

Review

Influence of Conformational Entropy on the Protein Folding Rate

Oxana V. Galzitskaya

Institute of Protein Research, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia; E-Mail: ogalzit@vega.protres.ru

Received: 30 December 2009; in revised form: 15 March 2010 / Accepted: 30 March 2010 /

Published: 16 April 2010

Abstract: One of the most important questions in molecular biology is what determines folding pathways: native structure or protein sequence. There are many proteins that have similar structures but very different sequences, and a relevant question is whether such proteins have similar or different folding mechanisms. To explain the differences in folding rates of various proteins, the search for the factors affecting the protein folding process goes on. Here, based on known experimental data, and using theoretical modeling of protein folding based on a capillarity model, we demonstrate that the relation between the average conformational entropy and the average energy of contacts per residue, that is the entropy capacity, will determine the possibility of the given chain to fold to a particular topology. The difference in the folding rate for proteins sharing more ball-like and less ball-like folds is the result of differences in the conformational entropy due to a larger surface of the boundary between folded and unfolded phases in the transition state for proteins with a more ball-like fold. The result is in agreement with the experimental folding rates for 67 proteins. Proteins with high or low side chain entropy would have extended unfolded regions and would require some additional agents for complete folding. Such proteins are common in nature, and their structural properties are of biological importance.

Keywords: native topology; conformational entropy; protein stability; folding rate; energy of residue-residue contacts

1. Introduction

Understanding the folding of biological macromolecules is one of the main problems of molecular biology and biophysics. Protein molecules have an enormous number of possible conformations, but this does not hinder finding their unique stable three-dimensional (3D) structure within much less time than required for an exhaustive sorting of all their conformations. Understanding the reason for fast folding of 3D structures of biological macromolecules is especially important for the *de-novo* design of proteins. For instance, for the *de-novo* design of a protein it is necessary to know what features of its primary structure define the stability of its 3D structure and what features of the sequences provide for its fast folding.

Some general trends and correlations are beginning to emerge between the structural, thermodynamic and kinetic properties of proteins [1–7]. There is an enormous diversity in the folding behavior for small proteins that fold with simple two-state kinetics, as well as for large proteins that fold with multi-state kinetics.

The apparent lack of a relationship between stabilities and folding rates of topologically diverse proteins indicates that the topology may be a critical determinant of folding kinetics [3]. But the topology itself cannot explain the differences in the refolding rates for some proteins sharing the same fold (SH3 domains, cold shock proteins, fibronectin domains, proteins belonging to the ferredoxin fold) [8–12].

On the other hand, a number of basic correlations between the protein size and folding rate have been suggested [1,13–15]. All of them point out that, as might be expected, the folding rate decreases with the protein size, but suggest different scaling laws for this decrease. It should be noted that simulations of folding of simplified off-lattice protein models [16] suggest that the logarithm of protein folding rate scales as $-L^{0.61 \pm 0.18}$, which is in accordance with both $-L^{2/3}$, following the theory of Finkelstein and Badretdinov [1], and $-L^{1/2}$, following the theory of Thirumalai [13]. The current statistical analysis of protein folding data shows that all the suggested scalings, from $\log(L)$ [14] to $-L^{1/2}$ and $-L^{2/3}$ correlate with the observed folding rates almost equally: the correlation between folding rates and protein sizes is about 70% (see Table 1) [17,18].

The relative contact order (*CO*) (equal to average chain separation of contacting residues in the native fold, divided by the chain length) was invented to compare differences in chain folds (rather than in size) between two-state folding proteins of different lengths [3]. This parameter is small for proteins stabilized mainly by local interactions and large when residues in a protein interact frequently with partners far away in the sequence. The latter should lead to slower folding [3,19]. Recently, however, it has been found that the relative contact order (*CO*) [3], computed from the shape of protein chain structure rather than from the lengths of the chains, fails to predict the protein folding rates (see Table 1) [18]. At the same time, it was found that the “absolute contact order” that combines *CO* and the chain length *L* in a form $Abs_CO = CO \times L$ (and scales as $\sim L^{0.7}$, on the average) is able to predict rather accurately folding rates for both two- and multi-state folding proteins, as well as for short peptides: the correlation reaches 77% for all 84 proteins and peptides (see Table 1) [18,20]. Moreover, it has been noted that in order to predict protein folding rates, any parameter should correlate well with the protein chain length [18]. A very illustrative example of this is observed with relative contact order (the normalized parameter, which has poor correlation with the logarithm of

protein folding rates, see Table 1) and absolute contact order, which includes both the size of protein and the average length of loops and correlates well with the logarithm of protein folding rates [18].

Table 1. Correlation coefficients between logarithms of folding rates in mid-transition and size of protein structure.

	$\ln k_f^{mt}$ 84 proteins	$\ln k_f^{mt\ multi}$ 26 proteins	$\ln k_f^{mt\ two}$ 58 proteins
Parameters of protein size and average size of protein loop			
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
Relative contact order is the parameter of an average size of the protein loop normalized to exclude dependence on protein size.			
CO	0.00 ± 0.11	-0.01 ± 0.20	-0.17 ± 0.13

It has been shown that folding rates of small single-domain proteins that fold through simple two-state kinetics correlate with the number of native contacts and can be estimated from contact predictions [21,22]. To predict the protein folding rate using *Abs_CO*, one should have its 3D structure. The question arises: is it possible to predict the protein folding rate from the sequence alone?

One important question in molecular biology is what determines folding pathways: native structure or protein sequence. There are many proteins that have similar structures but very different sequences, and a relevant question is whether such proteins have similar or different folding mechanisms. To explain the differences in folding rates of various proteins, the theory of protein folding rate should be developed further. Therefore, the search for the factors affecting the protein folding process goes on.

The empirical dependence of the folding rate on some features of amino acid sequences (secondary structure content in the 3D structures, hydrophobic content) has been published for two-state proteins [23–25]. Having considered 69 proteins and peptides Naganathan and Munoz obtained 74% correlation between experimental folding times and the square root of the number of residues [26]. As the authors underlined, an important practical implication of the finding is that one can predict folding times with a precision of ~ 1.1 time decades by just knowing the size of a protein [26].

Punta and Rost have calculated in a recent publication that if one considers 27 two-state proteins having a significant content of beta strands, the correlation between the effective length and the folding rate becomes insignificant (13% in a jack-knife experiment) [22]. Such a result demonstrates that the theory of protein folding rate requires further development. Therefore, the search for the factors affecting the protein folding process continues.

The capillarity model [1] gave rise to the hypothesis that protein folding rates are determined by the average "entropy capacity" (the entropy capacity for the given protein structure is defined as the average conformational entropy per residue divided by the average contact energy per residue; thus,

this value is, in a sense, reciprocal to the expected melting temperature) [27]. Consideration of the compactness specifically addresses the issue of why some proteins fold more rapidly than others. Statistical analysis demonstrates that four main structural classes [28] of proteins (all- α , all- β , α/β , $\alpha+\beta$) differ from one another in a statistically significant manner with respect to the number of rotatable angles ϕ , ψ and χ and the average number of contacts per residue [29]. On the whole, it has been shown that among proteins of the same size, α/β proteins have, on average, a greater number of contacts per residue due to their more compact (more “spherical,” since a sphere is the most compact geometrical body having the minimal accessible surface area in comparison with other geometrical bodies of equal volumes) structure, rather than to tighter packing [30]. For 75 proteins for which both folding rates and tertiary structures are known, α -helical proteins have on average the fastest folding kinetics and the smallest number of contacts per residue (they are less compact than others), whereas α/β proteins have on average the slowest folding kinetics and the largest number of contacts (they are more compact than others) [30]. One explanation is that the expected surface of the boundary between folded and unfolded phases in the transition state [18,31] for a more spherical protein is larger than that for a non-spherical protein. Thus, the fact that α/β proteins are more spherical explains both the more average number of contacts per residue [30] and the slower folding kinetics.

The more spherical the protein is, the slower folding it exhibits. Proteins with multi-state kinetics, on average, are more spherical than proteins with two-state kinetics [18]. The obtained result suggests that shape properties are important determinants of the folding kinetics. Rapidly and slowly folding proteins have additional structural diversities that contribute to their differing folding rates. From our previous analysis, we can conclude that the barrier height for folding of large proteins is defined by the size of the boundary between folded and unfolded phases in the transition state. This boundary is larger for a ball-like protein globule than for an elongated one [18,31].

In this work, based on the known experimental data and using theoretical modeling of protein folding process, we demonstrate that an optimal relationship exists between the average conformational entropy and the average energy of contacts per residue, *i.e.*, an entropy capacity, for fast protein folding. Consideration of the entropy capacity specifically addresses the issue of why some proteins fold more rapidly than others. Statistical analysis demonstrates that four main structural classes of proteins (all- α , all- β , α/β , $\alpha+\beta$) differ from one another in a statistically significant manner with respect to the number of rotatable angles ϕ , ψ and χ and the average number of contacts per residue. These parameters offer an explanation to the vast range of experimentally observed folding rates. The obtained result suggests that structural and sequence properties are important determinants of the folding rate.

2. Results and Discussion

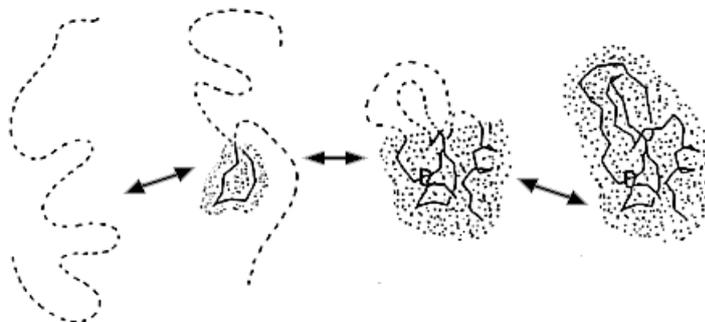
2.1. Entropy Capacity for Proteins with Given Topology

The formation of a sufficient number of interactions between amino acid residues is necessary to compensate for the loss of the conformational entropy during protein folding. Therefore structural uniqueness of the native state is a result of the balance between the conformational entropy and the energy of interactions between residues. Taking these phenomena into account we consider the entropy capacity for a given protein chain with a chosen topology [27]. This parameter is the relation

between conformational entropy and energy of residue-residue interactions. We will show that the relationship between these values determines the ability of the given chain to fold to a particular topology and the rate of the process.

To demonstrate the importance of the balance between the conformational entropy and energy of residue-residue interactions and its influence on the folding rate, we consider a theoretical model of protein folding proceeding from the analytical theory of this process [1,15]. In this model, the folding and unfolding pathways are treated as a sequential insertion of residues from the unfolded state to their native positions according to the 3D native structure or removal of residues from the native position to the coil, respectively, so that the pathways of folding and unfolding would coincide at the point of equilibrium between the native and unfolded states according to the principle of detailed balance (Figure 1). The removed (inserted) residues are assumed to lose (gain) all the non-bonded interactions and gain (lose) the coil entropy, except that spent to close the disordered loops protruding from the remaining globule. A general assumption of this model is that the residues remaining in the globule keep their native position and the unfolded regions do not fold to another, non-native globule. Thus, we neglect non-native interactions; this makes the considered model similar to that of Gō [32].

Figure 1. Sequential folding (viewed from left to right) and unfolding (viewed from right to left) pathways. The folded part (dotted) is native-like. The bold line shows the backbone fixed in this part; the fixed side chains are not shown for the sake of simplicity (the volume that they occupy is dotted). The dashed line shows the unfolded chain.



2.2. Optimal Value of Entropy Capacity for Fast Protein Folding

First-order phase transitions proceed by a nucleation mechanism in which a small droplet of the ordered phase is formed within the metastable disordered phase [33]. A similar mechanism could apply to the folding of a protein with the ordered phase identified as a native state and the disordered state identified as an unfolded phase [34]. A number of recent experimental and theoretical studies suggest that a crucial step of folding is the formation of a native-like folding nucleus [1,18,35–40]. We can estimate the changing of free energy during the protein folding by considering the capillarity model [1,41]. From such a consideration, a general question arises: is there a minimal value of the free energy barrier for the protein folding? To answer this question, let us consider several general equations from the capillarity model where

$$G_N = N\varepsilon - \mu N^{2/3} \varepsilon \quad (1)$$

is the free energy of the native state;

$$G_0 = -TNs \quad (2)$$

is the free energy of the unfolded state;

$$G_n = n\varepsilon - \mu n^{2/3}\varepsilon - T(N-n)s \quad (3)$$

is the free energy of the intermediate state with n fixed residues in the native topology. Here, N is the total number of residues in the protein molecule; ε is the average contact energy per residue; $\mu n^{2/3}$ takes into account the surface residues having less interactions than the internal ones, where $\mu = 1.5$ for a ball-like body [1]; s is the average conformational entropy per residue for an unfolded chain, and T is the temperature.

$$\Delta G_n = G_n - G_0 = n\varepsilon - \mu n^{2/3}\varepsilon + Tns \quad (4)$$

is the free energy difference between intermediate and unfolded states. We can simplify this expression using $\xi = -\varepsilon/RT > 0$ and $\sigma = s/R > 0$. Then we have:

$$g_n = \Delta G_n / RT = -n\xi + \mu n^{2/3}\xi + n\sigma \quad (5)$$

The transition state on this pathway is $g^\# = \max\{g_n\}$. To find the transition state (the maximum), we take the derivative of Equation 5 with respect to n : $n_0 = \left(\frac{2}{3} \times \frac{\mu}{1 - \sigma/\xi}\right)^3$. If we substitute the value of n_0 in Equation 5 and introduce the “entropy capacity” $C = \sigma/\xi$ ($0 < C < 1$) we obtain the equation

$$g^\#(C) = \frac{4}{27} \times \frac{\mu^3 \sigma}{C(1-C)^2} \quad (6)$$

The question about the minimal value of a free energy barrier is reformulated from Equation 6 as follows: at what value of ξ at given σ would the value of $g^\#$ be the minimal? If we try to find the optimal values of C ($0 < C < 1$) from this equation, we obtain $C = 1$ (this value is not applicable) and $C = 1/3$. Namely, with this relationship between the energy and entropy we will have the minimal value of the barrier on the pathway of folding in our model. If we take $C_{\text{opt}} = 1/3$ in Equation 6, we obtain $g_{\text{opt}} = \mu^3 \sigma$.

This equation means that for proteins with a similar topology, the difference in the folding rates will be determined by the difference in the average conformational entropy between native and unfolded states. This result is in agreement with our previous result where we demonstrated that there exists an optimal value of side-chain entropy for fast protein folding of proteins sharing the same fold [27].

At the point of mid-transition we have:

$$-N(\xi - \sigma) + \mu N^{2/3}\xi = 0 \quad (7)$$

and $C_{\text{eq}} = \sigma/\xi = 1 - \mu N^{-1/3}$. If $N = 125$ and $\mu = 1.5$, we obtain $C_{\text{eq}} = 0.7$ and

$$g_{\text{eq}}^\# = \frac{4}{27} \times \frac{\mu\sigma N^{2/3}}{1 - \mu N^{-1/3}} \quad (8)$$

It should be mentioned that such a relationship between entropy and energy (1/3) is possible only at very low temperatures (if we consider that the relationship between entropy and energy does not change with changing the temperature, we have $C = (s/-\varepsilon)T$ and $C_{\text{eq}}/C_{\text{opt}} = T_{\text{eq}}/T_{\text{opt}} = 0.7 \times 3 = 2.1$).

To answer the question about the changing of the free energy barrier with two major parameters, ξ and σ , let us write the changing of the free energy barrier in partial derivatives in ξ and σ :

$$dg^{\#} = g^{\#} \times \frac{1}{\left(1 - \frac{\sigma}{\xi}\right)} \left[\left(2 \frac{\sigma}{\xi}\right) \frac{d\sigma}{\sigma} + \left(1 - 3 \frac{\sigma}{\xi}\right) \frac{d\xi}{\xi} \right] \quad (9)$$

At the point of equilibrium we have $G_N - G_0 = 0$, see Equation 7 and $\sigma/\xi = 1 - \mu N^{-1/3}$. Putting this equation in Equation 9 and neglecting $\mu N^{-1/3}$ (which is not large in comparison with 1 at large N , where N is the size of protein domain in the range of 100–300 amino acid residues) we obtain:

$$dg^{\#} = \left(\frac{g^{\#}}{\mu} N^{1/3}\right) \times \left[2 \frac{d\sigma}{\sigma} - 2 \frac{d\xi}{\xi} \right] = \left(2 \frac{g^{\#}}{\mu} N^{1/3}\right) \times \frac{\xi}{\sigma} d \frac{\sigma}{\xi} \quad (10)$$

Introducing the entropy capacity $C = \sigma/\xi$ we obtain the dependence of the changing of free-energy barrier on the changing of entropy capacity at an equal value of N :

$$dg^{\#} = \left(2 \frac{g^{\#}}{\mu} N^{1/3}\right) \times d \ln C \quad (11)$$

Further, we try to demonstrate using the statistical and experimental data of the protein folding rate that each class of proteins has its own class-specific average number of contacts and average conformational entropy per residue, and these class-specific features determine the folding rates.

One of the main determinants of the protein folding rate is the number of degrees of freedom that are to be fixed during the rate-limiting step of folding. Therefore, the difference in the folding rate for proteins sharing the spherical and less spherical fold is the result of differences in the conformational entropy due to a larger surface of the boundary between folded and unfolded phases in the transition state for a protein with a spherical fold (see Equation 11, μ is larger for a non ball-like body, which causes diminishing of the free energy barrier). Moreover, the unfolded closed loops protruding from the folded part (the nucleus) cause an additional surface tension.

2.3. Statistical Analysis of Average Conformational Entropy and Average Number of Contacts per Residue for Different Classes of Proteins

To examine whether there is a significant difference between the mean values of a number of contacts and conformational entropy for four general structural classes, we analyzed 1133 all- α proteins from class *a*, 1644 all- β proteins from class *b*, 1617 α/β proteins from class *c*, and 1435 $\alpha+\beta$ proteins from class *d* (according to the SCOP classification [28]).

Calculations of the average number of contacts per residue in proteins were done separately for each class (see Figure 2). In our case, two residues have a contact if any pair of their heavy atoms is at a distance of less than 8.0 Å from each other. The average number of contacts per residue is calculated from the 3D structure of a protein as a sum of the number of contacts of all residues, divided by the number of residues in the protein. The value of the average conformational entropy, ν , was calculated as a summation of the individual number of rotatable angles ϕ , ψ and χ for each residue (see Table 2) over its complete sequence, normalized by the number of residues in the protein. In Table 3 and

Figure 2, one can see that each class of proteins has its own class-specific average number of contacts and the average conformational entropy. The *c* class has the largest number of contacts.

Table 2. Number of rotatable angles for amino acid residues of 20 types.

Type of residue	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE
Degrees of freedom	2	6	4	4	3	5	5	3*	4	4
Type of residue	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
Degrees of freedom	4	6	5	4	1	3	3	4	4	3

The set of conformations for the main chain of glycine is larger than that of other residues; this is accounted for by assigning three rather than two rotatable angles for glycine.

Table 3. Structural characteristics of proteins.

	SCOP 1.61 (5829 proteins)			
	Class <i>a</i> (1133)	Class <i>b</i> (1644)	Class <i>c</i> (1617)	Class <i>d</i> (1435)
$\langle L \rangle$	151 ± 3	156 ± 2	256 ± 3	163 ± 2
$\langle m(8 \text{ \AA}) \rangle$	18.33 ± 0.09	19.48 ± 0.06	21.56 ± 0.06	19.39 ± 0.07
$\langle \nu \rangle$	3.87 ± 0.01	3.71 ± 0.01	3.76 ± 0.01	3.80 ± 0.01
$\langle C_{\text{mod}} = \nu/m \rangle$	0.219 ± 0.001	0.196 ± 0.002	0.178 ± 0.001	0.200 ± 0.001

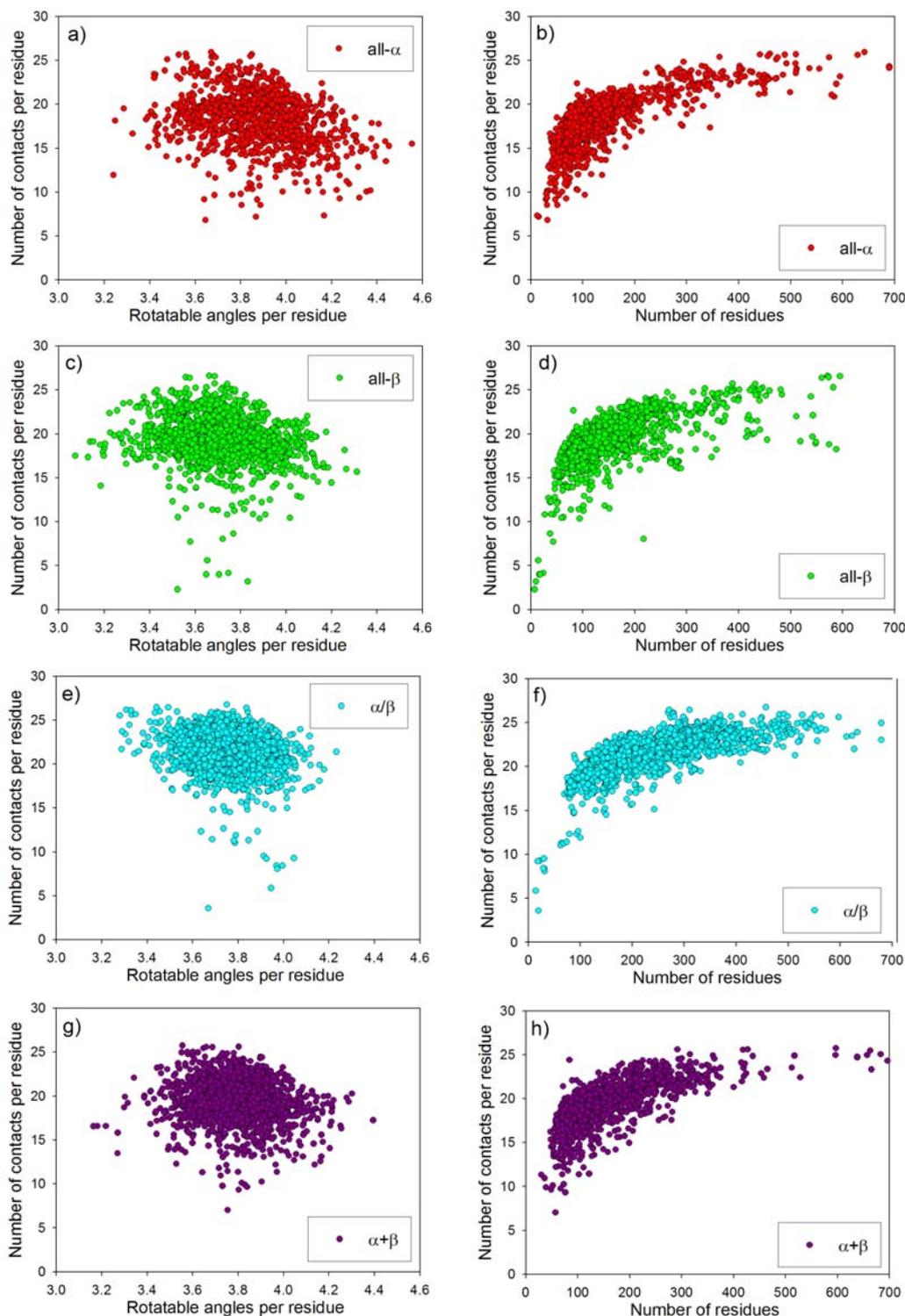
L is the number of residues in the protein; $m(8 \text{ \AA})$ is the number of contacts (at a contact distance of 8 Å) per residue; ν is the number of rotatable angles per residue; C_{mod} is the model entropy capacity which is the relation between ν and m . Averaging is done for each class of proteins.

To prove that the result is not a consequence of possible differences in length distribution between proteins of class *c* and others, we analyzed the number of contacts in each structural class within a given size range where the average length of proteins is nearly the same in each structural class (see Figure 3c,d). It turns out that the proteins of class *c* indeed have the largest number of contacts than proteins of all other structural classes (all- α , all- β , $\alpha+\beta$), and this trend persists in all length windows (see Figure 3c,d) where the number of proteins is sufficient for statistics (see Figure 3a) with the exception of only one site (200-250 residues in the protein chain) where four classes have practically the same number of contacts. We checked that this result (site 200-250) is the consequence of using the database of proteins with homology less than 80%. If one considers a database of proteins with homology less than 25%, then at the site of 200-250 residues the *c* class of proteins has the largest number of contacts than other classes (the result is not presented).

The reasons for the observed differences in the number of contacts per residue have been investigated [30]. Apart from the protein size, the number of contacts also depends on the tightness of packing and on the protein shape. It turns out that, between two structures of the same size and shape, the structure with a tighter packing will have a greater number of contacts, while between two structures of the same size and tightness of packing, the structure with a less surface area (*i.e.*, a more

“spherical” structure) will have a greater number of contacts. We have explored separately the influence of tightness of packing and shape.

Figure 2. (a, c, e, g) Distribution of proteins on the average number of rotatable angles ϕ , ψ and χ per residue and the average number of contacts per residue (the contact radius is 8 Å) for four general classes. (b, d, f, h) Distribution of the number of contacts per residue on the protein size for four general classes. Different colors correspond to various structural classes (red—class *a*, green—class *b*, blue—class *c*, and violet—class *d*).



To this end, the solvent-accessible surface area S and volume V surrounded by this surface for each protein considered have been calculated by program YASARA [<http://yasara.org>], using 1.4 Å as the probe radius of a water molecule. The analysis of protein structures having from 51 to 350 residues revealed that the packing efficiency, indeed, is the same for proteins from different structural classes evidenced by the molecular volume per atom (for all- α proteins – $18.520 \pm 0.010 \text{ \AA}^3$, for all- β proteins – $18.577 \pm 0.009 \text{ \AA}^3$, for α/β proteins – $18.618 \pm 0.007 \text{ \AA}^3$, and for $\alpha+\beta$ proteins – $18.598 \pm 0.009 \text{ \AA}^3$) [18]. Therefore, the average molecular volume per atom V , which excludes water, is practically the same for all classes of proteins. This means that the tightness of packing is practically the same for all the four investigated protein classes.

Next, we explored the influence of the protein shape on the average number of contacts per residue. We used two ratios: V/S for different geometric bodies of equal volumes, this ratio should be maximal for a ball, and this value is proportional to the radius of the cross-section [18]) and S_0/S which we call compactness (*i.e.*, the ratio of the surface area of a sphere of the same volume as that of the protein to the accessible surface area of the protein; for a sphere, this ratio will be 1). The average values of these parameters are statistically higher for class c than for other classes (Figure 4), which holds true for proteins of all sizes. This means that proteins from class c , on average, are more compact than proteins from other classes.

On the whole, the presented analysis shows that among proteins of the same size, α/β proteins have, on average, a greater number of contacts per residue due to their more compact (more ball-like, since a ball is the most compact geometrical body having the minimal accessible surface area in comparison with other geometrical bodies of equal volumes) structure, rather than to tighter packing.

The value of the average conformational entropy, ν , was calculated as a summation of the individual number of rotatable angles ϕ , ψ and χ for each residue (see Table 2) over its complete sequence normalized by the number of residues in the protein. Figure 2 demonstrates that topologically diverse proteins fall in a definable region of the average conformational entropy, 3.4–4.2 [27].

It is noteworthy that proteins with the smallest value of the average number of rotatable angles per residue belong to class b (see Figure 2), and proteins with the highest value of the average number of rotatable angles belong to class a (as a rule some α -helical DNA-binding proteins, such as transcription factors, have a high conformational entropy). Moreover, the averaging of the obtained values over the proteins of each structural class revealed (see Table 3) that class a has the highest average number of rotatable angles while this value is minimal for class b .

We also analyzed the number of rotatable angles per residue in each structural class within a given size range where the average length of proteins is nearly the same in each structural class (see Figure 3b). Our differential picture for conformational entropy just confirms our results obtained from the consideration of all proteins: class a has the highest average number of rotatable angles while this value is minimal for class b . This result supports the hypothesis that the loss of side-chain entropy is a major determinant of the helix-forming propensities of naturally occurring residues [42].

Figure 3. (a) Distribution of proteins composing our database by their size. Two vertical lines indicate the range of sizes (proteins 51-350 residues long) where the number of proteins is sufficient for our analysis. (b) Number of rotatable angles per residue in the proteins of given size. (c) Number of residue-residue contacts (per residue) at a contact distance of 6 Å in proteins of given size. (d) Number of residue-residue contacts (per residue) at a contact distance of 8 Å. The colors are the same as used in Figure 2.

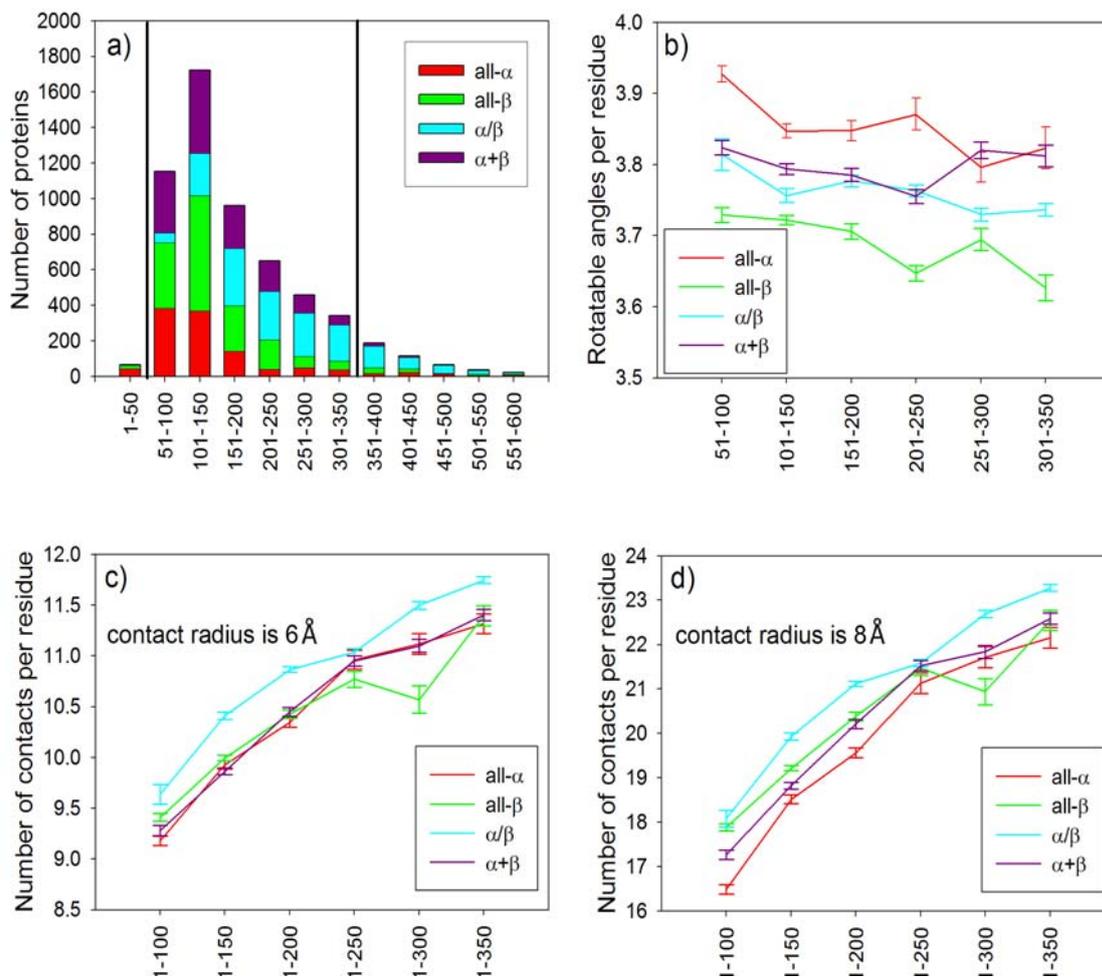
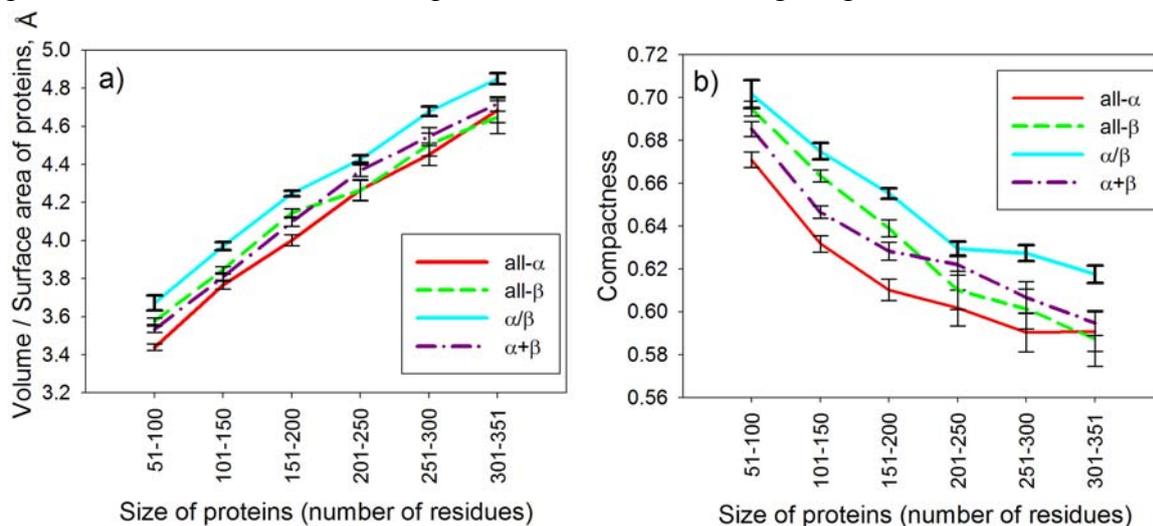


Table 4. Calculated parameters and experimentally measured folding rates of proteins.

SCOP fold	Name of protein	PDB entry	Degrees of freedom per residue	Number of contacts per residue	$C_{mod=v/m}$ (8 Å)	k_f in water, s^{-1}
b.1	Tenascin (short form) [43]	1TEN	3.74	18.52	0.202	3
	Fibronectin 10th FN3 module [44]	1FNF	3.36	18.43	0.182	244.7
b.34	SH3 domain (PI3 kinase) [45]	1PNJ	3.91	16.53	0.236	0.3
	SH3 domain (c-src protein tyrosine kinase) [46]	1SRL	3.70	16.79	0.220	54.6
b.40	CspA (<i>Escherichia coli</i>) [47]	1MJC	3.67	18.00	0.204	200.3
	CspB (<i>Bacillus subtilis</i>) [48,49]	1CSP	3.90	17.67	0.220	897.8
d.58	HypF-N [25]	1GXT	3.82	19.35	0.197	81.5
	Spliceosomal protein U1A [50]	1URN	4.00	19.38	0.206	330.3

Figure 4. (a) Average ratio of volume to accessible surface area (*i.e.*, radius of cross-section) in each size range for the four general SCOP classes. (b) Average compactness, *i.e.*, ratio of the accessible surface area of a sphere of the same volume as the protein to the surface area of the protein. The error of average is given.



One of the main determinants of the protein folding rate is the number of degrees of freedom that are to be fixed during the rate-limiting step of folding. Therefore, the difference in the folding rate for proteins sharing the same fold (SH3 domains, cold shock proteins, fibronectin domains, proteins belonging to the ferredoxin fold [8–12]) is the result of differences in the conformational entropy. In our case, the difference in the number of rotatable bonds ranges in terms of energy from 0.03 to 0.06 kcal/mol per residues, which could account for a factor of 100 and 10,000 in the spread of folding times for proteins of ~ 100 residues. To demonstrate this, we have analyzed the proteins from the database of 67 proteins which belong to the same fold (according to SCOP) separately. Proteins that have a similar number of contacts per residue are listed in Table 4. One can see that the folding rates of the proteins significantly differ: it is a factor of four in the case of cold shock proteins (fold b.40), while in the fold b.34 one protein folds 180 times faster than the other one. This difference in the folding rates can be explained by a different number of degrees of freedom per residue for these proteins (see Table 4).

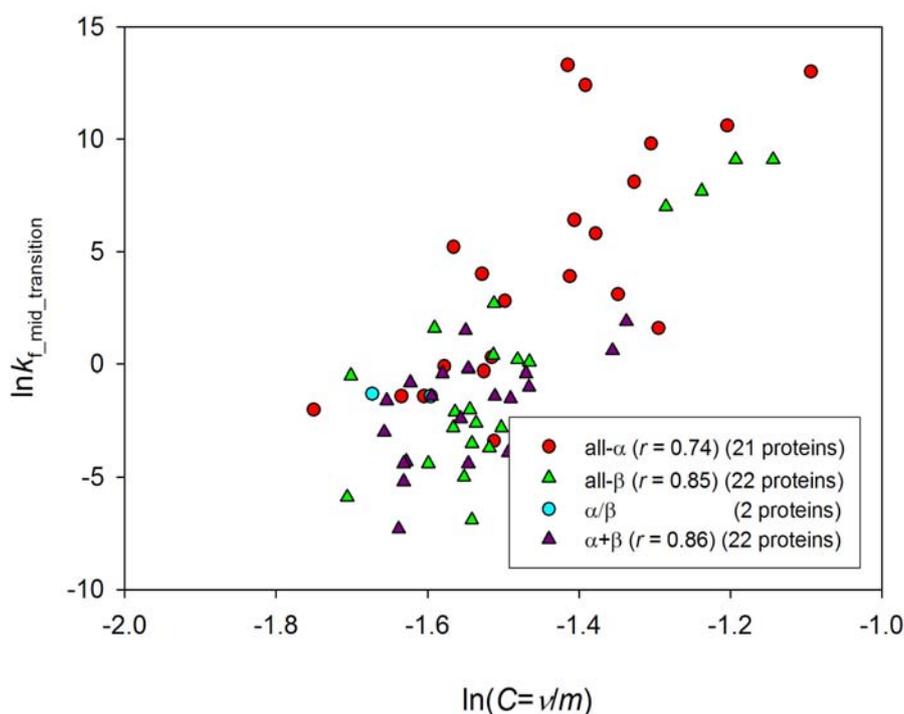
2.4. Statistical Analysis of Average Conformational Entropy and Average Number of Contacts per Residue for Different Classes of Proteins

Now we are interested only in the dependence of the change of the free-energy barrier dg^\ddagger on the change of value C , which depends on the properties of amino-acid residues in the protein (see Equation 11). Since the value of conformational entropy is proportional to the number of degrees of freedom (*i.e.*, the number of rotatable angles φ , ψ and χ) and the value of energy per residue proportional to the number of atom-atom (or residue-residue) contacts in protein, then the entropy capacity is proportional to the relation of the above values: $C \approx \text{const} \times C_{\text{mod}}$ where model entropy capacity $C_{\text{mod}} = \nu/m$ with m being the number of residue-residue contacts in the native structure (two residues were considered in contact with each other if any atom-atom pairs were at a distance of less

than 8\AA , the contacts between adjacent residues were not taken into account), and ν being the number of rotatable angles φ , ψ and χ (see Table 2). The model entropy capacity has been calculated for 67 proteins with known experimental data for the folding rate and for which the Monte-Carlo simulations has been done for comparison of two methods for prediction of protein folding rate [51]. Figure 5 demonstrates that the logarithm of entropy capacity correlates by about 78% with the logarithm of protein folding rate in mid-transition. Using model values for entropy capacity we can see that helical proteins are more frequently selected for fast folding than other classes of proteins.

It should be noted that entropy capacity correlates with folding rate in mid-transition practically at the same level as the first passage times obtained from Monte Carlo simulations (82%, see Table 5) [51]. The capillarity model allows us to predict the folding rate at the same level of correlation as by Monte Carlo simulations but without time-consuming calculations required for large proteins. The correlation between the logarithm of the first passage times in Monte Carlo simulations and the logarithm of folding rates for separate classes of proteins is sufficiently high: -0.74 for class *a*, -0.83 for class *b*, -0.84 for class *d* (see Table 5, and <http://phys.protres.ru/resources/compact.html>, the correlation for class *c* is absent, so only two proteins are included in the set of 67 proteins).

Figure 5. Correlation between the computed model entropy capacity and the experimentally measured folding rate at the mid-transition for 67 proteins. Both are represented in a logarithmic scale. The correlation coefficient is 0.78 ± 0.05 .



We compare the correlation between folding rates and a set of structural parameters for 67 proteins: contact order, length and the different functional forms of length (see Table 5). It is worth underlining here, that the correlation between folding rates and such structural parameters as contact order is absent for all proteins. The correlation with the length reaches 70% as has been mentioned in the Introduction. Although, the dependence of the logarithm of folding rate proportional to L has

Levinthal's estimation of folding time, it is not applicable for consideration. Such a good correlation means that there is a rather great scatter in the data, or the range of variation is too small. It is surprising that the correlation between the folding rate and the different functional forms of lengths ($\ln(L)$, $L^{2/3}$, $L^{1/2}$) is high, which means that the size of the proteins is one of the important and determining parameters of protein folding [4,18] (see Table 5). But the protein size cannot explain the difference in the folding rates for proteins of a similar size (or proteins sharing a similar fold, practically with a similar number of amino acid residues). All other considered parameters are used to explain the differences in protein folding for proteins of practically the same size. We demonstrate that the relation between the folding rate and entropy capacity has class-specific features. It is surprising that average structural properties give a high correlation with the folding rate in each structural class. At the same time the radius of the cross-section (V/S) has a worse correlation coefficient with the folding rate than the size of proteins for the considered set of proteins.

Table 5. Correlation coefficients between logarithms of folding rates in mid-transition and parameters connected with protein size and shape for three protein classes (*a*, *b*, *d*).

	$\ln k_{mt}$ 67 proteins	$\ln k_{mt}^a$ 21 proteins	$\ln k_{mt}^b$ 22 proteins	$\ln k_{mt}^d$ 22 proteins
<i>L</i>	-0.69 ± 0.06	-0.70 ± 0.11	-0.76 ± 0.09	-0.72 ± 0.10
$\ln(L)$	-0.76 ± 0.05	-0.73 ± 0.10	-0.85 ± 0.06	-0.82 ± 0.07
$L^{1/2}$	-0.73 ± 0.06	-0.72 ± 0.11	-0.82 ± 0.07	-0.77 ± 0.09
$L^{2/3}$	-0.72 ± 0.06	-0.72 ± 0.11	-0.80 ± 0.08	-0.75 ± 0.09
<i>AbsCO</i>	-0.81 ± 0.04	-0.76 ± 0.09	-0.88 ± 0.05	-0.74 ± 0.10
<i>RelativeCO</i>	-0.12 ± 0.12	0.37 ± 0.19	0.44 ± 0.17	-0.38 ± 0.18
Normalized radius of gyration, R_g/R_g^* [18]	0.08 ± 0.12	-0.34 ± 0.19	0.41 ± 0.18	0.32 ± 0.19
Radius of cross-section, V/S	-0.53 ± 0.09	-0.47 ± 0.17	-0.55 ± 0.15	-0.48 ± 0.16
$\ln(\text{Entropy capacity})$	0.78 ± 0.05	0.74 ± 0.10	0.85 ± 0.06	0.86 ± 0.06
$\ln(t_{1/2})$, Monte-Carlo steps	-0.82 ± 0.04	-0.74 ± 0.10	-0.83 ± 0.07	-0.84 ± 0.06

We also analyzed the value of model entropy capacity in each structural class within a given size range where the average length of proteins is nearly the same in each structural class (see Figure 6). The differential pattern of entropy capacity only confirms our results obtained from the consideration of all proteins: class *a* comprises the fastest folding proteins, then follow proteins of classes *b* and *d*, and class *c* comprises the slowest folding proteins. The number of contacts depends on the choice of the contact radius; therefore we checked whether the agreement between folding rates for four general classes is conserved at a varying contact radius. Figure 6 also demonstrates the value of model entropy capacity for two different values of the contact radius: 6 Å and 8 Å. In all cases, we observed that proteins of class *a* are the fastest folding ones, then follow proteins of classes *b* and *d*. The slowest folding proteins are those from class *c*.

We used a *t*-test to compare the average values of the model entropy capacity in different structural classes. The results of this comparison are shown in Table 6. There is a significant difference in average between all classes, whereas the difference between β and $\alpha+\beta$ proteins is not so significant.

From this result we conclude that each class has a class-specific average number of contacts and conformational entropy. In addition to the relationship observed here between a larger number of contacts per residue and slower folding for class *c* proteins, other properties specific to class *c* proteins have been observed in previous works. It has been shown that α/β proteins are relatively older than proteins from other classes [52]. This is supported by the observation that eukaryotic innovation domains (which only exist in eukaryotes and, therefore, are believed to have evolved later) are statistically less compact than the older, prokaryotic-only domains [53]. Also, fold's contact trace in proteins is proportional to protein designability (the ability of a protein to accept mutations that do not destroy the original fold and is correlated with the logarithm of the average gene family size) [54,55]. This, too, is consistent with our results because class *c* with class *b* includes the largest number of proteins (see Table 3). The higher trace conformations exhibit a marked tendency toward greater structural regularity and symmetry [54,55]. This is typical of proteins from class *c*.

Therefore, the work presented here demonstrates that a more compact structure and slower folding can be expected for α/β proteins, which are older and more designable than proteins from the other structural classes.

Figure 6. Dependence of the model entropy capacity on protein size for four general classes (large database) when contact radius is 8 Å (lower four curves) and when it is 6 Å (upper four curves). The average values of the model entropy capacity for all proteins of four classes are presented in the inset.

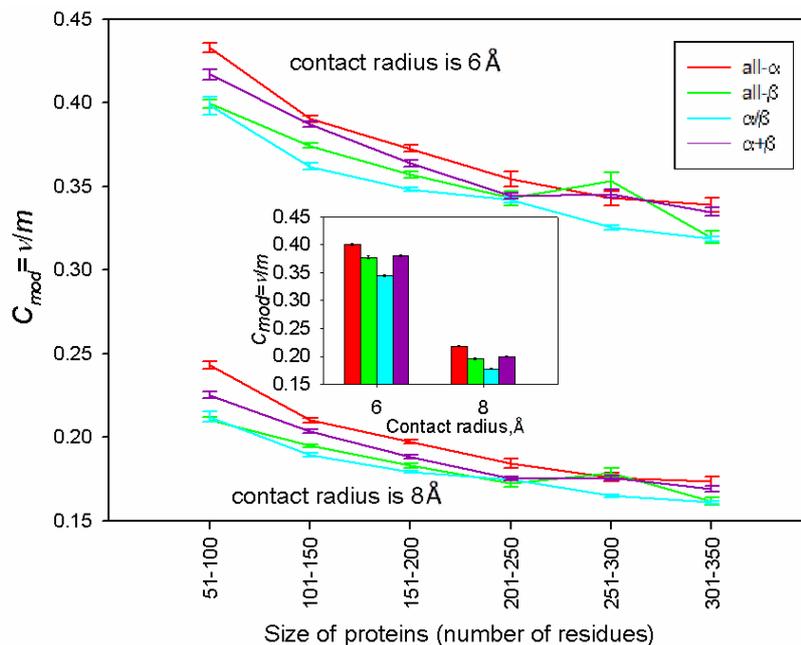


Table 6. Two-sample t-test performed on structural classes.

	Class <i>a</i>	Class <i>b</i>	Class <i>c</i>	Class <i>d</i>
Class <i>a</i>		1.3×10^{-25}	8.4×10^{-109}	5.3×10^{-27}
Class <i>b</i>	1.3×10^{-25}		1.5×10^{-25}	4.3×10^{-02}
Class <i>c</i>	8.4×10^{-109}	1.5×10^{-25}		2.1×10^{-64}
Class <i>d</i>	5.3×10^{-27}	4.3×10^{-02}	2.1×10^{-64}	

2.5. Behavior of Proteins with a High and Low Number of Contacts and Side-Chain Entropy

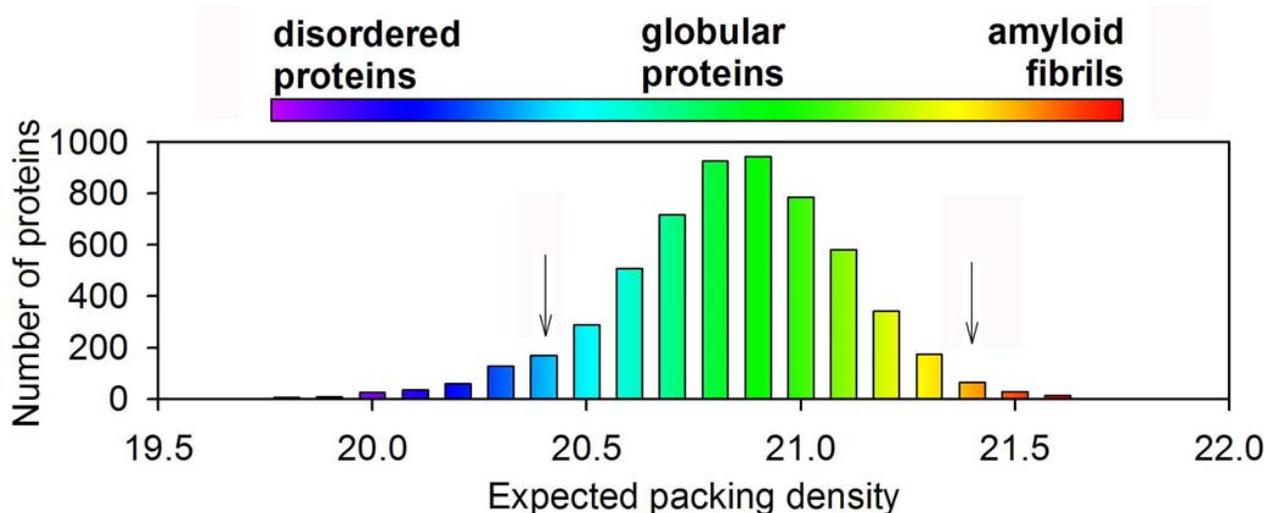
The competition of the side-chain entropy of the given protein, and favorable inter-residue interactions for the given topology, will determine the possibility of the given chain to fold to a particular topology. The formation of sufficient residue-residue interactions is required to compensate side-chain conformational entropy. Therefore, structural uniqueness of native proteins is the result of the balance between the conformational entropy of side-chains and the energy of residue interactions. Proteins with higher or lower side-chain entropy have, as a rule, extended unfolded regions and/or require some additional agents for complete folding. For example: some DNA-binding proteins such as transcription factors (with high side-chain entropy) have regions which only became structured upon DNA binding [56,57]; α -lytic protease and subtilysin (with low side-chain entropy because of the relatively high *Gly* and *Ala* content which results in a high conformational entropy of the backbone chain) use a pro region to promote folding [58,59]; *N*-terminal region of prions [60] with low side-chain entropy can adopt a more stable conformation which is induced by interaction with other prions with that conformation [61]. It seems that proteins with large conformational entropy (out of the optimal region) have no sufficient energetic interactions to compensate such large entropy. Therefore, enhanced stabilization for them is achieved by additional interactions with other agents or by oligomerization.

It seems that disordered regions in a protein chain do not have a sufficient number of interactions to compensate for the loss of conformational entropy, which results from the formation of a globular state. On the other hand, a large increase in the energy of interactions will lead to a loss of the unique structure because the strengthening of contact energy will speed up folding, but it is also likely to lead to erroneous folds (for example, to amyloid fibrils). It has been suggested that the lack of a rigid globular structure under physiological conditions might represent a considerable functional advantage for intrinsically disordered proteins, as their large plasticity allows them to interact efficiently with several different targets, as compared with a folded protein with limited conformational flexibility [62–67]. It has been shown that disordered regions are involved in DNA binding and other types of molecular recognition [68].

The need to obtain a definite balance between conformational entropy and energy of interactions is one of the general conditions to achieve the functional active form of the protein. For some proteins (with high conformational entropy or low energy of residue-residue interactions) such a balance can be achieved only by oligomerization with the same proteins and, for other proteins, only by interaction with additional agents. The shift of this balance caused by changing the external condition (pH or temperature) can result in the formation of stable intermediates as a molten globule-like one [69] or fibrils which may play a pathological role in a cell [70–72].

The obtained results may be useful in protein design. When we attempt to create de-novo protein that has a unique structure, it is important to answer the question whether a protein will be folded or natively unfolded. Such a parameter as the mean packing density is able to detect both amyloidogenic and disordered regions in the protein sequence [66,67]. (Each residue in a protein sequence is first assigned an expected number of contacts (ENC, or expected packing density) found from the statistics of globular structures).

Figure 7. Histogram representing the distribution of 5829 globular protein domains as a function of the expected packing density. Arrows indicate upper and lower thresholds which correspond to unusually strong and unusually weak expected packing density.



The observed average number of contacts in a globular state for each of the 20 residue types was calculated using 5829 protein structures from the SCOP protein database. Then an average value of ENC_s was computed and assigned to the central residue in each window of several sequential residues. If the average ENC value of a residue is larger than the threshold value of 20.4, the residue is considered to be rigid otherwise it is flexible [67]). It has been shown that regions with strong ENC_s (expected number of contacts) are responsible for the amyloid formation and regions with weak expected packing density are responsible for the appearance of disordered regions (see Figure 7) [73].

3. Conclusion

The existence of the optimal region of the entropy capacity, relating two important factors of protein folding such as the conformational entropy and the energy of residue-residue interactions, suggests some optimum balance between them for fast folding.

A definite balance between the conformational entropy and the energy of interactions is one of the general conditions to get a functionally active form of the protein. For some proteins (with a high conformational entropy or low energy of residue-residue interactions), such a balance can be achieved only by oligomerization with the same proteins, and, for other proteins, only by interaction with additional agents. Strong contact energy will speed up the folding rate, but it is likely to result in misfolds. Namely, taking into account the existence of such a balance, we can explain why sometimes less stable proteins fold faster than more stable proteins with the same topology.

We consider such intrinsic structural properties of protein structure that provide information about intrinsic preferences for the given class of protein structures: the number of rotatable angles and the number of contacts per residue. It should be noted that the protein classes differ from one another in a statistically significant manner with respect to the model entropy capacity in our case. The result suggests that structural and sequence properties are important determinants of the folding rate. Such

topological features as the number of contacts and number of rotatable angles (which is defined by the content and connectivity of secondary structures) can significantly affect the protein folding rates.

The understanding of the folding mechanisms and the reason for fast protein folding can help in the design of new proteins and in the understanding of protein misfolding. When we attempt to create a *de novo* protein that has a unique structure, it is important to consider the relationship between the number of rotatable angles and the average number of contacts for optimal folding of this protein.

Acknowledgements

I am grateful to A.V. Finkelstein for some discussion and S.O. Garbuzynskiy, D.N. Ivankov, M.Yu. Lobanov, and N.S. Bogatyreva for their assistance in preparation of some figures and 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), the grant from the Federal Agency for Science and Innovations (#02.740.11.0295), and by the "Russian Science Support Foundation".

References and Notes

1. Finkelstein, A.V.; Badretdinov, A.Ya. Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. *Fold. Des.* **1997**, *2*, 115–121.
2. Jackson, S.E. How do small single-domain proteins fold? *Fold. Des.* **1998**, *3*, R81–R91.
3. Plaxco, K.W.; Simons, K.W.; Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **1998**, *277*, 985–994.
4. De Sancho, D.; Doshi, U.; Munoz, V. Protein folding rates and stability: How much is there beyond size? *J. Am. Chem. Soc.* **2009**, *131*, 2074–2075.
5. Galzitskaya, O.V.; Garbuzynskiy, S.O.; Ivankov, D.N.; Finkelstein, A.V. Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. *Proteins* **2003**, *51*, 162–166.
6. Kuznetsov, I.B.; Rackovsky, S. Class-Specific correlations between protein folding rate, structure-derived, and sequence-derived descriptors. *Proteins* **2004**, *54*, 333–341.
7. Finkelstein, A.V.; Galzitskaya, O.V. Physics of protein folding. *Phys. Life Rev.* **2004**, *1*, 23–56.
8. Guijarro, J.I.; Morton, C.J.; Plaxco, K.W.; Campbell, I.D.; Dobson, C.M. Folding kinetics of the SH3 domain of PI3 kinase by Real-Time NMR combined with optical spectroscopy. *J. Mol. Biol.* **1998**, *276*, 657–667.
9. Plaxco, K.W.; Guijarro, J.I.; Morton, C.J.; Pitkeathly, M.; Campbell, I.D.; Dobson, C.M. The folding kinetics and thermodynamics of the Fyn-SH3 domain. *Biochemistry* **1998**, *37*, 2529–2537.
10. Perl, D.; Welker, Ch.; Schindler, Th.; Schroder, K.; Marahiel, M.A.; Jaenicke, R.; Schmid, F.X. Conservation of rapid two-state folding in Mesophilic, Thermophilic and Hyperthermophilic cold shock proteins. *Nature Struct. Biol.* **1998**, *5*, 229–235.
11. Van Nuland, N.A.J.; Chiti, F.; Taddei, N.; Raugei, G.; Ramponi, G.; Dobson, C.M. Slow folding of muscle acylphosphatase in the absence of intermediates. *J. Mol. Biol.* **1998**, *283*, 883–891.

12. Zerovnik, E.; Virden, R.; Jerala, R.; Turk, V.; Waltho, J.P. On the mechanism of human Stefin B folding: I. comparison to homologous Stefin A. influence of Ph and Trifluoroethanol on the fast and slow folding phases. *Proteins* **1998**, *32*, 296–303.
13. Thirumalai, D. From minimal models to real proteins: Time Scales for Protein Folding Kinetics. *J. Phys. Orsay Fr.* **1995**, *5*, 1457–1467.
14. Gutin, A.M.; Abkevich, V.I.; Shakhnovich, E.I. Chain length scaling of protein folding time. *Phys. Rev. Lett.* **1996**, *77*, 5433–5456.
15. Finkelstein, A.V.; Badretdinov, A.Ya. Physical reasons for fast folding of stable spatial structure of proteins: A solution of the Levinthal Paradox. *Mol. Biol. (Russia) Engl. Edition* **1997**, *31*, 391–398.
16. Koga, N.; Takada, S. Roles of native topology and chain-length scaling in protein folding: A simulation study with a Gō-Like model. *J. Mol. Biol.* **2001**, *313*, 171–180.
17. Galzitskaia, O.V.; Ivankov, D.N.; Finkelstein, A.V. Folding nuclei in proteins. *FEBS Lett.* **2001**, *489*, 113–118.
18. Ivankov, D.N.; Bogatyreva, N.S.; Lobanov, M.Yu.; Galzitskaya, O.V. Coupling between properties of the protein shape and the rate of protein folding. *PLoS ONE* **2009**, *4*, e6476.
19. Fersht, A.R. Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 1525–1529.
20. Ivankov, D.N.; Garbuzynskiy, S.O.; Alm, E.; Plaxco, K.W.; Baker, D.; Finkelstein, A.V. Contact order revisited: influence of protein size on the folding rate. *Protein Sci.* **2003**, *12*, 2057–2062.
21. Makarov, D.E.; Keller, C.A.; Plaxco, K.W.; Metiu, H. How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. *Proc. Natl. Acad. Sci. USA* **2002**, *9*, 3535–3539.
22. Punta, M.; Rost, B. Protein folding rates estimated from contact predictions. *J. Mol. Biol.* **2005**, *348*, 507–512.
23. Gong, H.; Isom, D.G.; Srinivasan, R.; Rose, G.D. Local secondary structure content predicts folding rates for simple, two-state proteins. *J. Mol. Biol.* **2003**, *327*, 1149–1154.
24. Shao, H.; Peng, Y.; Zeng, Z.-H. A simple parameter relating sequences with folding rates of small alpha helical proteins. *Protein Pept. Lett.* **2003**, *10*, 277–280.
25. Calloni, G.; Taddei, N.; Plaxco, K.W.; Ramponi, G.; Stefani, M.; Chiti, F. Comparison of the folding processes of distantly related proteins. Importance of hydrophobic content in folding. *J. Mol. Biol.* **2003**, *330*, 577–591.
26. Naganathan, A.N.; Munoz, V. Scaling of folding times with protein size. *J. Am. Chem. Soc.* **2005**, *127*, 480–481.
27. Galzitskaya, O.V.; Surin, A.K.; Nakamura, H. Optimal region of average side-chain entropy for fast protein folding. *Protein Sci.* **2000**, *9*, 580–586.
28. Murzin, A.G.; Brenner, S.E.; Hubbard, T.; Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247*, 536–540.
29. Galzitskaya, O.V.; Garbuzynskiy, S.O. Entropy capacity determines protein folding. *Proteins* **2006**, *63*, 144–154.

30. Galzitskaya, O.V.; Danielle, C.; Reifsnnyder, D.C.; Bogatyreva, N.S.; Ivankov, D.N.; Garbuzynskiy, S.O. More compact protein globules exhibit slower folding rates. *Proteins* **2008**, *70*, 329–332.
31. Galzitskaya, O.V.; Bogatyreva, N.S.; Ivankov, D.N. Compactness determines protein folding type. *J. Bioinform. Comput. Biol.* **2008**, *6*, 667–680.
32. Taketomi, H.; Ueda, Y.; Gō, N. Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Protein Res.* **1975**, *7*, 445–459.
33. Lifshits, E.M.; Pitaevskii, L.P. *Physical Kinetics*; Pergamon: Oxford, UK, 1991.
34. Pande, V.S.; Grosberg, A.Yu.; Tanaka, T.; Rokhsar, D.S. Pathways for protein folding: Is a new view needed? *Curr. Opin. Struct. Biol.* **1998**, *8*, 68–79.
35. Fersht, A.R. Characterizing transition states in protein folding: An essential step in the puzzle. *Curr. Opin. Struct. Biol.* **1995**, *5*, 79–84.
36. Fersht, A.R. Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **1997**, *7*, 3–9.
37. Matouschek, J.T.; Kellis, J.T.Jr.; Serrano, L.; Fersht, A.R. Mapping the transition state and pathway of protein folding by protein engineering. *Nature* **1989**, *340*, 122–126.
38. Matouschek, A.; Kellis, J.T.Jr.; Serrano, L.; Bycroft, M.; Fersht, A.R. Transient folding intermediates characterized by protein engineering. *Nature* **1990**, *346*, 440–445.
39. Fersht, A.R.; Matouschek, A.; Serrano, L. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* **1992**, *224*, 771–782.
40. Abkevich, V.I.; Gutin, A.M.; Shakhnovich, E.I. Specific nucleus as a transition state for protein folding: an evidence from lattice model. *Biochemistry* **1994**, *33*, 10026–10036.
41. Wolynes, P.G. Folding funnels and energy landscapes of larger proteins within the capillarity approximation. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 6170–6175.
42. Creamer, T.P.; Rose, G.D. Side-chain entropy opposes alpha-Helix formation but rationalizes experimentally determined Helix-forming propensities. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 5937–5941.
43. Clarke, J.; Hamill, S.J.; Johnson, C.M. Folding and stability of a fibronectin type III domain of human tenascin. *J. Mol. Biol.* **1997**, *270*, 771–778.
44. Cota, E.; Clarke, J. Folding of beta-sandwich proteins: Three-state transition of a Fibronectin type III module. *Protein Sci.* **2000**, *9*, 112–120.
45. Guijarro, J.I.; Morton, C.J.; Plaxco, K.W.; Campbell, I.D.; Dobson, C.M. Folding kinetics of the SH3 domain of PI3 kinase by Real-Time NMR combined with optical spectroscopy. *J. Mol. Biol.* **1998**, *276*, 657–667.
46. Grantcharova, V.P.; Baker, D. Folding dynamics of the Src SH3 domain. *Biochemistry* **1997**, *36*, 15685–15692.
47. Reid, K.L.; Rodriguez, H.M.; Hillier, B.J.; Gregoret, L.M. Stability and folding properties of a model beta-sheet protein, *Escherichia coli* Cspa. *Protein Sci.* **1998**, *7*, 470–479.
48. Perl, D.; Welker, C.; Schindler, T.; Schroder, K.; Marahiel, M.A.; Jaenicke, R.; Schmid, F.X. Conservation of rapid two-state folding in Mesophilic, Thermophilic and Hyperthermophilic cold shock proteins. *Nature Struct. Biol.* **1998**, *5*, 229–235.

49. Schindler, T.; Herrler, M.; Marahiel, M.A.; Schmid, F.X. Extremely rapid protein folding in the absence of intermediates. *Nature Struct. Biol.* **1995**, *2*, 663–673.
50. Silow, M.; Oliveberg, M. High-energy channeling in protein folding. *Biochemistry* **1997**, *36*, 7633–7637.
51. Galzitskaya, O.V. Estimation of protein folding rate from Monte-Carlo simulations and entropy capacity. *Curr. Protein Pept. Sci.* In press.
52. Winstanley, H.F.; Abeln, S.; Deane, C.M. How old is your fold? *Bioinformatics* **2005**, *21*, i449–i458.
53. Tiana, G.; Shakhnovich, B.E.; Dokholyan, N.V.; Shakhnovich, E.I. Imprint of evolution on protein structures. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 2846–2851.
54. Shakhnovich, B.E.; Deeds, E.; Delisi, C.; Shakhnovich, E. Protein structure and evolutionary history determine sequence space topology. *Genome. Res.* **2005**, *15*, 385–392.
55. England, J.L.; Shakhnovich, B.E.; Shakhnovich, E.I. Natural selection of more designable folds: A mechanism for thermophilic adaptation. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 8727–8731.
56. McIntosh, P.B.; Frenkiel, T.A.; Wollborn, U.; McCormick, J.E.; Klempnauer, K.H.; Feeney, J.; Carr, M.D. Solution structure of the B-Myb DNA-binding domain: A possible role for conformational instability of the protein in DNA binding and control of gene expression. *Biochemistry* **1998**, *37*, 9619–9629.
57. Ippel, H.; Larsson, G.; Behravan, G.; Zdunek, J.; Lundqvist, M.; Schleucher, J.; Lycksell, P.O.; Wijmenga, S. The solution structure of the homeodomain of the rat insulin-gene enhancer protein Isl-1. Comparison with other homeodomains. *J. Mol. Biol.* **1999**, *288*, 689–703.
58. Baker, D. Metastable states and folding free energy barriers. *Nature Struct. Biol.* **1998**, *5*, 1021–1024.
59. Sohl, J.L.; Jaswal, S.S.; Agard, D.A. Unfolded conformations of α -Lytic protease are more stable than its native state. *Nature* **1998**, *395*, 817–819.
60. James, T.L.; Liu, H.; Ulyanov, N.B.; Farr-Jones, S.; Zhang, H.; Donne, D.G.; Kaneko, K.; Groth, D.; Mehlhorn, I.; Prusiner, S.B.; Cohen, F.E. Solution structure of a 142-residue recombinant prion protein corresponding to the infectious fragment of the scrapie isoform. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 10086–10091.
61. Taylor, K.L.; Cheng, N.; Williams, R.W.; Steven, A.C.; Wickner, R.B. Prion domain initiation of amyloid formation *in vitro* from native Ure2p. *Science* **1999**, *283*, 1339–1343.
62. Wright, P.E.; Dyson, H.J. Intrinsically unstructured proteins: Reassessing the protein structure–function paradigm. *J. Mol. Biol.* **1999**, *293*, 321–331.
63. Romero, P.; Obradovic, Z.; Dunker, A.K. Natively disordered proteins: Functions and predictions. *Appl. Bioinformatics* **2004**, *3*, 105–113.
64. Fink, A.L. Natively unfolded proteins. *Curr. Opin. Struct. Biol.* **2005**, *15*, 35–41.
65. Dunker, A.K.; Cortese, M.S.; Romero, P.; Iakoucheva, L.M.; Uversky, V.N. Flexible Nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.* **2005**, *272*, 5129–5148.
66. Galzitskaya, O.V.; Garbuzynskiy, S.O.; Lobanov, M.Y. Prediction of natively unfolded regions in protein chain. *Mol. Biol. (Moscow)* **2006**, *40*, 341–348.
67. Galzitskaya, O.V.; Garbuzynskiy, S.O.; Lobanov, M.Y. FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics* **2006**, *22*, 2948–2949.

68. Dyson, H.J.; Wright, P.E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell. Biol.* **2005**, *6*, 197–208.
69. Ptitsyn, O.B. Molten globule and protein folding. *Adv. Protein Chem.* **1995**, *47*, 83–229.
70. Chiti, F.; Webster, P.; Taddei, N.; Clark, A.; Stefani, M.; Ramponi, G.; Dobson, C.M. Designing conditions for *in vitro* formation of amyloid protofilaments and fibrils. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 3590–3594.
71. Harrison, P.M.; Chan, H.S.; Prusiner, S.B.; Cohen, F.E. Thermodynamics of model prions and its implications for the problem of prion protein folding. *J. Mol. Biol.* **1999**, *286*, 593–606.
72. Jimenez, J.L.; Guijarro, J.I.; Orlova, E.; Zurdo, J.; Dobson, C.M.; Sunde, M.; Saibil, H.R. Cryo-electron microscopy structure of an SH3 amyloid fibril and model of the molecular packing. *EMBO J.* **1999**, *18*, 815–821.
73. Galzitskaya, O.V.; Garbuzynskiy, S.O.; Lobanov, M.Y. Prediction of amyloidogenic and disordered regions in protein chains. *PLoS Comput. Biol.* **2006**, *2*, e177.

© 2010 by the authors; licensee Molecular Diversity Preservation International, Basel, Switzerland. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).