

Influence of Organization of Native Protein Structure on Its Folding: Modeling of the Folding of α -Helical Proteins

A. V. Glyakina¹ and O. V. Galzitskaya^{2*}

¹*Institute of Mathematical Problems of Biology, Russian Academy of Sciences,
ul. Institutskaya 4, 142290 Pushchino, Moscow Region, Russia*

²*Institute of Protein Research, Russian Academy of Sciences, ul. Institutskaya 4, 142290 Pushchino,
Moscow Region, Russia; fax: (84967) 318-435; E-mail: ogalzit@vega.protres.ru*

Received March 31, 2010

Revision received April 14, 2010

Abstract—An important question that is addressed here is whether the modeling of protein folding can catch the difference between the folding of proteins with similar structures but with different folding mechanisms. In this work, the modeling of folding of four α -helical proteins from the homeodomain family, which are similar in size, was done using the Monte Carlo and dynamic programming methods. A frequently observed order of folding of α -helices for each protein was determined using the Monte Carlo method. A correlation between the experimental folding rate and the number of Monte Carlo steps was also demonstrated. Amino acid residues that are important for the folding were determined using the dynamic programming method. The defined regions correlate with the order of folding of secondary-structure elements in the proteins both in experiments and in modeling.

DOI: 10.1134/S0006297910080079

Key words: protein folding, folding intermediates, folding nuclei, nucleation mechanism, diffusion–collision mechanism

During experimental investigations in the beginning of the 1990s it was discovered that there are proteins that fold in one stage without accumulation of metastable states in all regions of external conditions [1, 2].

It is this group of one-stage proteins, for which the folding looks especially simple, that researchers analyzed to reveal characteristics of spatial protein structure affecting its folding rate. Plaxco et al. [3] were the first to compare folding rates predicted from protein structures with the experimental rates. It was demonstrated that the folding rates for 12 one-stage proteins anticorrelate on the level of 0.81 with the relative contact order parameter (or simply “contact order” – *CO*) calculated from protein structure, which is equal to normalized for the protein chain length average distance along the chain between atoms contacting each other in the native structure [3-5]:

$$CO = \frac{1}{LN} \sum_{(i < j)}^N \Delta L_{ij}, \quad (1)$$

where *N* is the number of noncovalent contacts (at a limit distance of 6 Å) between non-hydrogen atoms in the pro-

tein, *L* is the number of amino acid residues in the protein chain, and ΔL_{ij} is the number of residues which separate the interacting pair of non-hydrogen atoms (*i* < *j*) (it is assumed that atoms of residues neighboring in the chain are divided by one residue and interacting atoms are separated by no more than 6 Å, i.e. the threshold distance limit; summation is done over all such pairs). This parameter is small for proteins whose structure is stabilized mainly by local contacts (i.e. for α -proteins) and large for proteins in whose native structure there are many contacts distant along the chain (first of all, for β -structured proteins). *CO* was specially invented for comparison of the arrangement (rather than the size) of different proteins; it depends on the secondary structure content in the protein and on the interference of α -helices disposition, β -structural elements, and loops in the native protein structure [4].

However, it turned out that *CO* weakly correlated with folding rates of other proteins (multi-stage proteins forming metastable intermediates of folding). The correlation coefficient was 0.34 for the set of experimentally investigated multi-stage proteins known by 2003 [6], and –0.01 for 82 proteins (one-stage and multi-stage) and two peptides known by 2009 [7]. Therefore, an assumption has been proposed that the success of *CO* in

* To whom correspondence should be addressed.

prediction of folding rates for one-stage proteins is determined by the fact that the one-stage proteins experimentally investigated by 1998 were approximately of the same length. Thus, the topological parameter of CO became the primary one in determining the folding rates, the more so that there is no correlation between the folding rate and the size for one-stage proteins [6].

Therefore, to take into account the size of the protein the following equation was suggested [8]:

$$\ln k_f^w \sim -CO \times L^P, \quad (2)$$

where $\ln k_f^w$ is the logarithm of the folding rate in water. It is possible to obtain a similar equation when considering critically Finkelstein–Badretdinov's theory according to which the protein folding rate at the point of thermodynamic equilibrium is determined by equation $\ln k_f^{mt} \sim -(1 \pm 0.5)L^{2/3}$. Here, the coefficient independent of the protein size given in brackets is close to 0.5 if the protein is stabilized mainly by local interactions, and to 1.5 if the protein has many distant contacts along the chain so that many closed loops protrude from any nucleus of folding. It is clear that this coefficient is similar to CO in physical meaning: both are small for proteins with local contacts (in the first place, α -helical), and both are large for proteins with mainly distant contacts along the chain when the protein could not avoid the formation of closed loops protruding from the folded part of the globule during the folding of the protein. So, we again come to Eq. (2).

The best agreement of Eq. (2) with experimentally investigated folding rates is under $P = 1$. The correlation coefficient is -0.74 for the full set of proteins [8]. The new parameter predicts folding rates for multi-stage proteins at the same level (the correlation coefficient is -0.78) and worse for one-stage proteins (for proteins larger than 40 amino acid residues, the correlation coefficient is -0.51) [8].

During the investigation it was found that CO is proportional to $L^{-0.30 \pm 0.07}$ for the whole set of proteins and peptides. This means that the parameter $CO \times L$, which has the maximal correlation with experiment for all proteins and peptides, depends on the size of protein (length of protein chain) as $L^{0.70 \pm 0.07}$. It is much more than a good coincidence with estimation of Finkelstein and Badretdinov, $L^{2/3}$, and also with the estimation of folding $L^{0.61 \pm 0.18}$ of simplified models of protein chains on the lattice [9].

The smaller the protein the simpler the folding mechanism. Therefore, many experimental and theoretical works have been devoted to folding of small α -helical proteins [10]. Four proteins (En-HD, c-Myb, RAP1, and TRF1) belonging to the homeodomain family, each of them consisting of three α -helices, are studied here. It has been shown experimentally that protein En-HD has complex (multi-stage) folding [11], and the others are one-stage proteins. These one-stage proteins have a more

compact native-like transition state. At that the transition state of En-HD is more ordered than that of c-Myb, and the folding rate of En-HD is higher than the folding rates of other proteins (Table 1) in spite of the fact that it is a multi-stage protein [11].





Modeling of protein dynamics of En-HD, c-Myb, and TRF1 by molecular dynamics simulations has shown [12] that although the structures of transition states of these proteins are native-like, their accessible surface area increases by 17% for En-HD and c-Myb and by 20% for TRF1. Moreover, modeling showed that an intermediate occurs on the pathway of c-Myb folding, and this intermediate has not been detected experimentally. The folding intermediate has been detected experimentally only for special mutants of this protein. No folding intermediate of TRF1 has been detected [12].

In papers devoted to the folding of these proteins, it has been suggested that the protein folding rate is connected with predisposition of amino acid residues in the protein to form secondary structure. It has turned out that the probability of formation of a helical state calculated using the Agadir program (<http://agadir.crg.es>) [13] is larger for En-HD and smaller for TRF1, proteins c-Myb and RAP1 being situated between them. This correlates with the folding rates: En-HD folds faster than all the other considered proteins, and TRF1 folds slower. It has also turned out that the existence of an intermediate on the folding pathway only accelerates the folding [12].

Two mechanisms of protein folding are distinguished in the papers, nucleation–condensation and diffusion–collision (framework model) [14–16], which are different representations of the common mechanism of protein folding. Under the nucleation–condensation mechanism secondary and tertiary protein structures are formed simultaneously, but under the diffusion–collision mechanism of folding the elements of secondary structures are formed first, which then are arranged in space forming a tertiary structure. The nucleation–condensation mechanism is found when secondary structure is not stable in the absence of tertiary interactions, while the diffusion–collision mechanism becomes more probable with increasing stability of secondary structure [12].

With increasing probability of secondary structure formation, the folding mechanism shifts from the nucleation–condensation to the diffusion–collision (framework model). Thus, the folding of En-HD is described by the framework model. While the folding of c-Myb is described by mixed framework and nucleation–condensation models, the folding of TRF1 has a purely nucleation–condensation mechanism. The common property under folding/unfolding of these proteins is that their transition states are very native-like [12]. But the energetic balance of interactions and pathways for achieving the transition state usually depends on the predisposition of amino acid residues to form tertiary and secondary structures. In this work the folding process and structural

Table 1. Values of free energy barriers, the time during which half of trajectories fold, the experimental folding rates in water and at the point of equilibrium, contact order, flexible regions, and helical propensity for proteins En-HD, c-Myb, RAP1, and TRF1

Name of protein (PDB entry), size	F/RT , value of free energy barrier	$t_{1/2}$, number of Monte Carlo steps	$\ln k_f$ in water/ $\ln k_f^{mt}$ at point of equilibrium	Contact order (CO)/ $Abs(CO)$ /radius of cross-section, V_{ASA}/S_{ASA} , Å	Flexible regions	Helicity, % (Agadir)/Helicity, % (X-ray)
En-HD(1enh), 54 	6.96	37 828	10.5/8.1	13.63/736.2 V_{ASA}/S_{ASA} 3.49	— 0 residues	13.88/70.37
c-Myb(1gv2), 47 	7.72	43 161	8.7/3.1	12.3/578.1 V_{ASA}/S_{ASA} 3.37	5-9 32-38, 12 residues	2.09/63.83
RAP1(1fex), 59 	4.86	71 215	8.2/3.9	13.04/769.6 V_{ASA}/S_{ASA} 3.42	19-30 37-41, 17 residues	1.45/52.54
TRF1(1ba5), 49 	9.11	79 791	5.9/1.6	14.52/711.6 V_{ASA}/S_{ASA} 3.45	10-15 21-26, 12 residues	1.33/65.31

properties (connected with folding) of these four proteins have been investigated using Monte Carlo and dynamic programming methods. It has been shown that Φ -values calculated by dynamic programming for all proteins correlate with the order of folding of secondary structure elements of these proteins obtained by the Monte Carlo method. Moreover, a good correlation has been demonstrated between experimental folding rates at the point of thermodynamic equilibrium and helical propensities (0.94) and the number of Monte Carlo steps (−0.71), and there is absence of correlation with such parameter as contact order (−0.09).

METHODS OF INVESTIGATION

Test subjects of this work are four proteins each consisting of three α -helices: En-HD (PDB entry 1enh, amino acid residues from 3 to 56), c-Myb (PDB entry 1gy2, residues from 144 to 190), RAP1 (PDB entry 1fex, residues from 1 to 59), and TRF1 (PDB entry 1ba5, residues from 5 to 53). The types of organisms in which these proteins are found are different. Protein En-HD is from *Drosophila melanogaster* and is related to the *Arthropoda* type, but proteins c-Myb, RAP1, and TRF1 are from organisms related to the *Chordata* type.

Helical propensity of proteins has been calculated using the Agadir program (<http://agadir.crg.es>) [13] at pH = 7, ionic strength 0.1, and temperature $T = 300$ K.

Flexible regions have been calculated using the FoldUnfold program (<http://omega.protres.ru/~mlobanov/ogu/ogu.cgi>) [17] with an averaging window of five residues.

The radius of cross-section, contact order, and absolute contact order have been calculated by us earlier and has been taken from the web site <http://phys.protres.ru/resources/compact.html>.

Estimation of free energy. Let us consider the process of sequential folding/unfolding of native structure of a

protein chain consisting of L links (Fig. 1). This protein chain has a fully folded native state I_0 , a fully unfolded state I_L , and an ensemble of partly unfolded intermediate structures I_ν , including ν disordered fragments, and a native-like globular part with $L - \nu$ links ($\nu = 0$ for native state I_0 , $\nu = L$ for fully unfolded state I_L , $\nu = 1, \dots, L - 1$ for partly unfolded structures). Structures with non-native-like globular parts are not considered here.

Free energy of structure I is represented by the equation:

$$F(I) = n_I \times \varepsilon - T[\nu_I \times \sigma + \sum_{\text{loop} \in I} S_{\text{loop}}], \quad (3)$$

where n_I is the number of atom–atom contacts in the native-like part of structure 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 the 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 [18], where R is the gas constant); S_{loop} is described by Eq. (5) (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 are in contact if the distance between their centers is not more than 6 Å. When modeling with account of 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 $n_0 \times \varepsilon = TN\sigma$, where n_0 is the number of contacts in the native structure and N is the total number of amino acid residues in the protein.

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

$$\varepsilon = -TN\sigma/n_0. \quad (4)$$

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

The entropy spent to close a disordered loop protruding from the globule between the still fixed residues k and l is estimated [19] as:

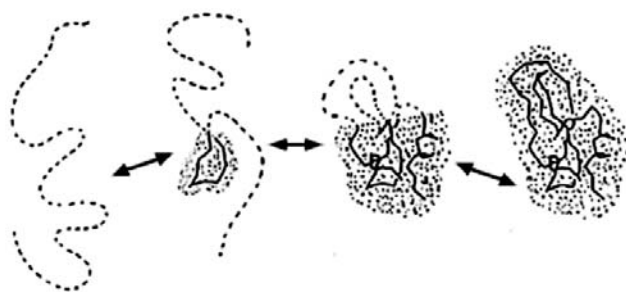


Fig. 1. One of the pathways of sequential unfolding and folding of native 3D protein structure (I_0). I_L is the coil where all L links of the protein chain are disordered. In each of various intermediates of type I_ν , the ν chain links (shown by a dashed line) are unfolded, while the other $L - \nu$ links keep their native positions and conformations (they are shown as a solid line).

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

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

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

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

$$\Phi_f = \frac{\Delta \ln k_f}{\Delta \ln K}, \quad (6)$$

where k_f is the folding rate constant, $K = k_f/k_u$ is the folding–unfolding equilibrium constant, and Δ is the shift of the corresponding value induced by mutation of this residue. According to the model of a native-like folding nucleus [21, 22], $\Phi_f = 1$ means that the residue has its native conformation and environment already in the transition state (i.e. this residue is in the folding nucleus), while $\Phi_f = 0$ means that the residue remains unfolded in the transition state. The values $\Phi_f \approx 0.5$ are ambiguous: either the residue is at the surface of the nucleus, or it is in one of the alternative nuclei belonging to different folding pathways. It is noteworthy that the values $\Phi_f < 0$ and $\Phi_f > 1$ (which would be inconsistent with the model of a native-like folding nucleus) are extremely rare and never concern a residue with reliable measured $\Delta \ln K$.

According to Eqs. (3) and (6), the value $\Delta \ln K = \Delta_r[F(I_0) - F(I_L)]$ is equal to $\varepsilon \times \Delta_r(n_0^{nb})$, where ε is the contact energy, and $\Delta_r(n_0^{nb})$ is the residue r mutation-induced change in the number of contacts in the native state I_0 (since all native contacts are assumed to be equal, and no contacts are assumed to be present in the unfolded state I_L).

Correspondingly, $\Delta \ln k_f = \Delta_r[F(TS) - F(I_L)] = \varepsilon \times \Delta_r(n_I^{nb})$, where $\Delta_r(n_I^{nb})$ is the same residue r mutation-induced change in the number of native contacts in the transition state, averaged over the transition state ensemble $\{I^\#\}$. This change can be calculated as:

$$\Delta_r(n_I^{nb}) = \sum P(I^\#) \Delta_r(n_{I^\#}^{nb}), \quad (7)$$

where $P(I^\#)$ is the Boltzmann probability of microstate $I^\#$ in the transition state ensemble (see Eq. (8)) and $\Delta_r(n_{I^\#}^{nb})$ is the residue r mutation-induced change in the number of native contacts in microstate $I^\#$.

The values $\Delta_r(n_{I^\#}^{nb})$ can be calculated for each microstate I from atomic coordinates of a non-mutated

protein when we know what atoms are deleted or substituted in the mutant. However, this calculation assumes that the protein structure is not disturbed by mutation. Therefore, we have to consider only those mutations that do not insert new atomic groups. In this study we considered mutations by glycine for each residue.

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

The Φ -values were computed by the following equation:

$$\Phi = \frac{\langle \Delta_r(n_I^{nb}) \rangle_{I^\#}}{\Delta_r(n_0^{nb})}. \quad (8)$$

The effective value of free energy barrier was calculated by the following equation:

$$F_{\text{eff}} = -RT \ln \sum_{I^\#} \exp(-F_{I^\#}/RT), \quad (9)$$

where $F_{I^\#}$ is the value of free energy barrier of transition state I , and summation is done over the complete ensemble of transition states.

RESULTS AND DISCUSSION

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 the 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) as in the thermodynamic [18] and kinetic [23, 24] experiments, i.e. only two states of protein molecule (native and denatured) are observed, but intermediate states are not observed to a sufficient extent. In other words, under these conditions native and denatured states have (by definition at the point of thermodynamic equilibrium) equal free energies and the other states (including intermediates) are destabilized. Therefore, by modeling the protein folding and unfolding process at the point of thermodynamic equilibrium we can disregard misfolding structures (which in other conditions can be peculiar “dead ends”, traps that in principle are able to strongly affect the folding and unfolding of the protein molecule). According to the principle of detailed balance [25], pathways of direct and reverse reactions coincide if both reac-

tions occur under the same conditions. Consequently, we can represent the processes of protein folding and unfolding at the point of thermodynamic equilibrium as a reversible process of folding/unfolding.

Using the dynamic programming method we are able to analyze the full network of folding/unfolding pathways although presented in a rather rough resolution (the state includes no more than two loops; the chain link consists of several amino acid residues). But using the Monte Carlo method we are able to analyze separate folding pathways without restrictions inherent in the dynamic programming method.

As a basis we take the three dimensional structure of the protein in the native state from the database of spatial protein structures in PDB [26]. 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 represented as a simplified sequential unfolding of the protein (Fig. 1).

Each step on the unfolding pathway represents the removing of one chain link from the native spatial structure of the protein (the "link" can consist both of one amino acid residue and several residues along the chain without a break). The removed links are assumed to form an indigested coil, i.e. they lose all the non-bonded interactions and gain coil entropy excluding its relatively small part [19] 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 states, i.e. on the stability (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 the "links" consisting of not a single 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 for the dynamic programming method. 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 place the residue in its native position according to the coordinates from the protein data bank (see below "Modeling of folding by the Monte Carlo method").

Modeling of folding by the Monte Carlo method. The process of protein folding is modeled as "traveling" of one protein molecule along the network of folding/unfolding pathways [27]. The start is performed from the fully unfolded state; the finish corresponds to the native structure of the protein (i.e. a 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. At that, if the residue was in the native position it should "unfold", if it was unfolded it should "fold". The changing of free energy in the case of such an elementary step is calculated. If the free energy decreases such an elementary step is accepted; if the free energy grows such an elementary step is accepted with probability $\exp[-\Delta F/RT]$ (standard criterion of Metropolis [28]). For each of the four proteins (En-HD, c-Myb, RAP1, TRF1) 50 runs were performed with 10^8 steps.

The Monte Carlo method allows us to obtain the folding time of a separate molecule determined in Monte Carlo steps (see Fig. 2). As theoretically calculated time of protein folding we used time ($t_{1/2}$), for which half of the molecules fold (i.e. 25 from 50 performed runs resulted in the native structure formation [29]).

The typical folding kinetics for proteins traced using the Monte Carlo method is presented in Fig. 2. The protein molecule starts from a fully unfolded state, and for a sufficiently long time stays near the unfolded state, then it overcomes the free energy barrier and then quickly folds completely.

The folding pathways of the four proteins were investigated in detail. The profile of free energy and the order of folding of separate elements of secondary structure were obtained for each trajectory (Fig. 3).

From Table 1 one can see that the time during which half of the trajectories of protein En-HD folds during modeling is less than the folding time of the other proteins considered. These results correlate with the experimental data on the folding rates of these proteins in water and at the point of mid-transition (Table 1). The correlation coefficient between the experimental folding rate at the point of equilibrium (in water) and the number of Monte Carlo steps is -0.71 (-0.89).

Folding pathways for the four proteins obtained using the Monte Carlo method are different. A more often observed folding pathway for protein En-HD is as follows: first the middle α -helix folds, then the C-terminal helix, and the last is the N-terminal helix (in 29 cases, Fig. 3a), or the C- or N-terminal helices fold simultaneously (in 10 cases, Fig. 3b). It should be noted that the most frequent folding pathway of this protein during modeling with the Monte Carlo method corresponds to the observed folding pathway in the experiment with the folding intermediate consisting of the middle and C-terminal α -helices and the turn between them [12]. In protein TRF1 the middle α -helix folds first, then the N-terminal helix (in 15 cases, Fig. 3c) or the middle and N-terminal helices fold simultaneously (in 13 cases, Fig. 3d), and the last is the C-terminal helix (in both cases). In protein RAP1 the C-terminal α -helix folds first, then the middle α -helix (in 24 cases, Fig. 3e) or the middle and the C-terminal α -helices fold simultaneously (in 11

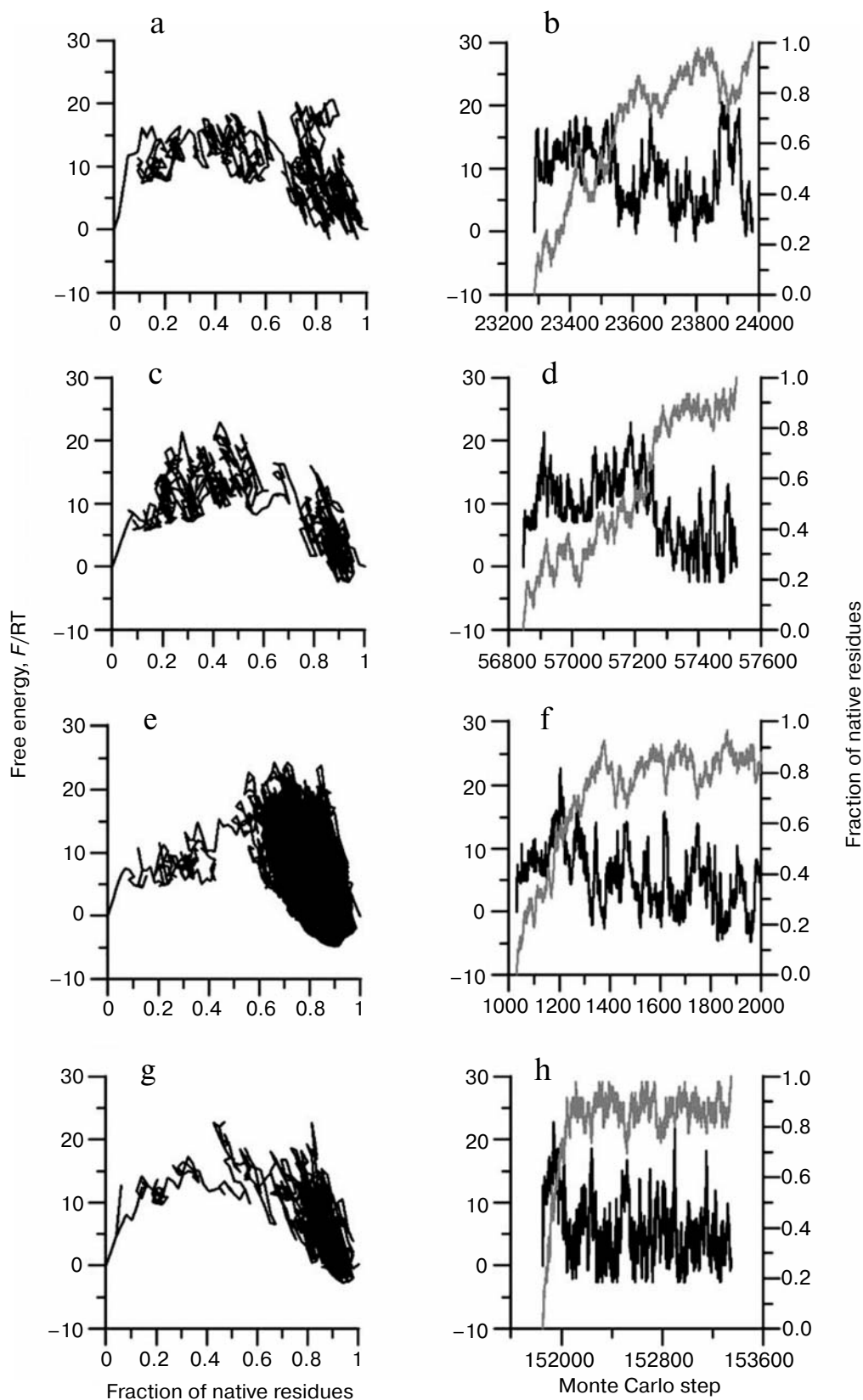


Fig. 2. Dependences of free energy on the fraction of folded amino acid residues (a, c, e, g), and free energy (black curve) and the fraction of native amino acid residues (gray curve) on the number of Monte Carlo step (b, d, f, h) for α -helical proteins: a, b) En-HD; c, d) c-Myb; e, f) RAP1; g, h) TRF1. (For the best representation on graph (f) only the initial 1000 (from 46946) Monte Carlo steps are demonstrated.)

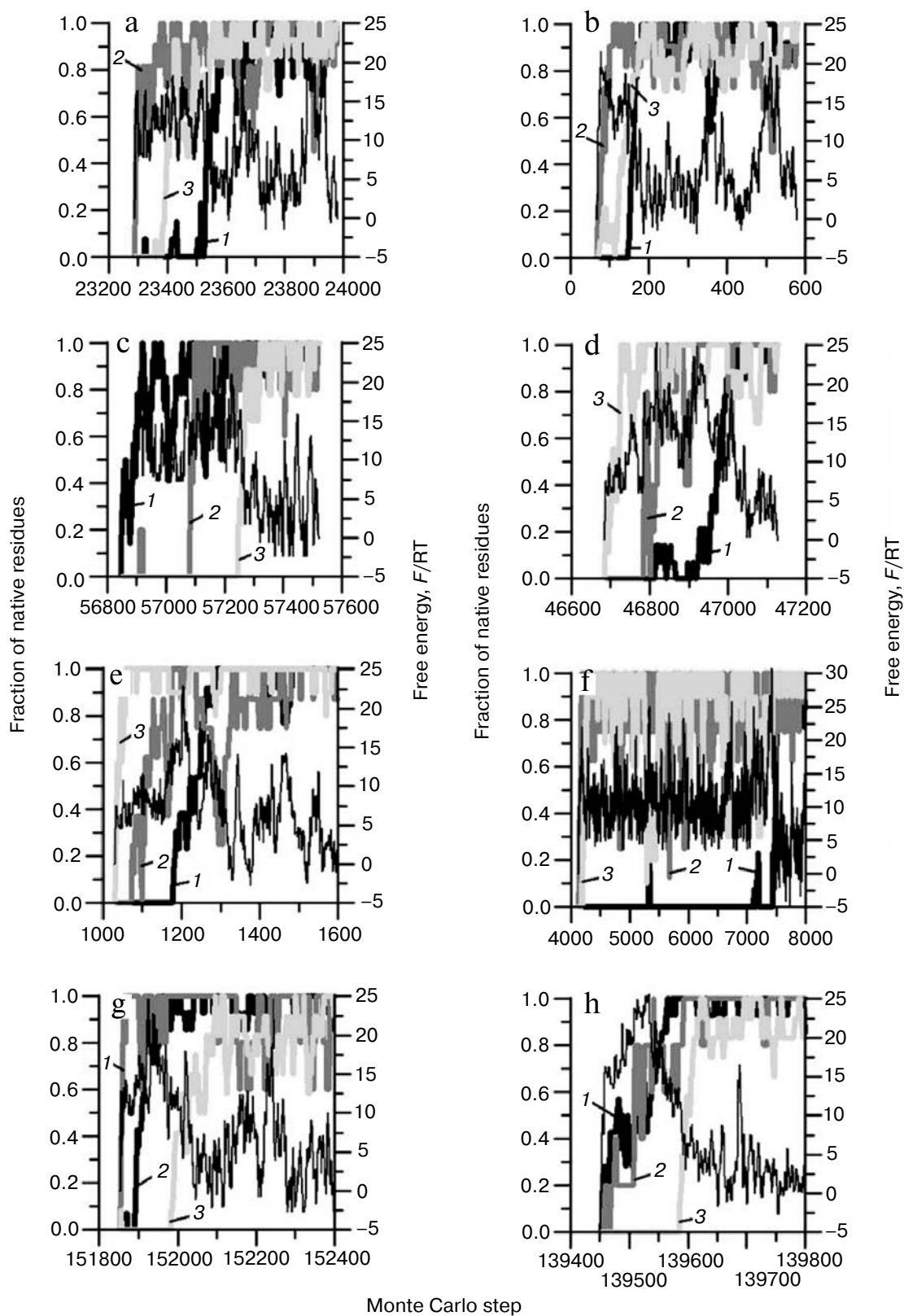


Fig. 3. Frequently observed folding pathways of α -helical proteins: a, b) En-HD; c, d) c-Myb; e, f) RAP1; g, h) TRF1. The graphs show fractions of amino acid residues that are fixed in the native positions for each secondary structure element (1-3 – the first, second, and third α -helix, respectively) and free energy (thin black curve).

cases, Fig. 3f), and last the N-terminal helix folds (in both cases). In protein c-Myb one can choose two opposite folding pathways. In the first case, the N-terminal helix folds first, then the middle helix, and the last is the C-terminal helix (in 16 trajectories, Fig. 3g). In the second case, the C-terminal helix folds first, then the middle one, and the last is the N-terminal helix (in 12 trajectories, Fig. 3h).

Search for the folding nuclei. The folding nuclei have been calculated for the four proteins considered here by the dynamic programming method (Fig. 4). Φ -values for these proteins correlate with the order of folding of secondary structure elements (Table 2 and Fig. 4). The middle α -helix folds first in protein En-HD, and the largest Φ -values for amino acid residues from this helix are observed in comparison with the terminal α -helices. The middle α -helix folds first in protein TRF1 or the middle and the N-terminal helices fold simultaneously, and on the profile of Φ -values one can see that the high average Φ -values are observed (0.62 for the middle α -helix and 0.35 for the N-terminal helix) in comparison with the C-terminal helix. The C-terminal α -helix folds first in protein RAP1, or the C-terminal and the middle α -helices fold simultaneously, and for these helices the high average Φ -values are observed (0.70 for the middle α -helix and 0.49 for the C-terminal one) in comparison with the N-terminal helix. Two completely opposite folding trajectories are observed in most cases for protein c-Myb. In one case the N-terminal helix folds first, in other cases the C-terminal one. But the highest Φ -values are observed for amino acid residues from the middle α -helix. This is due to the fact that Φ -values are average characteristics from which it is not always possible to choose parallel folding pathways. Correlation between experimental [12] and calculated Φ -values for protein En-HD is 0.52 (13 experimental points) and 0.37 for protein c-Myb (18 experimental points). It is worth noting that the value of the free energy barrier for these proteins obtained using the dynamic programming method correlates with the experimental folding rates at the point of thermodynamic equilibrium only at the level of 46%.

Correlation between folding rate and structural parameters. The amino acid sequences for the investigated proteins have been aligned using the BLAST program (<http://blast.ncbi.nlm.nih.gov>) and are shown in Fig. 5. From the figure one can see that the loop connecting helices 1 and 2 in protein RAP1 (1fex) is longer (11 amino acid residues) than in the remaining proteins (5 amino acid residues for En-HD, 3 amino acid residues for c-Myb, and 4 amino acid residues for TRF1). The small differences in the structure of these three α -helical proteins, which in general consist in the difference of angles between helices 1 and 3, have been revealed by superposition of their structures [12].

Flexible regions have been calculated using the FoldUnfold program [17]. These regions and the number

Table 2. Φ -values averaged over amino acid residues included in the secondary structure element

Name of protein (PDB entry)	Average Φ -value		
	N-terminal α -helix	middle α -helix	C-terminal α -helix
En-HD(1enh)	0.10	0.75	0.24
c-Myb(1gv2)	0.38	0.63	0.33
RAP1(1fex)	0.05	0.70	0.49
TRF1(1ba5)	0.35	0.62	0.07

of residues involved in these regions are presented in Table 1. One can see that the appearance of flexible regions results in decreasing of the folding rate and even in changing of the folding mechanism. It should be noted that the number of flexible regions increases as the protein origin becomes more complex (En-HD has a simpler origin (*Arthropoda*), than the other proteins (*Chordata*), but it folds faster).

Moreover, the good correlation between the experimental folding rate at the point of thermodynamic equilibrium (or in water) and helical propensity predicted by the Agadir program [13] has been demonstrated. The correlation coefficient between the experimental folding rate at the point of thermodynamic equilibrium (and in water) and the calculated helical propensity is 0.94 (0.79). It is only -0.09 for the contact order at the point of thermodynamic equilibrium and -0.48 in water. The correlation with absolute contact order is -0.01 at the point of thermodynamic equilibrium and 0.32 in water. The same low correlation is also found for the cross-section radius (see Table 1). Although it is difficult to judge about the presence or absence of any correlation using four points, one would rather speak about a tendency.

Modeling of folding of proteins close in size using the Monte Carlo and dynamic programming methods has shown that protein which folds with accumulation of the intermediate state indeed folds an order faster than their structural homologs. Φ -values calculated by the dynamic programming method for all proteins correlate with the order of folding of secondary structure elements obtained by the Monte Carlo method. Besides, a good correlation has been demonstrated between experimental folding rates at the point of thermodynamic equilibrium and the helical propensity calculated using the Agadir program (0.94) and the number of Monte Carlo steps (-0.71) for these proteins. The correlation between experimental and Φ -values for En-HD protein is 0.52 (13 experimental points) and 0.37 (18 experimental points) for c-Myb protein.

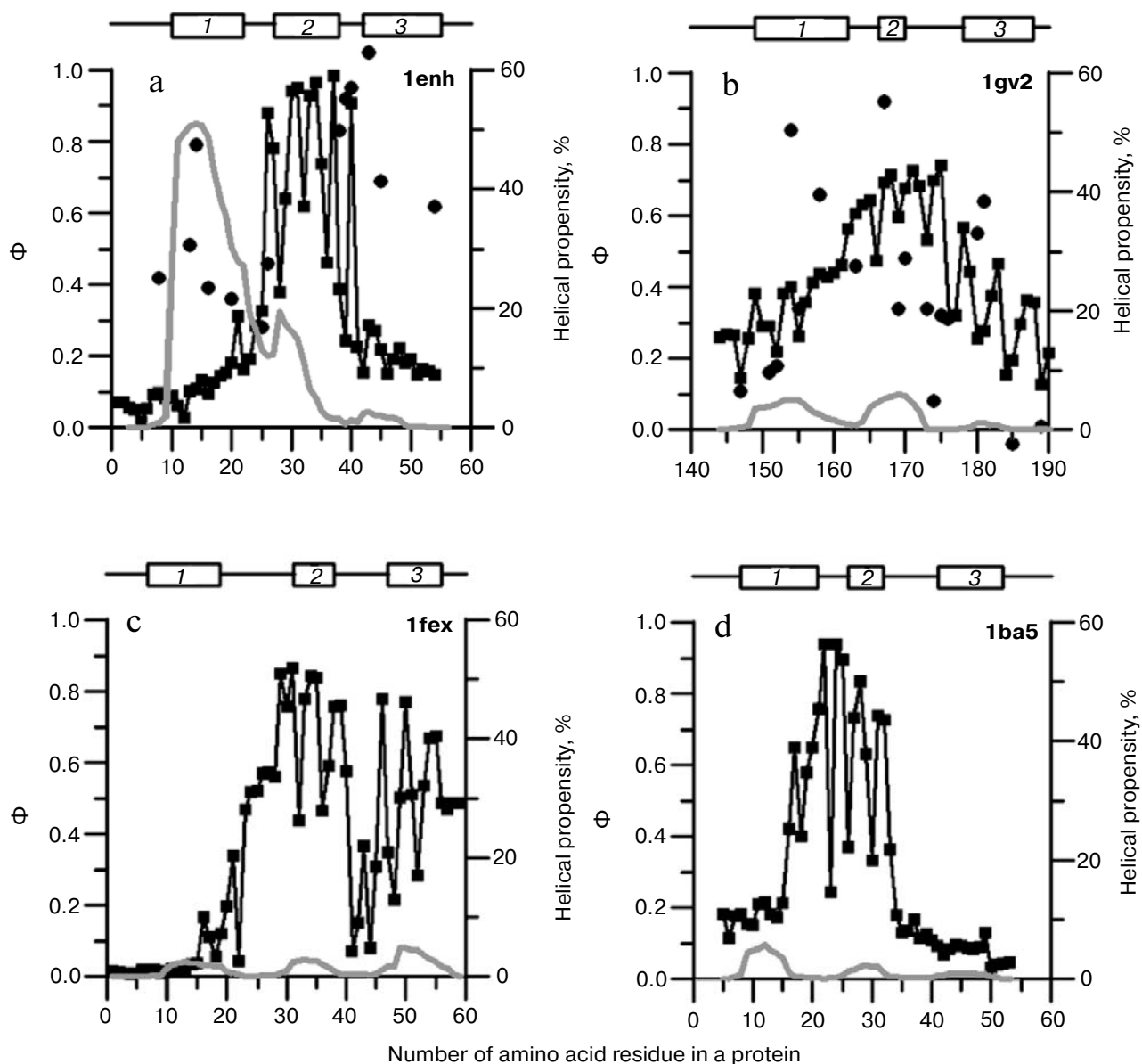


Fig. 4. Theoretical Φ -values (black curve) calculated using the method of dynamic programming and helical propensity (gray curve) for proteins: a) En-HD; b) c-Myb; c) RAP1; d) TRF1. Φ -values were calculated with the assumption that each amino acid residue was substituted by glycine. Black dots on graphs (a) and (b) are the experimental Φ -values, which were taken from [12]. Correlation between experimental and calculated Φ -values for protein En-HD is 0.52 (13 experimental points) and for protein c-Myb – 0.37 (18 experimental points). Helical propensity of proteins was calculated using the Agadir program (<http://agadir.crg.es>).



Fig. 5. Alignment of amino acid sequences of proteins En-HD, c-Myb, RAP1, and TRF1 using the BLAST program (<http://blast.ncbi.nlm.nih.gov>). Helical regions are defined based on X-ray data.

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

REFERENCES

- Jackson, S. E., and Fersht, A. R. (1991) *Biochemistry*, **30**, 10428-10435, 10436-10443.
- Jackson, S. E. (1998) *Fold. Des.*, **3**, R81-R91.
- Plaxco, K. W., Simons, K. W., and Baker, D. (1998) *J. Mol. Biol.*, **277**, 985-994.
- Plaxco, K. W., Gujjarro, J. I., Morton, C. J., Pitkeathly, M., Campbell, I. D., and Dobson, C. M. (1998) *Biochemistry*, **37**, 2529-2537.
- Grantcharova, V., Alm, E. J., Baker, D., and Horwich, A. L. (2001) *Curr. Opin. Struct. Biol.*, **11**, 70-82.
- Galzitskaya, O. V., Garbuzynskiy, S. O., Ivankov, D. N., and Finkelstein, A. V. (2003) *Proteins*, **51**, 162-166.
- Ivankov, D. N., Bogatyreva, N. S., Lobanov, M. Yu., and Galzitskaya, O. V. (2009) *PLoS ONE*, **4**, e6476.
- Ivankov, D. N., Garbuzynskiy, S. O., Alm, E., Plaxco, K. W., Baker, D., and Finkelstein, A. V. (2003) *Protein Sci.*, **12**, 2057-2062.
- Koga, N., and Takada, S. (2001) *J. Mol. Biol.*, **313**, 171-180.
- Dyer, R. B. (2007) *Curr. Opin. Struct. Biol.*, **17**, 38-47.
- Mayor, U., Guydosh, N. R., Johnson, C. M., Grossmann, J. G., Sato, S., Jas, G. S., Freund, S. M., Alonso, D. O., Daggett, V., and Fersht, A. R. (2003) *Nature*, **421**, 863-867.
- Gianni, S., Guydosh, N. R., Khan, F., Caldas, T. D., Mayor, U., White, G. W. N., DeMarco, M. L., Daggett, V., and Fersht, A. R. (2003) *Proc. Natl. Acad. Sci. USA*, **100**, 13286-13291.
- Munoz, V., and Serrano, L. (1994) *Nature Struct. Biol.*, **1**, 399-409.
- Islam, S. A., Karplus, M., and Weaver, D. L. (2002) *J. Mol. Biol.*, **318**, 199-215.
- Karplus, M., and Weaver, D. L. (1994) *Protein Sci.*, **3**, 650-668.
- Itzhaki, L. S., Otzen, D. E., and Fersht, A. R. (1995) *J. Mol. Biol.*, **254**, 260-288.
- Galzitskaya, O. V., Garbuzynskiy, S. O., and Lobanov, M. Y. (2006) *Bioinformatics*, **22**, 2948-2949.
- Privalov, P. L. (1979) *Adv. Protein Chem.*, **33**, 167.
- Galzitskaya, O. V., and Finkelstein, A. V. (1999) *Proc. Natl. Acad. Sci. USA*, **96**, 11299-11304.
- Flory, P. J. (1969) *Statistical Mechanics of Chain Molecules*, Interscience, New York.
- Matouschek, J. T., Kellis, Jr., Serrano, L., and Fersht, A. R. (1989) *Nature*, **340**, 122-126.
- Matouschek, J. T., Kellis, Jr., Serrano, L., Bycroft, M., and Fersht, A. R. (1990) *Nature*, **346**, 440-445.
- Fersht, A. R. (1995) *Curr. Opin. Struct. Biol.*, **5**, 79-84.
- Fersht, A. R. (1997) *Curr. Opin. Struct. Biol.*, **7**, 3-9.
- Landsberg, P. T. (1971) *Problems in Thermodynamics and Statistical Physics*, PION, London.
- 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.
- Galzitskaya, O. V., Surin, A. K., and Nakamura, H. (2000) *Protein Sci.*, **9**, 580-586.
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953) *J. Chem. Phys.*, **21**, 1087-1092.
- Galzitskaya, O. V., and Finkelstein, A. V. (1995) *Protein Eng.*, **8**, 883-892.