ISSN 0006-2979, Biochemistry (Moscow), 2011, Vol. 76, No. 2, pp. 225-235. © Pleiades Publishing, Ltd., 2011.
Original Russian Text © O. V. Galzitskaya, N. S. Bogatyreva, A. V. Glyakina, 2011, published in Biokhimiya, 2011, Vol. 76, No. 2, pp. 274-286.

Bacterial Proteins Fold Faster than Eukaryotic Proteins with Simple Folding Kinetics

O. V. Galzitskaya[1]*, N. S. Bogatyreva[1], and A. V. Glyakina[2]

[1]*Institute of Protein Research, Russian Academy of Sciences, Institutskaya ul. 4, 142290 Pushchino, Moscow Region, Russia; fax: (4967) 318-435; E-mail: ogalzit@vega.protres.ru*
[2]*Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Institutskaya ul. 4, 142290 Pushchino, Moscow Region, Russia*

Received April 28, 2010
Revision received September 16, 2010

Abstract—Protein domain frequency and distribution among kingdoms was statistically analyzed using the SCOP structural database. It appeared that among chosen protein domains with the best resolution, eukaryotic proteins more often belong to α-helical and β-structural proteins, while proteins of bacterial origin belong to α/β structural class. Statistical analysis of folding rates of 73 proteins with known experimental data revealed that bacterial proteins with simple kinetics (23 proteins) exhibit a higher folding rate compared to eukaryotic proteins with simple folding kinetics (27 proteins). Analysis of protein domain amino acid composition showed that the frequency of amino acid residues in proteins of eukaryotic and bacterial origin is different for proteins with simple and complex folding kinetics.

DOI: 10.1134/S000629791102009X

Key words: "all-or-none" transition, eukaryotic and bacterial proteins, folding intermediates, folding rate, protein folding

Stable spatial structure of biological molecules lasting for a biologically reasonable period is necessary for their normal functioning. The folding times of different proteins differs over many orders of magnitude, from microseconds to seconds and even hours. Small proteins usually fold quickly and without folding intermediates (i.e. it is a single-stage process) and "single-stage" kinetics is observed, while larger proteins fold at a lower rate, and metastable intermediates are often observed, i.e. folding is a multistage process and multistage kinetics are observed [1].

Studies of the dependence of protein chain folding rate on the chain length showed that the folding time should grow with the protein chain elongation, and this process is slower that exp(L). Such studies were carried out for different parameters associated with the length of a protein chain ($L^{2/3}$ [2], ln$L$ [3], $L^{1/2}$ [4], $L^{0.61}$ [5]). Concerning prediction of folding rate, all proposed folding rate dependences on the number of residues in a protein chain approximately equally correlate with observed protein folding rates: the coefficient of correlation is ~70% [5-7].

We have shown that the length of a protein chain first of all defines folding rate of multistage proteins [8], the coefficient of correlation being about −0.80 for dependence $\ln k_f^w \sim -L^P$, where P is an exponent ($0 \le P \le 1$), whereas no correlation was found for one-stage proteins (the coefficient of correlation was −0.07 [8]).

Thus, it was concluded on the basis of analytical theories [2, 3] and computer experiments [4, 5] that the length of the protein chain is one of the main factors that define rates of protein folding. This dependence helps one understand the reason for difference in folding rates of proteins of significantly different size, but it says little about why proteins of approximately equal size often exhibit different folding rates.

Besides protein size, there are factors which may additionally influence the rate of protein folding. Thus, it follows from the nucleation mechanism that topology of a transition state, the protein chain path in space, should be the same as in the native structure [9]. This means that the more contacts between remote residues there are in the chain, the more probable is it that in the transition state the protein will be not able to have loops escape from the native-like protein part, and folding will be slower. Just this is observed upon comparison with experiment: for proteins of approximately the same size, the folding rate logarithm decreases as the parameter of "contact order" (*CO*), which is equal to the average distance (in

* To whom correspondence should be addressed.

amino acid residues) between atoms along the chain contacting in the native structure and normalized for the number of amino acid residues in the protein chain, increases [10]. However, topology in itself cannot explain the difference in folding rates for some proteins with the same protein chain path (SH3 domains, cold shock proteins, fibronectin domains, as well as proteins of ferredoxin stacking) [11-15]. Since *CO* is independent of the protein size, the rate of protein folding cannot be predicted using this parameter.

The effect of protein shape on its folding rate was detected in works [16-19]. Under otherwise equal conditions, spherical proteins, independently of the way of their folding, cannot escape a large area of interphase between structured and unstructured parts (Fig. 1); oblong proteins are able to "choose" such a way for that the protein folding passes through a small area of the border and therefore, through a lower barrier. Therefore, more spherical proteins should fold more slowly. It was shown in works [18, 19] that the α/β class proteins are on average more "spherical" than those with a different type of secondary structure packing. It was also shown that α/β proteins, on average again, fold more slowly than the same length proteins of different structural classes. Since under otherwise equal conditions for more spherical proteins, like α/β proteins, the interphase surface in the transition state is larger, then their folding is a slower process. We found that the compactness parameters, associated with dimensions of globule cross section, on the average rather well predict rates of protein folding and unfolding as well as folding rates at the point of thermodynamic equilibrium. However, compactness parameters not associated with the protein size, which describe the shape of the protein globule, are to a significantly lower extent suitable for prediction of the rates of protein folding and unfolding [19].

During the past few years many parameters have been proposed for prediction of the rate of protein chain folding [8, 16, 20-27]. Thus, it was found that different folding rate dependences on amino acid composition [27] and on physicochemical properties of amino acids [25, 28] are observed for proteins with single-stage and multistage folding kinetics in water. Amino acids phenylalanine (F) and glycine (G) influence folding rate of proteins with simple folding kinetics, while cysteine (C), histidine (H), leucine (L), and arginine (R) influence folding rates of proteins with complex folding kinetics [27]. The authors do not explain why just these amino acids are so important for folding of protein chains. Will this dependence be retained for proteins of eukaryotic and bacterial origin?

Questions concerning evolution of protein structures have been discussed in many works [29], but nothing is known about the evolution of folding rates of protein structures. We now know that "relatively" rapid folding to the correct stable structure (ignoring the class of natively unfolded proteins) is important for protein molecules. As shown in a recent work [29], evolutionary selection of structures takes place at the level of mutations that still allow the protein chain to fold to a correct structure.

The necessity to preserve structural and functional integrity of an evolving protein strictly restricts the set of admissible amino acid substitutions. The limits of protein evolution were studied in a recent paper about evolution of protein sequences using data on sequence divergence

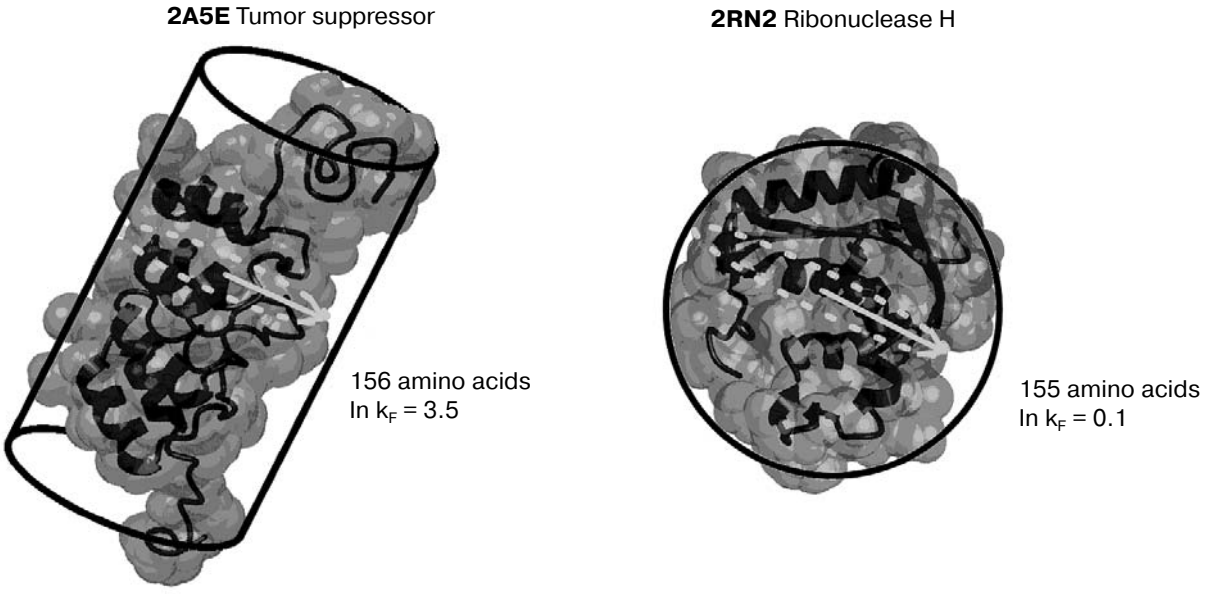

Fig. 1. An example of two proteins with oblong (PDB code 2A5E) and spherical surface (PDB code 2RN2) of the same size (number of amino acid residues) but with different folding rate.

[30]. The authors showed that ancient proteins still diverge from each other, which points to continued expansion of the protein sequence "universe". Evolution of protein sequences, in turn, leaves its trace on evolution of the rates of protein folding.

In our work we considered folding of separate protein molecules and their structural properties associated with folding. Folding rate dependence on protein size, number of stages, and origin were statistically analyzed. Presently there is not enough experimental material for unambiguous conclusions, but it is already possible to draw preliminary concepts on the basis of the results. Statistical analysis of folding rates of 73 proteins with known experimental data has shown that proteins of eukaryotic organisms (27 proteins) folding according to the all-or-none mechanism on average fold more slowly compared to bacterial proteins (23 proteins) following the same mechanism. In the case of proteins of eukaryotic and bacterial origin, the frequency of amino acid residues was different for proteins with simple and complex folding kinetics.

METHODS OF INVESTIGATION

Experimental data. We have statistically analyzed the structural database version 1.65 of SCOP proteins [31]. Upon database compilation, the main criterion for selection for us was the quality of the protein domain structure (structure with good resolution, the least number of ligands) rather than its origin. A total of 3337 protein domains belonging to four main structural classes (*a*-*d*) with identity below 25% were found, including 727 α proteins of *a* class, 816 β proteins of *b* class, 942 α/β proteins of class *c*, and 852 (α+β) proteins of class *d*.

After subdivision of protein domains according to their origin, we obtained 1446 protein domains of bacterial origin, 1545 eukaryotic domains, 154 domains of Archaeans, and 192 virus domains.

Since in this work protein folding kinetics will be considered, the following items will attract our attention: $\ln k_f$, logarithm of constant of the rate of protein chain folding in water. Index "two" of the rate constant means that the protein folding kinetics is described by two-state kinetics, and in this case one stage is observed and the protein is called "single stage". Index "multi" means that three or more states of the protein chain are necessary for kinetic description, and in this case more than one stage is observed and the protein is called "multistage". When the whole protein aggregate is considered without division into folding types, we shall omit the vicissitude index in the logarithm of the protein folding rate.

We used the database of proteins with known experimental folding rates (http://phys.protres.ru/resources/compact.html) [19]. Proteins with disulfide bonds and large ligands are not included in this database. The resulting set includes 25 proteins folding with accumulation of intermediate state, 57 proteins following the "all-or-none" folding mechanism, and two peptides. It appeared upon protein sorting according to their origin that there are 12 multistage proteins of bacterial origin, 11 multistage proteins of eukaryotic origin, 23 single-stage proteins of bacterial origin, and 27 single-stage proteins of eukaryotic origin (Table 1 and http://antares.protres.ru/rate_evo_src.html).

The frequency of amino acid residues in the proteins was calculated as the ratio of the number of each type amino acid residue to the total number of amino acid residues in the protein, and then the average value for all proteins was taken.

RESULTS AND DISCUSSION

Statistics of protein structures according to the SCOP database of protein domains. We analyzed structural domains within kingdoms of living organisms in the SCOP database. We considered the protein domain quality rather than its origin as the main criterion of selection for database compilation. It is interesting to see the protein distribution among kingdoms in this selection. Figure 2 shows that eukaryotic proteins are better represented in α-helical, β-structure, and α+β classes, while proteins of bacterial origin are more common in the α/β and α+β classes. Archaean proteins are better represented in α/β and α+β classes, while virus proteins are more abundant in the β structure class.

The authors of work [32] showed that proteins of α/β class are more "ancient' than the others. Actually, most proteins of α/β class belong to bacteria, unlike the other classes. This result does not contradict data of [32], according to which the α/β class proteins are more "ancient" than the others, and α-helical proteins are "younger".

As shown in [18], the number of contacts per amino acid residue for α/β class proteins is statistically higher than for proteins of the other structural classes. We studied the shape of single-domain globular proteins and individual protein domains and their relationship with the protein folding kinetics. On average, α/β proteins were more "spherical" compared to those with different type of secondary structural packing, and in this case α/β proteins, again on the average, fold at a lower rate than proteins of the same length of the other classes. One explanation for this fact is that, under otherwise comparable conditions, for more spherical proteins such as α/β proteins the minimal surmountable interface between structured and unstructured parts in the transition state is larger, and folding proceeds at a lower rate [19] (Fig. 1). Other things being equal, the spherical protein, independently of its folding pathway, cannot escape a large interface area, while an oblong protein is able to "choose" a way in which

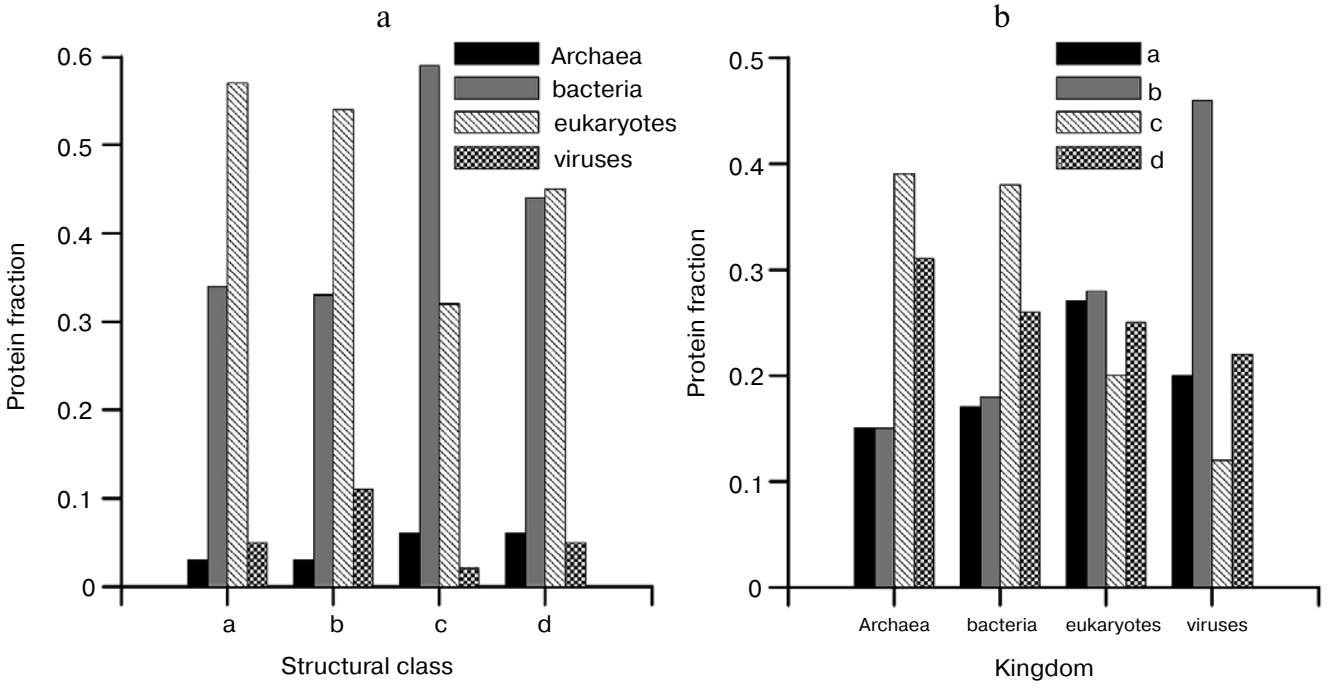

Fig. 2. Protein distribution: a) in structural class among kingdoms; b) within the kingdom among structural classes.

protein folding passes through the small interface area and therefore through a lower barrier. Therefore, more spherical proteins should fold more slowly. This means that on the average the "younger" α-helical proteins fold more rapidly compared to the more "ancient" α/β proteins. One may suppose that folding rate undergoes evolution along with evolution of proteins.

Statistical analysis of protein structures for which experimental folding data are available. Table 2 shows mean values on folding rate in water for proteins of different origins and having different folding mechanisms. It is seen that the slowly folding α/β class proteins are entirely absent from the group of single-stage proteins and did not get into the group of multistage eukaryotic proteins. We now have an approximately equal number of proteins with simple folding kinetics in each of three structural classes (Table 2). This allows us to conclude that bacterial proteins with simple kinetics fold on average at a higher rate compared to eukaryotic proteins with simple folding kinetics. On average, in our selection bacterial proteins with simple folding kinetics are even somewhat longer than eukaryotic proteins. Since there are still not many proteins with experimental data on folding rates for each structural class, it is reasonable to subdivide proteins into groups according to their folding type and origin (Tables 1 and 3).

We have obtained mean folding rate values separately for bacterial and eukaryotic proteins folding by the all-or-none mechanism and with intermediate accumulation. Table 3 shows averaged results. It was expected that the mean folding rate of proteins with multistage kinetics would be lower than that for proteins with single-stage folding kinetics. Bacterial proteins with simple folding kinetics, in turn, fold at a higher rate compared to eukaryotic proteins with simple folding kinetics (Table 3, $\ln(k_f)$ in water: 6.62 ± 0.67 against 4.86 ± 0.72, respectively). In this case it is seen in Table 3 that the difference in mean folding rate for bacterial proteins with simple and complex kinetics is two orders of magnitude (Table 3, $\ln(k_f)$ in water: 6.62 ± 0.67 against 0.28 ± 1.15, respectively). This much exceeds the difference for eukaryotic proteins with simple and complex folding kinetics (Table 3, $\ln(k_f)$ in water: 4.86 ± 0.72 against 3.25 ± 1.08, respectively). Such a high difference can be explained by the presence of slowly folding *c* class bacterial proteins. If these proteins are not considered, then the difference will be less pronounced (6.62 ± 0.67 against 4.70 ± 0.72, respectively).

Frequency of different types of amino acid residues in bacterial and eukaryotic proteins. We statistically analyzed the occurrence of the various amino acids in the studied proteins. Although for the total collection of proteins the conclusion of work [27] (such amino acids as phenylalanine (F) and glycine (G) influence the rate for proteins with simple folding kinetics, while cysteine, histidine, leucine, arginine (C, H, L, R) influence the rate for proteins with complex folding kinetics) does not contradict our data (data not shown), upon subdivision of the proteins into groups according to their origin we see a quite different situation. For eukaryotic and bacterial proteins the phenylalanine and glycine content is approximately

Table 1. Considered proteins and their characteristics: number of residues in protein, folding rate in water, cross section radius, contact order, absolute contact order, and origin

PDB code	Structural class	Number of residues in protein	$\ln(k_f)$, sec^{-1}	V_{ASA}/S_{ASA}, Å	Contact order (*CO*), %	Absolute contact order (Abs*CO*), %	Origin
1	2	3	4	5	6	7	8
Single-stage proteins of bacterial origin (23 proteins)							
2pdd	A	41	9.80	2.98	11	470	*Bacillus stearothermophilus* (*Geobacillus stearothermophilus*)
1w4e	A	45	8.7	2.53			
1prb	A	47	13.80	3.15	12	580	*Peptostreptococcus magnus*
1bdd	A	60	11.70	3.14	9	522	*Staphylococcus aureus*
1jyg	A	69	9.1	3.57	12	833	*Escherichia coli*
1imq	A	85	7.30	3.42	12	1044	*Escherichia coli*
256b	A	106	12.30	3.85	7	794	*Escherichia coli* (strain K12)
1l8w	A	294	2	4.52	8	2509	*Borrelia burgdorferi*
1c9o	B	66	7.20	3.45	17	1115	*Bacillus caldolyticus*
1g6p	B	66	6.30	3.33	18	1169	*Thermotoga maritima*
1csp	β	67	6.50	3.42	16	1098	*Bacillus subtilis*
1psf	β	69	3.20	3.24	17	1175	*Synechococcus* sp. (strain PCC 7002) (*Agmenellum quadruplicatum*)
1mjc	β	69	5.30	3.54	16	1102	*Escherichia coli* (strain K12)
1m9s	β	76	4.00	3.57	18	1388	*Listeria monocytogenes*
1k0s	β	143	7.40	3.79	13	1901	*Thermotoga maritima*
1lop	β	164	6.60	4.41	16	2571	*Escherichia coli* (strain K12)
1divN	α + β	56	6.60	3.29	13	709	*Bacillus stearothermophilus* (*Geobacillus stearothermophilus*)
3gb1	α + β	56	6.3	3.40	17	925	*Streptococcus* sp.
2ptl	α + β	60	4.1	3.47	18	1109	*Peptostreptococcus magnus*
1poh	α + β	85	2.70	3.80	18	1499	*Escherichia coli* (strain K12)
1divC	α + β	92	3.30	3.63	14	1271	*Bacillus stearothermophilus* (*Geobacillus stearothermophilus*)
1n88	α + β	96	2.00	3.46	14	1352	*Thermus thermophilus*
1ris	α + β	97	6.10	3.90	19	1838	*Thermus thermophilus*
Multistage proteins of bacterial origin (12 proteins)							
1ayi	α	85	7.20	3.68	10	862	*Escherichia coli*
1brs	α/β	89	3.40	4.04	12	1054	*Bacillus amyloliquefaciens*

Table 1 (Contd.)

1	2	3	4	5	6	7	8
3chy	α/β	128	1.00	4.26	9	1115	*Escherichia coli* (strain K12)
1aon	α/β	155	−1.50	4.18	14	2117	*Escherichia coli* (strain K12)
2rn2	α/β	155	0.10	4.10	12	1927	*Escherichia coli* (strain K12)
1ra9	α/β	159	−3.20	4.06	14	2231	*Escherichia coli* (strain K12)
1php1	α/β	175	2.30	4.40	12	2021	*Bacillus stearothermophilus* (*Geobacillus stearothermophilus*)
1php2	α/β	219	−3.50	4.41	8	1743	*Bacillus stearothermophilus* (*Geobacillus stearothermophilus*)
1qop1	α/β	267	−2.50	4.79	8	2228	*Salmonella typhimurium*
1qop2	α/β	390	−6.90	5.24	8	3253	*Salmonella typhimurium*
1gxt	α + β	88	4.40	3.88	21	1838	*Escherichia coli* (strain K12)
1bni	α + β	108	2.60	4.01	11	1226	*Bacillus amyloliquefaciens*
Single-stage proteins of eukaryotic origin (27 proteins)							
1vii	α	36	9.40	2.99	11	403	*Gallus gallus* (Chicken)
1gv2	α	47	8.70	3.34	12	578	*Mus musculus* (Mouse)
1ba5	α	49	5.90	3.41	15	712	*Homo sapiens* (Human)
1fex	α	59	8.20	3.39	13	770	*Homo sapiens* (Human)
1nti	α	86	7.00	3.60	12	1064	*Bos taurus* (Bovine)
1cun1	α	106	4.80	3.64	10	1097	*Gallus gallus* (Chicken)
1cun2	α	107	3.40	3.62	10	1094	*Gallus gallus* (Chicken)
1u5p	α	110	11.00	3.57	10	1134	*Gallus gallus* (Chicken)
1pin	β	34	9.50	2.99	19	647	*Homo sapiens* (Human)
1e0l	β	37	10.60	3.01	16	603	*Mus musculus* (Mouse)
1jmq	β	40	8.40	3.07	23	900	*Homo sapiens* (Human)
1rlq	β	56	4.40	3.28	20	1136	*Gallus gallus* (Chicken)
1shg	β	57	1.10	3.52	19	1089	*Gallus gallus* (Chicken)
1avz	β	57	4.90	3.51	20	1118	*Homo sapiens* (Human)
1jo8	β	58	2.50	3.58	19	1125	*Saccharomyces cerevisiae* (Baker's yeast)
1pnj	β	86	−1.00	3.59	16	1387	*Bos taurus* (Bovine)
1ten	β	89	1.10	3.85	17	1544	*Homo sapiens* (Human)
1fnf1	β	90	−0.90	3.75	18	1628	*Homo sapiens* (Human)
1wit	β	93	0.4	3.51	20	1890	*Caenorhabditis elegans* (Nematode)

Table 1 (Contd.)

1	2	3	4	5	6	7	8
1fnf2	β	94	5.50	3.62	17	1552	*Homo sapiens* (Human)
2ci2	α + β	64	5.80	3.50	16	1002	*Hordeum vulgare* (Barley)
1rfa	α + β	78	8.40	3.72	16	1261	*Homo sapiens* (Human)
1o6x	α + β	81	6.80	3.27	16	1323	*Homo sapiens* (Human)
1urn	α + β	96	4.60	3.84	17	1623	*Homo sapiens* (Human)
2acy	α + β	98	0.80	3.87	20	1962	*Bos taurus* (Bovine)
1aps	α + β	98	−1.60	3.96	21	2077	*Equus caballus* (Horse)
1fkf	α + β	107	1.60	3.93	18	1885	*Homo sapiens* (Human)
Multistage proteins of eukaryotic origin (11 proteins)							
1enh	α	54	10.50	3.46	14	736	*Drosophila melanogaster* (Fruit fly)
1a6n	α	151	1.10	4.30	8	1266	*Physeter catodon* (Sperm whale) (*Physeter macrocephalus*)
2a5e	α	156	3.50	3.67	5	832	*Homo sapiens* (Human)
1tit	β	89	3.60	3.66	18	1584	*Homo sapiens* (Human)
1hng	β	95	1.80	3.80	17	1603	*Rattus norvegicus* (Rat)
1eal	β	127	1.30	4.13	12	1575	*Sus scrofa* (Pig)
1ifc	β	131	3.40	4.32	14	1771	*Rattus norvegicus* (Rat)
1opa	β	133	1.40	4.29	14	1865	*Rattus norvegicus* (Rat)
1cbi	β	136	−3.20	4.05	14	1878	*Mus musculus* (Mouse)
1ubq	α + β	76	7.30	3.70	15	1147	*Homo sapiens* (Human)
2vik	α + β	126	5.00	3.72	12	1544	*Gallus gallus* (Chicken)

identical for multistage proteins. At the same time, for bacterial proteins the phenylalanine and glycine content is higher in proteins with simple folding kinetics, as it was noted in [27].

Statistical analysis on proteomes has shown that in bacteria (159 proteomes) alanine, glycine, and leucine are more frequent [33]. For our set of proteins, we observe an increase in the frequency fraction (the strongest difference compared to eukaryotic proteins) for such amino acids as alanine, glycine, and valine, and less pronounced difference for lysine, isoleucine, and phenylalanine (Fig. 3a). In eukaryotic proteomes (17 proteomes [33]) such amino acids as tryptophan, cysteine, phenylalanine, histidine, asparagine, glutamine, proline, serine, and threonine are more frequent. In our set of eukaryotic proteins we observe an increase in the frequency fraction compared to bacterial proteins for such amino acids as arginine, serine, threonine, tryptophan, tyrosine, proline, and histidine (Fig. 3a).

For 1446 bacterial and 1545 eukaryotic protein domains from the SCOP database we observe a frequency pattern of amino acid residues (Fig. 3d) that partially coincides with that described above: in bacterial proteins there are more alanines, glycines, and valines, while in eukaryotic proteins there are more cysteines, phenylalanines, lysines, serines, tryptophans, and threonines.

When amino acid compositions of multistage and single-stage bacterial and eukaryotic proteins are considered separately, then the following situation is obtained: multistage bacterial proteins contain more alanines, aspartic acid, proline, and valine, while in eukaryotic proteins there are more lysine, asparagines, and threo-

Table 2. Mean value of $\ln(k_f)$ in water for bacterial and eukaryotic proteins (subdivided into structural classes) following the "all-or-none" (single-stage proteins) folding mechanism and those with intermediate accumulation (multistage proteins)

	Bacteria			Eukaryotes		
	number of proteins	average number of amino acids in protein	$\ln(k_f)$ in water	number of proteins	average number of amino acids in protein	$\ln(k_f)$ in water
Multistage						
α	1	85	7.2	3	120 ± 33	5.0 ± 2.8
β	0	–	–	6	119 ± 9	1.4 ± 1.0
α/β	9	193 ± 30	−1.2 ± 1.1	0	–	–
α + β	2	98 ± 10	3.5 ± 0.9	2	101 ± 25	6.2 ± 1.2
Single-stage						
α	8	93 ± 30	9.3 ± 1.3	8	75 ± 11	7.3 ± 0.9
β	8	90 ± 14	5.8 ± 0.5	12	66 ± 7	3.9 ± 1.2
α/β	0	–	–	0	–	–
α + β	7	77 ± 7	4.4 ± 0.7	7	89 ± 6	3.8 ± 1.4

Table 3. Mean value of $\ln(k_f)$ in water for bacterial and eukaryotic proteins following the "all-or-none" folding mechanism and that with intermediate accumulation

Folding mechanism	"All-or-none"		With accumulation of folding intermediate	
Origin	bacteria	eukaryotes	bacteria	eukaryotes
Number of proteins	23 proteins	27 proteins	12 proteins	11 proteins
$\ln(k_f)$ in water	6.62 ± 0.67	4.86 ± 0.72	0.28 ± 1.15	3.25 ± 1.08

nine (Fig. 3b). At the same time, single-stage bacterial proteins contain more alanines and valines like multistage proteins, but in addition, the fraction of glycines and lysines also increases, while in eukaryotes there is more leucine, proline, arginine, serine, tryptophan, and tyrosine (Fig. 3c). In these lists amino acids that are prevalent in bacterial proteins with simple and complex folding kinetics, alanine and valine, can be distinguished. For eukaryotic proteins such amino acids are arginine, serine, and threonine. However, if one confines oneself only to proteins with simple kinetics, then **alanine, glycine, lysine,** and **valine** are prevalent in bacterial proteins, while **arginine, serine, proline,** and **tryptophan** are prevalent in eukaryotic proteins.

Comparison of folding rates for proteins with simple (single-stage) and complex (multistage) folding kinetics for proteins from different organisms. We compared the dependence of the protein folding rate in water on protein size and on a value proportional to the cross section radius. It appeared that for proteins following the single-stage folding mechanism (50 proteins) folding rate in water poorly correlates with these parameters: the correlation coefficient of folding rate with protein length was −0.32, while with a value proportional to the cross section radius (ratio of accessible protein volume to the area of its accessible surface, V_{ASA}/S_{ASA}) was −0.48. For single-stage eukaryotic proteins, correlation coefficient of folding rate with V_{ASA}/S_{ASA} was −0.68 (Fig. 4c). On the contrary, for proteins following the multistage folding mechanism (23 proteins, bacterial and eukaryotic), folding rate in water correlates well with the protein size and with a value proportional to cross sectional radius: the correlation coefficient of folding rate with protein length was −0.79 and with a value proportional to cross section radius it was

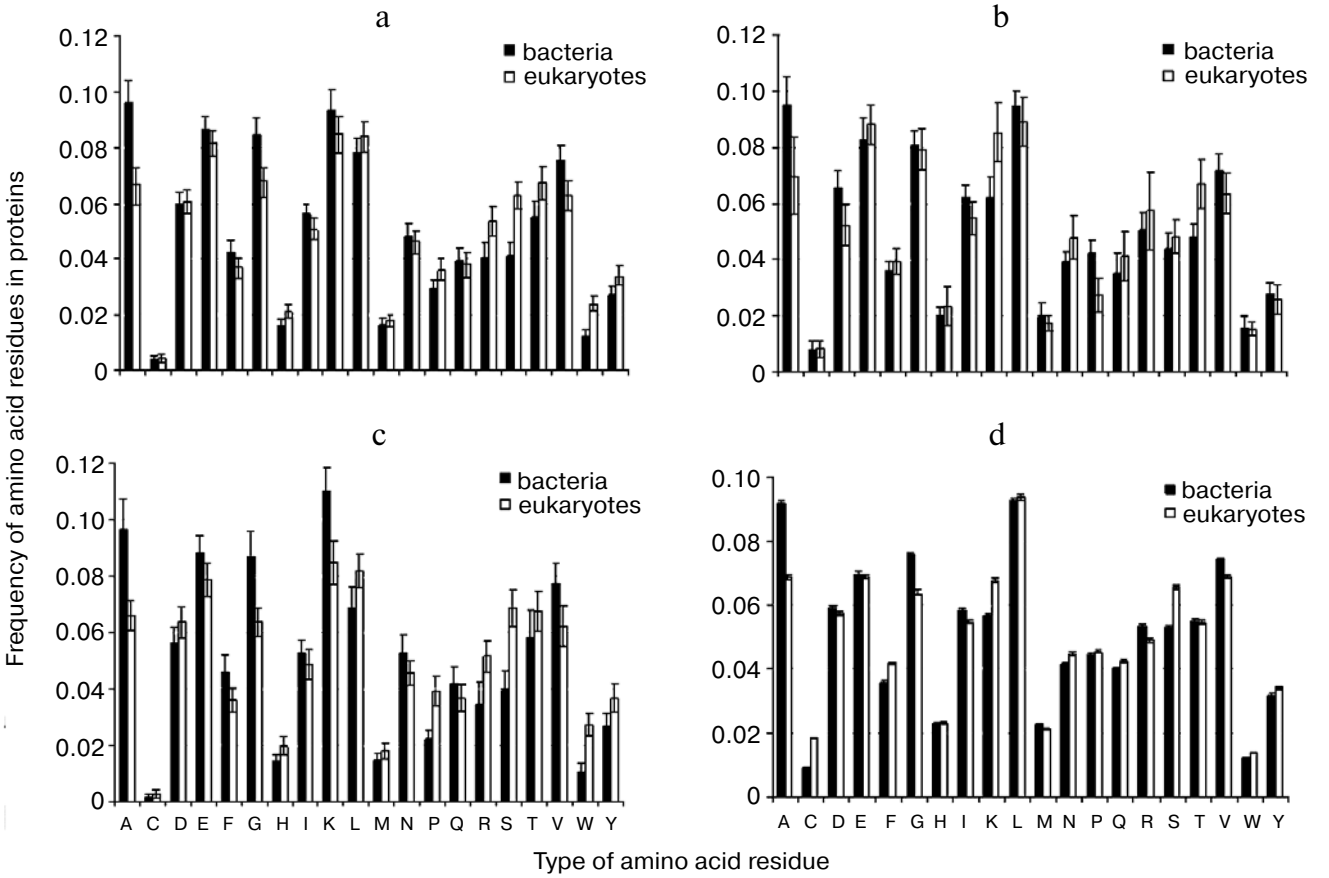

Fig. 3. Frequency of different types of amino acid residues in bacterial and eukaryotic proteins. a) All proteins (single-stage + multistage); b) multistage proteins; c) single-stage proteins; d) 1446 bacterial and 1545 eukaryotic proteins from the SCOP structural database.

also equal to −0.79. This result does not contradict our previous data [8]: protein size better correlates with folding rate of proteins following the multistage folding mechanism. It is seen in Fig. 4 (a, c, e) that single-stage bacterial proteins fold at a higher rate compared to single-stage eukaryotic proteins, although the mean size of bacterial proteins somewhat exceeds that of eukaryotic proteins. Figure 4 (b, d, f) shows that multistage eukaryotic proteins fold more rapidly than bacterial multistage proteins. This is so because the main contribution to folding of multistage bacterial proteins is made by class *c* structural proteins (α/β proteins exhibiting the slowest folding rate among all proteins).

When the folding rate dependence on protein size and on value proportional to cross section radius is analyzed after subdivision of the proteins into groups according to their origin, then bacterial proteins have very slightly better coefficients of correlation than eukaryotic proteins. The correlation coefficient of folding rate with protein length was −0.72 for bacterial proteins (35 proteins) with value proportional to cross section radius −0.74. For eukaryotic proteins (38 proteins) the correlation coefficient of folding rate with protein length was −0.56 and with value proportional to cross section radius it was −0.66.

Figure 4 (e, f) shows the relationship between logarithm of folding rate in water and logarithm of absolute order of contacts for single-stage (correlation coefficient of folding rate and absolute order of contacts were −0.67 for bacteria and −0.74 for eukaryotes) and multistage (−0.80 for bacteria and −0.70 for eukaryotes) proteins. It is seen from the graph that for single-stage eukaryotic proteins the parameter of absolute order of contacts is higher than in bacterial proteins (with the exception of three proteins). This is indicative of the existence of longer loops in the eukaryotic protein structures.

Synthesis of bacterial proteins is known to proceed at rates 4-10 times exceeding that of eukaryotic proteins [34]. The average duration of the elongation cycle in bacterial systems varies within the range from 0.05 to 0.1 sec at 37°C and it is ~3 times longer at 25°C. In eukaryotic systems elongation rate is lower and varies over broad limits due to existence of regulatory mechanisms. Usually the average time of a eukaryotic elongation cycle is from 0.1 to 0.5 sec [35]. One can suppose that for proteins with simple folding kinetics, *in vitro* folding rate correlates

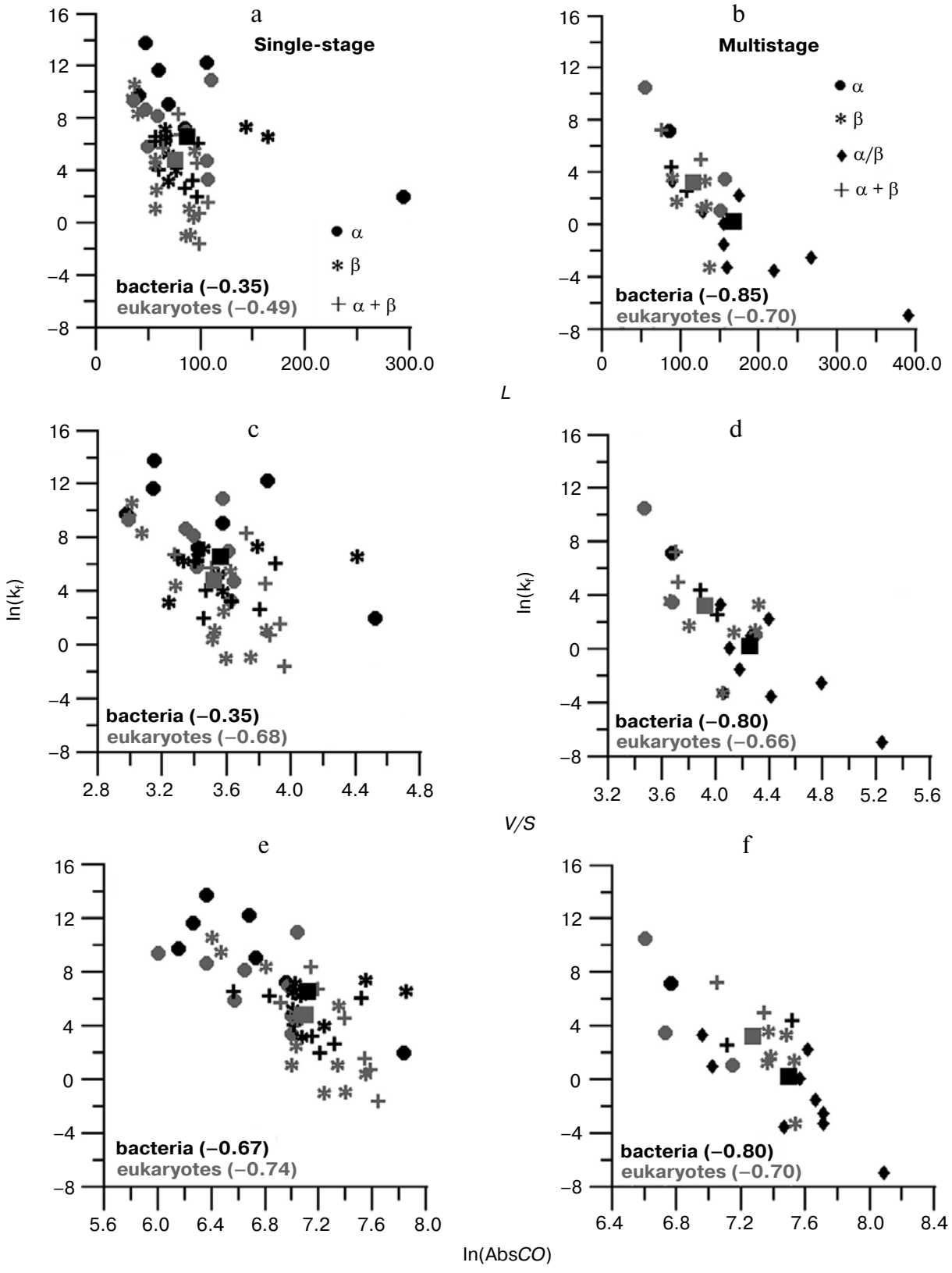

Fig. 4. Dependence of experimentally measured protein folding rates in water: a, b) on protein length (*L*); c, d) on ratio of accessible protein volume (*V*) to its accessible surface (*S*); e, f) on logarithm of absolute contact order (Abs*CO*%). Correlation is shown in parentheses. Rectangular symbols indicate mean values. Error does not exceed the symbol (square) size.

with *in vivo* rate of protein synthesis. It is reasonable to suppose that small protein domains that do not require additional regulatory mechanisms will be synthesized at a higher rate in bacterial systems, while complex proteins will be more efficiently and rapidly synthesized in eukaryotic systems. There is experimental confirmation for this fact. We know that eukaryotic genomes encode a much larger fraction of multidomain proteins compared to prokaryotic genomes. Chang et al. [36] checked whether eukaryotic and bacterial proteins differ in efficiency of multidomain protein folding caused by domain recombination. They compared folding of a set of recombinant proteins consisting of green fluorescent protein (GFP) joined to four different folding proteins via six different linkers during expression in *Escherichia coli* and *Saccharomyces cerevisiae*. It was found that, unlike yeasts, bacteria are not efficient in folding of such chimeric proteins even under comparable expression conditions. Experiments on folding *in vitro* and *in vivo* have shown that the GFP domain significantly restricts *de novo* folding of these partners in bacteria, which to a significant extent correlates with a posttranslational folding mechanism. They observed the accumulation of enzyme activity and found that the rate of emergence of correctly folded chimeric proteins on the ribosome is in fact significantly higher in yeasts than in bacteria.

In this work we have statistically analyzed protein domain frequency and distribution among kingdoms in the SCOP database of structural proteins. Folding rates of proteins with known experimental data were statistically analyzed. It is shown that the proteins of eukaryotic origin with single-stage kinetics fold, on average, more slowly than proteins of bacterial origin with single-stage kinetics. The frequency of amino acid residues for proteins of eukaryotic and bacterial origin was different for proteins with simple and complex folding kinetics.

This work was supported by the RFBR (grant 08-04-00561), the Russian Academy of Sciences (programs "Molecular and Cell Biology" (grant 01200959110) and "Fundamental Sciences for Medicine"), the Foundation for Cooperation with Russian Science, and the Federal Agency on Science and Innovations (02.740.11.0295).

REFERENCES

1. Jackson, S. E. (1998) *Fold. Des.*, **3**, R81-R91.
2. Finkelstein, A. V., and Badretdinov, A. Ya. (1997) *Fold. Des.*, **2**, 115-121.
3. Thirumalai, D. (1995) *J. Phys. Orsay Fr.*, **5**, 1457-1467.
4. Gutin, A. M., Abkevich, V. I., and Shakhnovich, E. I. (1996) *Phys. Rev. Lett.*, **77**, 5433-5436.
5. Koga, N., and Takada, S. (2001) *J. Mol. Biol.*, **313**, 171-180.
6. Finkelstein, A. V., and Galzitskaya, O. V. (2004) *Phys. Life Rev.*, **1**, 23-56.
7. Galzitskaya, O. V., Ivankov, D. N., and Finkelstein, A. V. (2001) *FEBS Lett.*, **489**, 113-118.
8. Galzitskaya, O. V., Garbuzynskiy, S. O., Ivankov, D. N., and Finkelstein, A. V. (2003) *Proteins*, **51**, 162-166.
9. Fersht, A. R. (1997) *Curr. Opin. Struct. Biol.*, **7**, 3-9.
10. Plaxco, K. W., Simons, K. W., and Baker, D. (1998) *J. Mol. Biol.*, **277**, 985-994.
11. Guijarro, J. I., Morton, C. J., Plaxco, K. W., Campbell, I. D., and Dobson, C. M. (1998) *J. Mol. Biol.*, **276**, 657-667.
12. Plaxco, K. W., Guijarro, J. I., Morton, C. J., Pitkeathly, M., Campbell, I. D., and Dobson, C. M. (1998) *Biochemistry*, **37**, 2529-2537.
13. Perl, D., Welker, Ch., Schindler, Th., Schroder, K., Marahiel, M. A., Jaenicke, R., and Schmid, F. X. (1998) *Nature Struct. Biol.*, **5**, 229-235.
14. Van Nuland, N. A. J., Chiti, F., Taddei, N., Raugei, G., Ramponi, G., and Dobson, C. M. (1998) *J. Mol. Biol.*, **283**, 883-891.
15. Zerovnik, E., Virden, R., Jerala, R., Turk, V., and Waltho, J. P. (1998) *Proteins*, **32**, 296-303.
16. Ivankov, D. N., Garbuzynskiy, S. O., Alm, E., Plaxco, K. W., Baker, D., and Finkelstein, A. V. (2003) *Protein Sci.*, **12**, 2057-2062.
17. Galzitskaya, O. V., Bogatyreva, N. S., and Ivankov, D. N. (2008) *J. Bioinform. Comput. Biol.*, **6**, 667-680.
18. Galzitskaya, O. V., Reifsnyder, D. C., Bogatyreva, N. S., Ivankov, D. N., and Garbuzynskiy, S. O. (2008) *Proteins*, **70**, 329-332.
19. Ivankov, D. N., Bogatyreva, N. S., Lobanov, M. Yu., and Galzitskaya, O. V. (2009) *PLoS ONE*, **4**, e6476.
20. Punta, M., and Rost, B. (2005) *J. Mol. Biol.*, **348**, 507-512.
21. Ivankov, D. N., and Finkelstein, A. V. (2004) *Proc. Natl. Acad. Sci. USA*, **101**, 8942-8944.
22. Zhou, H., and Zhou, Y. (2002) *Biophys. J.*, **82**, 458-463.
23. Gong, H., Isom, D. G., Srinivasan, R., and Rose, G. D. (2003) *J. Mol. Biol.*, **327**, 1149-1154.
24. Capriotti, E., and Casadio, R. (2007) *Bioinformatics*, **23**, 385-386.
25. Gromiha, M. M., Thangakani, A. M., and Selvaraj, S. (2006) *Nucleic Acids Res.*, **34**, W70-W74.
26. Gromiha, M. M., and Selvaraj, S. (2001) *J. Mol. Biol.*, **310**, 27-32.
27. Ma, B. G., Chen, L. L., and Zhang, H. Y. (2007) *J. Mol. Biol.*, **370**, 439-448.
28. Gromiha, M. M. (2005) *J. Chem. Inf. Model*, **45**, 494-501.
29. Lobkovsky, A. E., Wolf, Yu. I., and Koonin, E. V. (2010) *Proc. Natl. Acad. Sci. USA*, **107**, 2983-2988.
30. Povolotskaya, I. S., and Kondrashov, F. A. (2010) *Nature*, **465**, 922-927.
31. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) *J. Mol. Biol.*, **247**, 536-540.
32. Winstanley, H. F., Abeln, S., and Deane, C. M. (2005) *Bioinformatics*, **21**, i449-i458.
33. Bogatyreva, N. S., Finkelstein, A. V., and Galzitskaya, O. V. (2005) *J. Bioinform. Comput. Biol.*, **4**, 597-608.
34. Widmann, M., and Christen, P. (2000) *J. Biol. Chem.*, **275**, 18619-18622.
35. Spirin, A. S. (2010) *Molecular Biology: Ribosome Structure and Protein Biosynthesis* (in press).
36. Chang, H. C., Kaiser, C. M., Hartl, F. U., and Barral, J. M. (2005) *J. Mol. Biol.*, **353**, 397-409.