experimental data has been done. It has been shown that proteins with multi-state kinetics from the size range 50-100 a.a. fold as fast as proteins with two-state kinetics. In their turn, these proteins fold not faster than proteins with the same mechanism of folding from the size range 101-151 a.a.

In the group of homologous proteins from the size range 50-100 a.a., proteins with multi-state kinetics usually fold faster than proteins with simple folding kinetics. Modeling of folding of six proteins of similar size with a ferredoxin-like fold using Monte Carlo and dynamic programming methods demonstrates that indeed protein folding with accumulation of an intermediate state makes this an order of magnitude faster than their structural homologs.

I am grateful to A. V. Glyakina and S. O. Garbuzynskiy for their assistance in preparation of some figures and to N. S. Bogatyreva for assistance in calculations.

This work was supported by the programs "Molecular and Cellular Biology" (01200959110) and "Fundamental Sciences to Medicine", by the Russian Foundation for Basic Research (08-04-00561) and the Russian Science Support Foundation, and by the grant from the Federal Agency for Science and Innovations (No. 02.740.11.0295).

REFERENCES

1. Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Rothlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., 3rd, Hilvert, D., Houk, K. N., Stoddard, B. L., and Baker, D. (2008) *Science*, **319**, 1387-1391.
2. Rothlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S., and Baker, D. (2008) *Nature*, **453**, 190-195.
3. Chiti, F., Webster, P., Taddei, N., Clark, A., Stefani, M., Ramponi, G., and Dobson, C. M. (1999) *Proc. Natl. Acad. Sci. USA*, **96**, 3590-3594.
4. Fandrich, M., Fletcher, M. A., and Dobson, C. M. (2001) *Nature*, **410**, 165-166.
5. Thirumalai, D. (1995) *J. Phys. Orsay Fr.*, **5**, 1457-1467.
6. Gutin, A. M., Abkevich, V. I., and Shakhnovich, E. I. (1996) *Phys. Rev. Lett.*, **77**, 5433-5456.
7. Finkelstein, A. V., and Badretdinov, A. Ya. (1997) *Mol. Biol.* (Moscow), **31**, 391-398.
8. Koga, N., and Takada, S. (2001) *J. Mol. Biol.*, **313**, 171-180.
9. Finkelstein, A. V., and Galzitskaya, O. V. (2004) *Phys. Life Rev.*, **1**, 23-56.
10. Finkelstein, A. V., and Badretdinov, A. Ya. (1997) *Fold. Des.*, **2**, 115-121.
11. Jackson, S. E. (1998) *Fold. Des.*, **3**, R81-R91.
12. Plaxco, K. W., Simons, K. W., and Baker, D. (1998) *J. Mol. Biol.*, **277**, 985-994.
13. Munoz, V., and Eaton, W. A. (1999) *Proc. Natl. Acad. Sci. USA*, **96**, 11311-11316.
14. Ivankov, D. N., and Finkelstein, A. V. (2001) *Biochemistry*, **40**, 9957-9961.
15. Alm, E., Morozov, A. V., Kortemme, T., and Baker, D. (2002) *J. Mol. Biol.*, **322**, 463-476.
16. Garbuzynskiy, S. O., Finkelstein, A. V., and Galzitskaya, O. V. (2004) *J. Mol. Biol.*, **336**, 509-525.
17. Naganathan, A. N., and Munoz, V. (2005) *J. Am. Chem. Soc.*, **127**, 480-481.
18. Galzitskaya, O. V., Garbuzynskiy, S. O., Ivankov, D. N., and Finkelstein, A. V. (2003) *Proteins*, **51**, 162-166.
19. Fersht, A. R. (1997) *Curr. Opin. Struct. Biol.*, **7**, 3-9.
20. Guijarro, J. I., Morton, C. J., Plaxco, K. W., Campbell, I. D., and Dobson, C. M. (1998) *J. Mol. Biol.*, **276**, 657-667.
21. Plaxco, K. W., Guijarro, J. I., Morton, C. J., Pitkeathly, M., Campbell, I. D., and Dobson, C. M. (1998) *Biochemistry*, **37**, 2529-2537.
22. Perl, D., Welker, Ch., Schindler, Th., Schroder, K., Marahiel, M. A., Jaenicke, R., and Schmid, F. X. (1998) *Nature Struct. Biol.*, **5**, 229-235.
23. Van Nuland, N. A. J., Chiti, F., Taddei, N., Raugei, G., Ramponi, G., and Dobson, C. M. (1998) *J. Mol. Biol.*, **283**, 883-891.
24. Zerovnik, E., Virden, R., Jerala, R., Turk, V., and Waltho, J. P. (1998) *Proteins*, **32**, 296-303.
25. Ivankov, D. N., Garbuzynskiy, S. O., Alm, E., Plaxco, K. W., Baker, D., and Finkelstein, A. V. (2003) *Protein Sci.*, **12**, 2057-2062.
26. Galzitskaya, O. V., Bogatyreva, N. S., and Ivankov, D. N. (2008) *J. Bioinform. Comput. Biol.*, **6**, 667-680.
27. Galzitskaya, O. V., Danielle, C., Reifsnyder, D. C., Bogatyreva, N. S., Ivankov, D. N., and Garbuzynskiy, S. O. (2008) *Proteins*, **70**, 329-332.
28. Ivankov, D. N., Bogatyreva, N. S., Lobanov, M. Yu., and Galzitskaya, O. V. (2009) *PLoS ONE*, **4**, e6476.
29. Punta, M., and Rost, B. (2005) *J. Mol. Biol.*, **348**, 507-512.
30. Ivankov, D. N., and Finkelstein, A. V. (2004) *Proc. Natl. Acad. Sci. USA*, **101**, 8942-8944.
31. Zhou, H., and Zhou, Y. (2002) *Biophys. J.*, **82**, 458-463.
32. Gong, H., Isom, D. G., Srinivasan, R., and Rose, G. D. (2003) *J. Mol. Biol.*, **327**, 1149-1154.
33. Capriotti, E., and Casadio, R. (2007) *Bioinformatics*, **23**, 385-386.
34. Gromiha, M. M., Thangakani, A. M., and Selvaraj, S. (2006) *Nucleic Acids Res.*, **34**, W70-W74.
35. Gromiha, M. M., and Selvaraj, S. (2001) *J. Mol. Biol.*, **310**, 27-32.
36. Ma, B. G., Chen, L. L., and Zhang, H. Y. (2007) *J. Mol. Biol.*, **370**, 439-448.
37. Makarov, D. E., Keller, C. A., Plaxco, K. W., and Metiu, H. (2002) *Proc. Natl. Acad. Sci. USA*, **99**, 3535-3539.
38. Gromiha, M. M. (2005) *J. Chem. Inf. Model*, **45**, 494-501.
39. Galzitskaya, O. V., and Finkelstein, A. V. (1999) *Proc. Natl. Acad. Sci. USA*, **96**, 11299-11304.
40. Baker, D. (2000) *Nature*, **405**, 39-42.
41. Privalov, P. L. (1979) *Adv. Protein Chem.*, **33**, 167.

42. Fersht, A. R. (1995) *Curr. Opin. Struct. Biol.*, **5**, 79-84.
43. Fersht, A. R. (1997) *Curr. Opin. Struct. Biol.*, **7**, 3-9.
44. Landsberg, P. T. (1971) *Problems in Thermodynamics and Statistical Physics*, PION, London.
45. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rogers, J. R., et al. (1997) *Eur. J. Biochem.*, **80**, 319-324.
46. Flory, P. J. (1969) *Statistical Mechanics of Chain Molecules*, Interscience, New York.
47. Chiti, F., Taddei, N., White, P., Bucciantini, M., Magherini, F., Stefani, M., and Dobson, C. (1999) *Nature Struct. Biol.*, **6**, 1005-1009.
48. Taddei, N., Chiti, F., Fiaschi, T., Bucciantini, M., Capanni, C., Stefani, M., Serrano, L., Dobson, C. M., and Ramponi, G. (2000) *J. Mol. Biol.*, **300**, 633-647.
49. Ternstrom, T., Mayor, U., Akke, M., and Oliveberg, M. (1999) *Proc. Natl. Acad. Sci. USA*, **96**, 14854-14859.
50. Oliveberg, M. (2001) *Curr. Opin. Struct. Biol.*, **11**, 94-100.
51. Galzitskaya, O. V., Surin, A. K., and Nakamura, H. (2000) *Protein Sci.*, **9**, 580-586.
52. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953) *J. Chem. Phys.*, **21**, 1087-1092.
53. Galzitskaya, O. V., and Finkelstein, A. V. (1995) *Protein Eng.*, **8**, 883-892.