Description of the model and predictions of C-values for Protein

Free energy estimation

At first, we consider sequential unfolding of the native structure of a chain that consists of L links. This chain has the native structure I_0, the entirely unfolded structure I_L, and many intermediate semifolded microstates I_v, each of which includes v disordered links and a native-like globular portion consisting of L-v links. For I_0, $v = 0$; I_L, $v = L$ and for partially unfolded structures, $v = 1, \ldots L\text{-}1$. As mentioned above, we do not consider microstates with globular portions dissimilar to the native one.

The free energy of state I is expressed as:

$$F(I) = n_I \times £ - T[¬_I \times ± + \sum_{loops \in I} S_{loop}] \quad (1)$$

where n_I is the number of atom-atom contacts in the native-like portion of I (contacts between neighboring residues are neglected because they exist in the unfolded state as well); ε is the energy of one atom-atom contact in the protein (all contacts are assumed to have equal energies, and the interaction with the solvent is implicitly taken into account in ε); T is temperature; ν_I is the number of residues in the unfolded portion of the microstate; σ is the difference in entropy between the unfolded and folded states of the residue (according to experimental data reported in [1], σ for any residue is assumed to be $2.3R$, where R is the universal gas constant); S_{loop} is the entropy cost of the closing of a disordered loop protruding from the native-like portion of I (described by Eq. 3 presented below, and summed up for all loops present in the microstate).

In modeling the folding/unfolding of proteins that lack disulfide bonds, free-energy calculations concern only the thermodynamic equilibrium point between the native state I_0 and the coil I_L. At this point, $F(I_0) = F(I_L)$; i.e., the energy of one intraprotein contact and temperature obey the following equation:

$$£ = -TN± / n_0 \quad (2)$$

where n_0 is the number of atom-atom contacts in the native protein (that is, in state I_0), and N is the number of residues in the protein. Thus, we can express all free energies in terms of RT units. Note that n_I and n_0 (as well as ε) significantly depend on whether hydrogen atoms are included into calculation, in addition to heavier atoms (see below).

The entropy cost of the closing of a disordered loop protruding from the globule between specified residues k and l is estimated in [2] to be

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

where r_{kl} is the distance between the C_α atoms of residues k and l; a = 3.8 Å is the distance between C_α atoms neighboring in the chain; and A is the persistent length of the polypeptide (assumed to be A = 20 Å according to [3]). Based on the data reported in [4], we chose the following computation parameters: contact distance between heavy atoms $r_{hea\text{-}hea}$ = 6 Å; between hydrogen atoms, r_{HH} = 4 Å; and between an H atom and a heavy atom, $r_{H\text{-}hea}$ = ($r_{hea\text{-}hea}$ + r_{HH}) / 2 = 5 Å.

Complete set of possible transition states found by dynamic programming and calculation of Φ values

It is beyond reason to state that a protein folds and unfolds solely through the intermediate state with the lowest free energy. The molecule can cross the free-energy barrier through other saddles, not optimal but, probably, numerous. Therefore, their diversity should be evaluated.

State I corresponds to a saddle on the barrier if its own free energy $F(I)$ equals $F^{\#}(I)$, the height of the lowest free-energy barrier on pathways running from I_0 to I_L via I. For this reason, we select states with $F(I)=F^{\#}(I)$ and, in doing so, obtain the complete set $\{I^{\#}\}$ of all possible saddles on the free-energy barrier separating I_0 and I_L. This set provides for the upper limit of the diversity of TSs. It can be redundant, because one pathway to TS of a high free energy can pass through another TS with a lower free energy. The lower limit of the diversity is provided by the set $\{I^{\#}_{\min}\}$, which includes only intermediate states with the minimal free energy $F^{\#}_{\min}$. To outline the nucleus, we analyze both these sets.

To examine the set TS = $\{I^{\#}\}$ (or TS = $\{I^{\#}_{\min}\}$), we calculate Boltzmann's weights for all TSs of the set:

$$P(I^{\#}) = \frac{\exp(-F^{\#}(I^{\#})RT)}{\sum_{I \in \Pi C} \exp(-F^{\#}(I)/RT)} \quad (4)$$

The sum comprises all states forming the set TS = $\{I^{\#}\}$ (or TS = $\{I^{\#}_{\min}\}$ if we consider only intermediates with the lowest free energy). All intermediates of the set $\{I^{\#}_{\min}\}$ have equal free energies and, therefore, equal Boltzmann's weights. For the members of $\{I^{\#}\}$, free energies and, correspondingly, the $P(I^{\#})$ values can differ. In this case, the higher $P(I^{\#})$, the more rapid the way through the saddle $I^{\#}$, in agreement with the common exponential dependence of the reaction rate on the TS free energy; thus, the more often the chain crosses this saddle during its folding/unfolding.

The experimental Φ_f -value for residue r is determined in [5] as

$$\Phi_f = \Delta_r[F(TS) - F(I_L)] / \Delta_r[F(I_0) - F(I_L)], \quad (5)$$

where $\Delta_r[F(I_0) - F(I_L)]$ is the change in the free-energy difference between the native (I_0) and denatured (I_L) states caused by a mutation in residue r, and $\Delta_r[F(TS) - F(I_L)]$ is the change of the free-energy difference between the transition (TS) and denatured (I_L) states caused by the same mutation.

It is believed that these experimental Φ-values show the "nativity" of the residue and its neighborhood in the transition state [6,7] (Fersht, 1995; 1997). If Φ = 1, the residue under consideration is stated to be native together with its neighborhood; if Φ = 0, the residue has lost either its own native structure or its native neighborhood. Intermediate Φ values indicate that the neighborhood is only partially native.

These Φ-values can also be obtained theoretically. According to Eq. (1), $\Delta_r[F(I_0) - F(I_L)] = \varepsilon \times \Delta_r(n_0)$, where ε is the energy of one contact, and $\Delta_r(n_0)$ is the change in the number of contacts in the native state I_0 caused by a mutation in residue r, all native contacts are assumed to be equal and the entirely unfolded state is assumed to be devoid of contacts.
Similarly, $\Delta_r[F(TS) - F(I_L)] = \varepsilon \times \langle\Delta_r(n_I)\rangle_{I\#}$, where $\langle\Delta_r(n_I)\rangle_{I\#}$ is the change in the number of native contacts in the TS caused by the mutation and averaged over the TS ensemble (TS = $\{I^{\#}\}$ or TS = $\{I^{\#}_{\min}\}$). This change is calculated as follows:

$$\langle\Delta_r(n_I)\rangle_{I\#} = \Sigma_{I\#} P(I^{\#}) \Delta_r(n_{I\#}), \quad (6)$$

where $P(I^{\#})$ is the Boltzmann probability of microstate $I^{\#}$ in the TS ensemble (Eq. (4)), and $\Delta_r(n_{I\#})$ is the change of the number of native contacts in microstate $I^{\#}$ caused by the mutation of residue r.

The known change in the atomic structure of the mutated residue r allows calculation of the resulting $\Delta_r(n_I)$ values for each state I of the mutant protein starting from the atomic coordinates of the wild-type protein. Such calculations assume, though, that the coordinates of other atoms remain the same; i.e., that the mutation does not alter the structure of the native state. Hence, we should consider only mutations that do not add new "overcrowding" atoms to residue r. This is generally taken into account in experimental studies of folding nuclei [7,8].

The calculated values

$$\Phi = \langle\Delta_r(n_I^{nb})\rangle_{I\#} / \Delta_r(n_0^{nb}) \quad (7)$$

can be compared to the experimental Φ-values, and correlation between the prediction and the experiment can be estimated (see Fig.1). As seldom occurring values $\Phi < 0$ and $\Phi > 1$ have no structural interpretation, very rare mutations for which $\Phi < 0$ and $\Phi > 1$, were omitted from the comparison with the results of calculations.

In proteins containing disulfide bonds not cleaved during denaturation, we do not consider Φ-values for links containing cysteine residues involved in these bonds, because such links retain their "native" contact in all semiunfolded states.

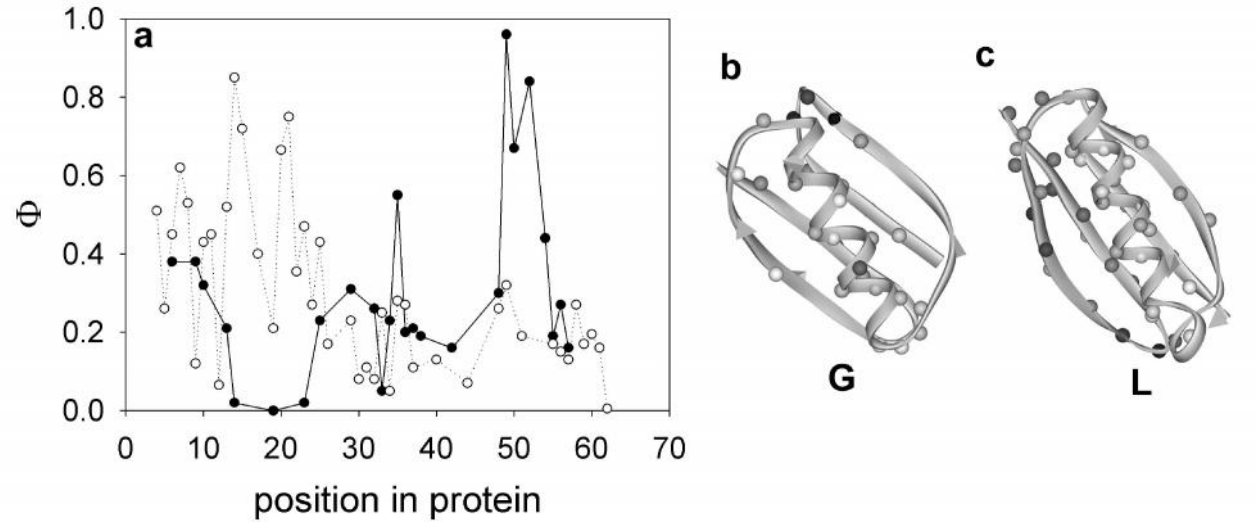


Fig.1. (a) Profiles of experimental Φ-values obtained for B1 domains of protein G (filled circles) and of protein L (open circles). **(b-c)** Schemes of 3D structures of these domains colored according to the Φ-values of the amino-acid residues, from white (Φ=0) to black (Φ=1). The experimentally studies residues are shown as beads against the background of the native chain fold. Φ-values are given for them only.

A comparison of calculation results with experimental data on the folding nuclei of 21 proteins shows that the model provides good Φ-value predictions for protein structures determined by X-ray analysis and, less successfully, for structures determined by nuclear magnetic resonance [4] (see Tables 1, 2, 3 below).

References

1. Privalov P.L. (1979) Stability of proteins: Small globular proteins. *Adv. Protein Chem.* **33**, 167–241.
2. Finkelstein A.V., Badretdinov A.Ya. (1997) Physical factors providing for rapid self-organization of stable spatial protein structure: A solution to the Lewinthal paradox. *Mol. Biol.* **31**, 469–477.
3. Flory P. 1971. *Statistical Mechanics of Chain Molecules* [Russian translation]. Moscow: Mir.
4. Garbuzynskiy S.O., Finkelstein A.V., Galzitskaya O.V. (2004) Outlining folding nuclei in globular proteins. *J. Mol. Biol.* **336**, 509–525.
5. Matouschek, J.T., Kellis, Jr., Serrano, L., Bycroft, M., and Fersht, A.R. (1990) *Nature,* **346**, 440–445.
6. Fersht A.R. (1995) Characterizing transition states in protein folding: An essential step in the puzzle. *Curr. Opin. Struct. Biol.* **5**, 79–84.
7. Fersht A.R. (1997) Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3–9.
8. Chiti F., Taddei N., White P.M., Bucciantini M., Magherini F., Stefani M., Dobson C.M. (1999) Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Struct. Biol.* **11**, 1005-1009.

Table 1. Correlation between predicted and experimental c=- values.

Protein name (PDB entry, method)	Number of residues in the protein	Total number of experiment-ally studied residues	Correlation coefficient (size of link is 5residues)				Minimal size of link	Correlation coefficient (Minimal size of link)			
			Optimal folding nucleus		The whole ensemble of folding nuclei			Optimal folding nucleus		The whole ensemble of folding nuclei	
			Without hydrogen atoms	With hydrogen atoms	Without hydrogen atoms	With hydrogen atoms		Without hydrogen atoms	With hydrogen atoms	Without hydrogen atoms	With hydrogen atoms
U1A (1urn, X-ray)	96	10	-0.83	0.80	-0.71	0.90	3	0.75	0.80	0.38	0.88
Domain SH3 of α- spectrin (1shg, X-ray)	57	13	0.73	0.71	0.81	0.82	2	0.75	0.74	0.77	0.77
Sso7d (1bf4, X-ray)	63	15	0.53	0.48	0.58	0.81	2	0.22	0.55	0.65	0.68
B1 domain of protein G (1pgb, X-ray)	56	20	0.68	0.70	0.74	0.76	2	0.84	0.78	0.79	0.83
Domain TNfn3 of tenascin (1ten, X-ray)	90	25	0.62	0.60	0.54	0.64	3	0.18	0.17	0.52	0.63
CheY (3chy, X-ray)	128	9	0.57	0.72	0.65	0.67	4	–*	0.92	0.68	0.61
Barnase (1rnb, X-ray)	109	19	0.59	0.60	0.59	0.68	4	0.42	0.45	0.55	0.68
suc1 (c, X-ray)	96	35	0.23	0.23	0.52	0.56	3	0.48	0.36	0.46	0.49
S6 (1ris, X-ray)	97	10	0.02	-0.01	0.41	0.49	3	0.36	0.15	0.46	0.51
CI2 (2ci2, X-ray)	64	34	0.18	0.51	0.41	0.50	2	0.58	0.59	0.60	0.63
FKBP12 (1fkb, X-ray)	107	22	0.22	0.44	0.27	0.34	4	-0.13	0.40	0.16	0.22
Im7 (1cei, X-ray)	85	19	-0.46	0.00	-0.26	-0.21	3	-0.29	-0.33	-0.14	-0.18
Mean for protein structures obtained by X-ray analysis (and its error)	**87.3±6.6**	**19.3±2.5**	**0.26±0.14**	**0.48±0.08**	**0.38±0.13**	**0.58±0.09**	**2.9±0.2**	**0.35±0.10**	**0.47±0.10**	**0.49±0.08**	**0.56±0.08**
Domain TI I27 of titin (1tiu, NMR)	89	22	0.33	0.59	0.68	0.70	3	0.34	0.34	0.57	0.52
src SH3 domain (1srm, NMR)	56	26	0.56	0.07	0.63	0.51	2	0.55	0.57	0.42	0.50
Villin 14T (2vil, NMR)	126	18	0.12	0.17	0.31	0.43	4	0.26	0.44	0.48	0.53
B1 IgG-binding domain of protein L (2ptl, NMR)	64	38	-0.35	0.19	-0.06	0.25	2	-0.14	-0.08	0.12	0.15
Barstar (1btb, NMR)	89	22	-0.32	-0.39	0.26	0.33	3	0.41	0.40	0.49	0.47
Domain FNfn10 of fibronectin (1ttf, NMR)	94	16	-0.05	-0.05	-0.27	-0.26	3	-0.17	-0.39	-0.49	-0.33
Acylphosphatase (1aps, NMR)	98	17	0.26	-0.06	0.48	0.29	3	0.52	0.19	0.54	0.49
Im9 (1imp, NMR)	86	19	-0.58	-0.31	-0.14	-0.09	3	-0.24	0.26	0.04	-0.09
Domain B of protein A (1bdc, NMR)	60	15	0.13	0.11	0.28	0.39	2	–*	0.41	0.61	0.62
Mean for protein structures obtained by NMR (and its error)	**84.7±7.3**	**21.4±2.4**	**0.01±0.12**	**0.04±0.10**	**0.24±0.11**	**0.28±0.10**	**2.8±0.2**	**0.17±0.10**	**0.24±0.10**	**0.31±0.11**	**0.32±0.11**
Mean for all proteins (and its error)	**86.2±4.8**	**20.2±1.7**	**0.15±0.10**	**0.29±0.08**	**0.32±0.09**	**0.45±0.07**	**2.9±0.2**	**0.27±0.08**	**0.37±0.07**	**0.41±0.07**	**0.46±0.07**

* Correlation coefficient cannot be calculated because the all residues with known experimental Φ-values are outside of the predicted folding nucleus. Under averaging the correlation coefficient for these cases were taken as 0.

Table 2. The level of correlation for X-ray proteins

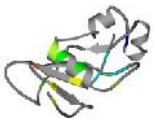
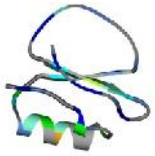
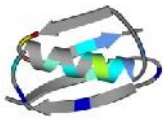
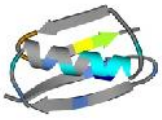
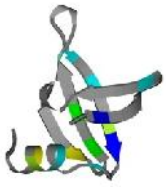
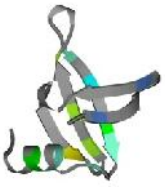
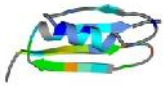
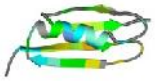
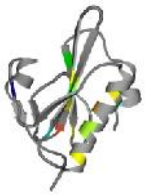
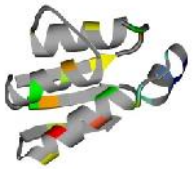
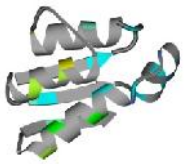
PDB	number of residues	number of experi-mental points	structure with experimental Φ - values[a]	structure with theoretical Φ - values[a]	coefficient of correlation	
					original structure	structure with hydrogens
barnase (1rnb)	109	21			0.59	0.68
CheY (3chy)	128	9			0.65	0.67
CI2 (2ci2)	64	41			0.41	0.50
protein G (1pgb)	56	20			0.74	0.76
SH3 domain of α-spectrin (1shg)	57	13			0.81	0.82
tenascin (1ten)	90	30			0.54	0.64
sso7D (1bf4)	63	15			0.58	0.79
average for X-ray proteins	**81**	**21.3**			**0.62±0.05**	**0.69±0.04**

Table 3. The level of correlation for NMR proteins

PDB	number of residues	number of experi-mental points	structure with experimental Φ - values[a]	structure with theoretical Φ - values[a]	coefficient of correlation	
					structure without hydrogen atoms	original structure
ACBP (2abd)	86	14			-0.02	-0.10
FKBP (1fkr)	107	29			-0.25	-0.16
protein L (2ptl)	64	41			-0.06	0.25
villin 14T (2vil)	126	18			0.31	0.43
barstar (1btb)	89	22			0.26	0.33
SH3 domain of src tyrosine kinase transforming protein (1srm)	56	26			0.63	0.51
average for NMR proteins	**88**	**25**			**0.15±0.14**	**0.21±0.12**
average for all proteins	**84.2**	**23**			**0.40±0.10**	**0.47±0.09**