Journal of Bioinformatics and Computational Biology
Vol. 4, No. 2 (2006) 597–608
© Imperial College Press

TREND OF AMINO ACID COMPOSITION OF PROTEINS OF DIFFERENT TAXA

NATALYA S. BOGATYREVA*, ALEXEI V. FINKELSTEIN† and OXANA V. GALZITSKAYA‡

Institute of Protein Research, Russian Academy of Sciences
Institutskaya str., 4, Pushchino, Moscow Region, 142290, Russia
**bogat@alpha.protres.ru*
†afinkel@vega.protres.ru
‡ogalzit@vega.protres.ru

Received 15 September 2005
Revised 7 December 2005
Accepted 23 December 2005

Archaea, bacteria and eukaryotes represent the main kingdoms of life. Is there any trend for amino acid compositions of proteins found in full genomes of species of different kingdoms? What is the percentage of totally unstructured proteins in various proteomes? We obtained amino acid frequencies for different taxa using 195 known proteomes and all annotated sequences from the Swiss–Prot data base. Investigation of the two data bases (proteomes and Swiss–Prot) shows that the amino acid compositions of proteins differ substantially for different kingdoms of life, and this difference is larger between different proteomes than between different kingdoms of life. Our data demonstrate that there is a surprisingly small selection for the amino acid composition of proteins for higher organisms (eukaryotes) and their viruses in comparison with the "random" frequency following from a uniform usage of codons of the universal genetic code. On the contrary, lower organisms (bacteria and especially archaea) demonstrate an enhanced selection of amino acids. Moreover, according to our estimates, 12%, 3% and 2% of the proteins in eukaryotic, bacterial and archaean proteomes are totally disordered, and long (> 41 residues) disordered segments are found to occur in 16% of arhaean, 20% of eubacterial and 43% of eukaryotic proteins for 19 archaean, 159 bacterial and 17 eukaryotic proteomes, respectively. A correlation between amino acid compositions of proteins of various taxa, show that the highest correlation is observed between eukaryotes and their viruses (the correlation coefficient is 0.98), and bacteria and their viruses (the correlation coefficient is 0.96), while correlation between eukaryotes and archaea is 0.85 only.

Keywords: Proteome; eukaryotes; viruses; uniform usage of codone.

1. Introduction

Recent successes in the whole-genome sequencing allow for the first detailed proteomic investigations of organisms. Comparative analysis of proteomes is a powerful method in prediction of structures and functions of proteins.[1] The distribution of amino acid residues is also a key element in bioinformatics. Investigation of proteomes shows that the amino acid compositions of proteins differ substantially for different kingdoms of life.[2] A question of general scientific interest is universal trends for amino acid compositions for proteins from different kingdoms.

Interrelations within the amino acid composition of proteins belonging to different taxa were studied at the very beginning of the molecular biology era. It is reasonable to renew such investigations in view of enormous enlargement of databases. Due to the fully exploited wealth of available data, some sequence features that are of interest in comparative genomics and proteomics studies can be revealed: charge clusters, alternating patterns of charge residues, histidine residues and etc.

Recently the universal trends in ongoing changes of amino acid frequencies have been reported.[3] Sets of ortologous proteins encoded by triplets of closely related genomes from 15 taxa representing all three domains of life have been compared, and phylogenies have been used to polarize amino acid substitutions: Cys, Met, His, Ser and Phe accrue in at least 14 taxa, whereas Pro, Ala, Glu and Gly are lost.[3]

In this work we have calculated and compared the amino acid composition for taxa representing all domains of life (eukaryotes, bacteria, archaea, viruses of eukaryotes and viruses of bacteria) from 195 known proteomes and all annotated sequences from the Swiss–Prot data base.[4] Our analysis shows that each of the superkingdoms is compositionally distinct. Moreover, we have compared the amino acid composition of different taxa with the frequencies following from a uniform usage of codons of the universal genetic code for 20 natural amino acids. Our results show that there are a surprisingly small selection for amino acid composition of proteins of eukaryotes and their viruses and enhanced selection of codon usage observed for archaea proteins in comparison with the frequency following from a uniform usage of codons of the universal genetic code. Some amino acid frequencies obtained from the known proteomes have differences in comparison with the data obtained from the Swiss–Prot data base. We have demonstrated that more complex organisms have more random distribution of amino acids. Another interesting result is that the differences between proteomes are larger than the differences between different kingdoms of life.

A substantial number of proteins in any proteome are likely unstructured. It was suggested that the lack of rigid globular structure under physiological conditions might represent a considerable functional advantage for "natively unfolded" proteins, as their large plasticity allows them to interact efficiently with several different targets as compared to a folded protein with limited conformational flexibility.[5,6] It was shown that disordered regions are involved in DNA-binding and other types of molecular recognition and a large portion of the sequences of "natively unfolded" proteins contain segments of low complexity and high predicted flexibility.[7–14] It also indicated that a combination of low overall hydrophobicity and a large net charge represent a structural feature of "natively unfolded" proteins in comparison with small globular proteins.[15,16] We have suggested a simple indicator of "natively unfolded" proteins.[17] It is the expected average number of contacts per residue calculated from the amino acid sequence alone. In this work we estimated the percentage of totally and partially unstructured proteins in proteomes of different taxa considering the scale of the average number of contacts per residues calculated for 20 amino acids in globular state.[17] We have demonstrated that the more complex are organisms, the more proteins and segments are disordered.

2. Methods

2.1. *Databases*

We consider 19 archaean proteomes, 159 bacterial and 17 eukaryotic proteomes. The protein sequences were downloaded from the EBI ftp server (ftp://ftp.ebi.ac.uk/pub/databases/SPproteomes/swissprot_files/proteomes/). The names of proteomes are available at http://phys.protres.ru/aminoacidcomposition.html. We used one of the last versions of Swiss–Prot data base[4] — Swiss–Prot 44.

2.2. *Estimation of error*

Error bars in the figures represent standard deviations from average values in the frequency of amino acids,

$$\frac{\sum_n x_i}{n}. \tag{1}$$

Here $\sum x_i$ is the number of occurrences of the given amino acid in a complete proteome, n is the total length of all sequences from the given proteome. The standard deviation for (1) is calculated[18] as $\sigma/\sqrt{n}$, where σ is the root-mean-square deviation for x_i values:

$$\sigma = \sqrt{\frac{n \sum_n x_i^2 - (\sum_n x_i)^2}{n(n-1)}}. \tag{2}$$

2.3. *The average number of contacts per residue in protein*

The average number of contacts per residue in globular state for 20 amino acids are presented in Table 1.[17]

The expected average number of contacts per residue from the amino acid sequence alone is calculated as a sum of the average number of contacts of all residues divided by the number of residues in the amino acid sequence. We used this property, i.e. the average number of contacts per residue, to predict the state of protein with an unknown three-dimensional structure: either folded or unfolded. If the expected average number of contacts per residue in protein is less than 20.4 then

Table 1. The average number of contacts per residue in globular state.

G	P	A	D	E	K	S	N	Q	T
17.1	17.4	19.9	17.4	17.5	17.7	18.2	18.5	19.2	19.8
R	H	C	V	M	L	I	Y	F	W
21.0	21.7	23.5	23.9	24.8	25.4	25.7	25.9	27.2	28.5

this protein is predicted as being in the unfolded form. However, if the expected average number of contacts is larger than 20.4, then we find disordered segments satisfying the criteria that the expected average number of contacts within the given segments is less than 20.4 and the size of this segment is equal or larger than the window-size of 11 and 41 residues, respectively.

3. Results and Discussion

3.1. *Amino acid composition of proteins of different taxa in comparison with the frequency following from a uniform codon usage*

We have calculated the amino acid composition for taxa representing all domains of life (eukaryotes, bacteria, archaea) from 195 known proteomes and all annotated sequences from Swiss–Prot data base. Tables 2 and 3 list a number of organisms, proteins and amino acids in these proteins for each taxon.

We have compared the amino acid composition of proteins encoded by the genomes from different kingdoms of life (eukaryotes, bacteria, archaea, viruses of eukaryotes and viruses of bacteria) with the frequencies following from a uniform usage of codons of the universal genetic code for 20 natural amino acids (the latter means, e.g. that Trp, coded by one codon of 61 amino acid encoding triplets, has

Table 2. Correlation between amino acid compositions of proteins of various taxa calculated from the Swiss–Prot data base and the amino acid composition following from uniform codon usage.* plus/minus root-mean-square deviation of correlation coefficients found for individual species from the mean correlation coefficient found for the taxa.

Name of Taxon	Number of Organisms	Number of Proteins	Number of Amino Acids	Correlation Coefficient*
Archaea	86	8744	2495787	0.51 ± 0.03
Viruses of bacteria	133	1272	275600	0.62 ± 0.04
Bacteria	1241	66231	21613528	0.65 ± 0.03
Viruses of eukaryotes	547	7351	3399163	0.82 ± 0.02
Eukaryotes	4555	70273	28824081	0.77 ± 0.04

Table 3. Correlation between the amino acid compositions of proteins of various taxa calculated from 195 known proteomes and amino acid composition following from uniform codon usage.

Name of Taxon	Number of Organisms	Number of Proteins	Number of Amino Acids	Correlation Coefficient
Archaea	19	41708	11853213	0.61 ± 0.02
Bacteria	159	478967	149102019	0.70 ± 0.03
Eukaryotes	17	185349	87328914	0.77 ± 0.05

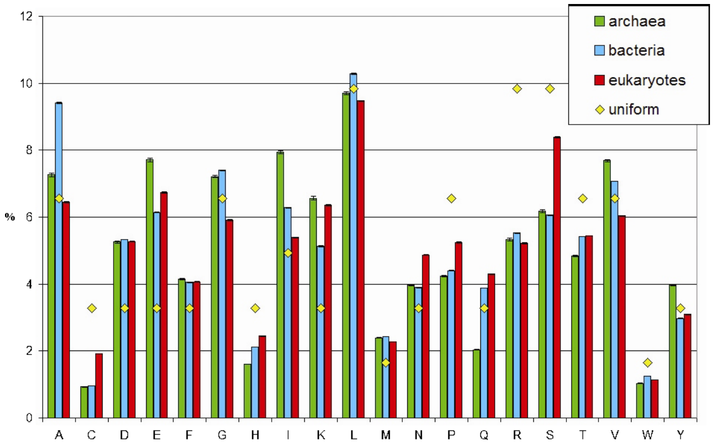

Fig. 1. Frequencies of amino acids in proteins for three main taxa calculated from the known proteomes (amino acid frequencies following from uniform codon usage are shown in triangles).

a random frequency of 1/61, while Ala, coded by four codons, has a random frequency of 4/61, etc.; the table of codon usage is taken from the work of Crick[19]). For this comparison we used all protein sequences from 195 known proteomes and the Swiss–Prot data base, which have no "open reading frames" that may not be expressed as proteins. Tables 2 and 3 represent the correlations between the amino acid composition in a given taxon and frequencies following from a uniform codon usage.

Tables 2, 3 and Fig. 1 show a surprisingly small selection for amino acid composition of proteins of eukaryotes and their viruses in comparison with the frequency following from a uniform usage of codons of the universal genetic code (i.e. eukaryotic proteins look more "random composed" than the others) and with an enhanced selection of codon usage observed for archaea proteins. We have not considered this result from a viewpoint of evolution. On the contrary to the results obtained by Takeuchi *et al.*,[20] in our case, archaea and bacteria are distinguishable by the difference between the real frequency in the translated proteins and theoretical frequencies (the expected frequency calculated from the ratio of nucleotides).

Figure 1 demonstrates the results for frequencies of individual amino acids in proteins from bacteria, archaea, and eukaryotes, representing three main domains of life, from 195 known proteomes. It should be noted that the frequencies calculated from the proteomes slightly changed for eukaryotes and most of all changed for archaea in comparison with the frequencies following from the Swiss–Prot data base. These differences can be seen from the correlation coefficient between frequencies of amino acid and random frequencies following from a uniform usage of codons of the universal genetic code. This correlation is 0.77 for eukaryotes both for 17 proteomes including 185 349 proteins and for the Swiss–Prot data base including 70 273 proteins.

For archaea this correlation is 0.61 for 19 proteomes, including 41708 proteins, and 0.5 for the Swiss–Prot data base including 8744 proteins. For bacteria this correlation is 0.70 for 159 proteomes, including 478 967 proteins, 0.65 for the Swiss–Prot database including 66 231 proteins, which accords with the earlier observations.[2]

It is surprising that the correlation does not change for eukaryotes despite of a different number of considered proteins, and changes for bacteria and archaea.

According to our data the changes in the amino acid frequencies concern such residues as Ala, Cys, Gly and Ser for eukaryotes. Frequences for Ala, Cys and Gly decrease, but for Ser it increases in the case of proteomes in comparison with the Swiss–Prot data base. Archaea's frequencies change for such residues as Glu, Phe, Lys, Leu, Ser, Val and Tyr.

Our data for the amino acid composition for three big domains of life obtained from the consideration of proteomes differ from recent published data[2] especially for such residues as Ala, Gly, Ile, Lys, Asn, Pro, Arg, Ser, Val.

If one were to consider how to represent amino acids in proteomes in view of the trend which has been revealed,[3] it is interesting that such amino acids found as losers in 14 taxa such as Ala, Gly are over represented (except Pro) in archaea and bacteria proteomes in average, and Glu in archaea, bacteria and eukaryotes proteomes relative to the frequencies corresponding to equal usage of all codons (see Fig. 1). At the same time amino acids which have been found as gainers such as Cys, His and Ser (except Phe and Met) are under-represented in archaea, bacteria and eukaryotes.

The comparison of amino acid contents of proteins found for organisms of different taxa (Table 4) shows that eukaryotic viruses are very close to eukaryotes (the correlation coefficient is 0.97), and viruses of bacteria are very close to bacteria (correlation coefficient is 0.96), while archaea is most different from all other taxa in this respect.

Figure 2 represents the correlation between the amino acid composition and theoretical frequency for each proteome. The highest correlation among archaea is 0.88 for *Aeropyrum pernix*, for bacteria 0.87 for *Rhodopirellula baltica*, and for eukaryotes 0.84 for *Homo sapiens*. It should be noted that the differences between proteomes are larger than the differences between taxa.

Table 4. Correlation between the amino acid compositions of proteins of various taxa from the Swiss–Prot database.

Name of Taxon	Archaea	Viruses of Bacteria	Bacteria	Viruses of Eukaryotes	Eukaryotes
Archaea	1.00	0.93	0.93	0.83	0.85
Viruses of bacteria	0.93	1.00	0.96	0.92	0.94
Bacteria	0.93	0.96	1.00	0.91	0.94
Viruses of eukaryotes	0.83	0.92	0.91	1.00	0.97
Eukaryotes	0.85	0.94	0.94	0.97	1.00

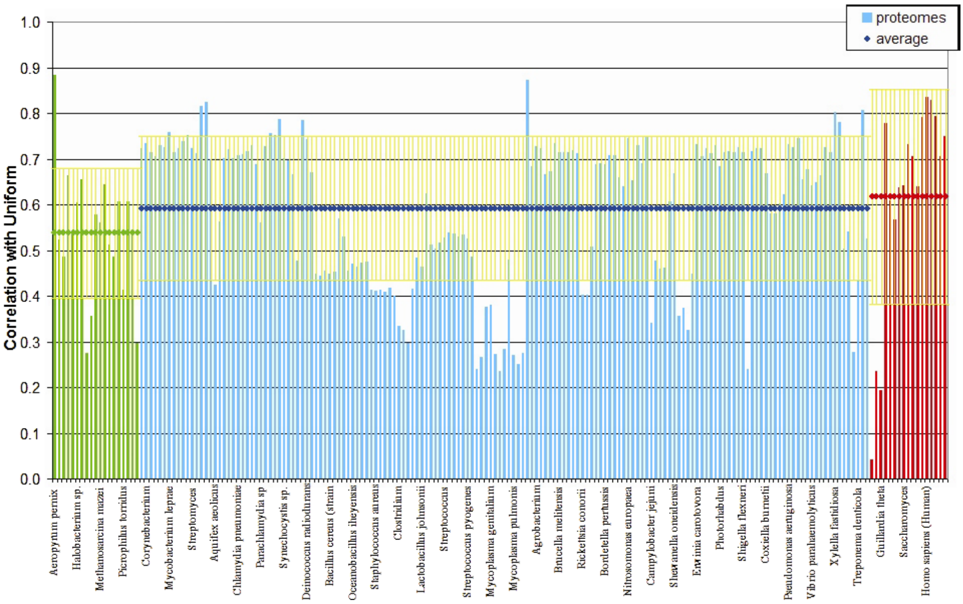

Fig. 2. Correlation between the amino acid compositions of each considered proteome and the amino acid composition following from the uniform codon usage. Arhaean proteomes are given in green, bacterial proteomes in blue and eukaryotic proteomes in red colors. Error bars of average are given in yellow color.

3.2. *Percentage of totally and partially unstructured proteins in various proteomes*

A receiver operator characteristic (ROC) curve for our method has been obtained (see Fig. 3), so the choice of significance thresholds are highly influenced by the rate of false disordered prediction. The true positive rate was calculated as the percentage of residues predicted as disordered on the unfolded list (sensitivity); the false positive rate is the percentage of predicted disordered residues on the folded set, also called specificity. The best result corresponds to the case when we construct the expected contact profile smoothed over the window-size of 41 residues, and the averaging for disordered regions is done over residues (see Table 5). Our method has been tested on datasets of globular proteins (559 proteins) and long disordered protein segments (129 proteins),[21] and showed improved performance over some other widely used methods, such as DISOPRED,[22] PONDR VL3H,[13] IUPred,[21] GlobPlot[23] (see Table 5).

Knowing the amino acid composition of each protein we can calculate the average number of contacts per residue to predict the state of the protein with an unknown three-dimensional structure: either folded or unfolded. To find the disordered regions, we can construct for each protein an expected contact profile smoothed over 41 residues taking into account the order of amino acids in

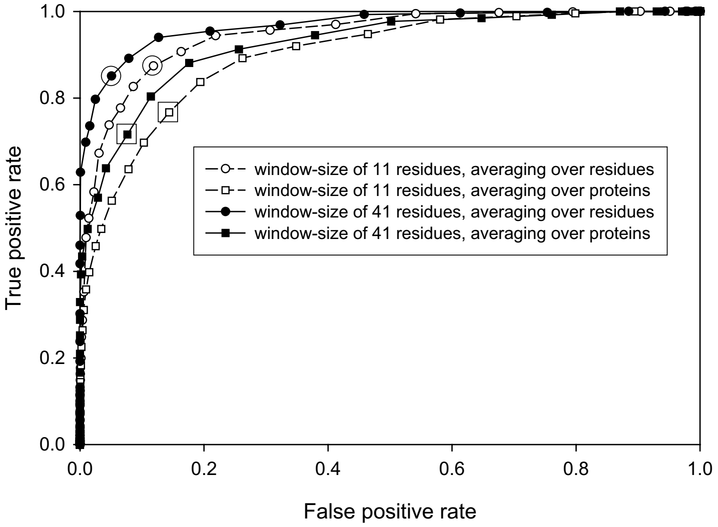

Fig. 3. Receiver operator characteristic (ROC) curves for our method. The true positive rate was calculated as the percentage of residues predicted as disordered on the unfolded list (129 proteins). The false positive rate is the percentage of predicted disordered residues on the folded set (559 proteins). Light circle — the expected contact profile is smoothed over a window-size of 11 residues, averaging is done over residues; black circle — the profile is smoothed over a window-size of 41 residues, averaging is done over residues; light square — the profile is smoothed over a window-size of 11 residues, averaging is done over proteins; black square — the profile is smoothed over a window-size of 41 residues, averaging is done over proteins. Large circles and squares are underlined points corresponding the threshold 20.4 expected number of contacts.

Table 5. Performance of disorder prediction methods.

	True Positive Rate Averaging is Done Over		False Positive Rate Averaging is Done Over	
Method	Residues	Proteins	Residues	Proteins
FoldUnfold[24] (we)	0.851	0.716	0.051	0.076
IUPred[21]	0.763	0.679	0.053	0.055
PONDRVL3H[13]	0.663	0.607	0.050	0.078
DISOPRED2[22]	0.664	0.491	0.050	0.069
GlobPlot[23]	0.330	0.304	0.181	0.197

protein chain. First, we calculated the percentage of fully unfolded proteins for 19 archaean, 159 bacterial and 17 eukaryotic proteomes: correspondingly, 2%, 3% and 12%. Second, we calculated the percentage of proteins, where there are disordered regions which are equal or larger than 41 residues for the same proteomes (partially unfolded proteins): 16%, 20% and 43%, respectively. Third, we calculated the percentage of disordered residues (including wholly and partially unfolded proteins): correspondingly, 8%, 11% and 25%. Figure 4 shows the estimated fraction of

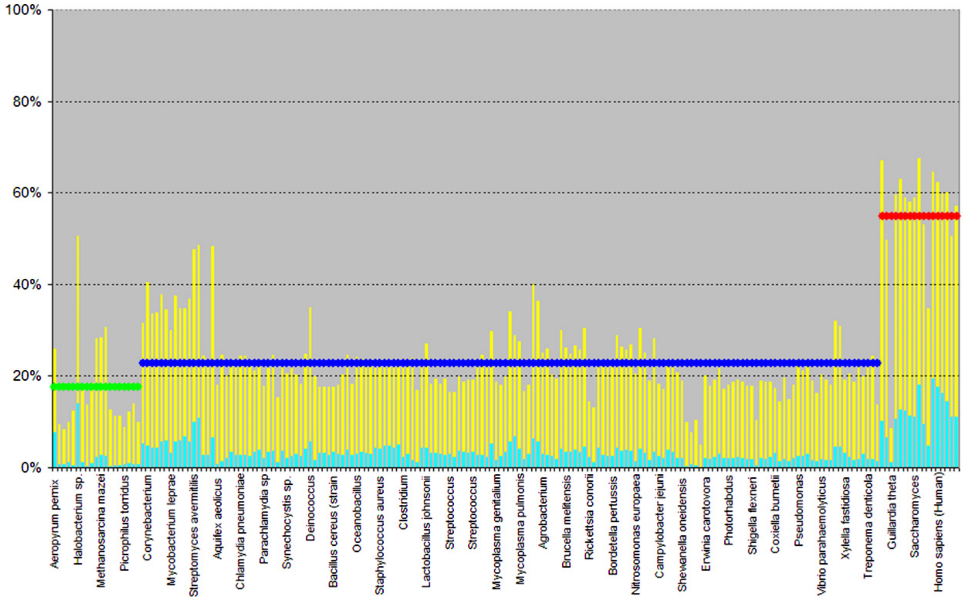

Fig. 4. Estimated fraction of wholly unstructured proteins (cyan color) and fraction of partially unfolded proteins (yellow) for 19 archaean, 159 bacterial and 17 eukaryotic proteomes. The average over both kinds of proteins is given by a green line for Arhaean proteomes, a blue line for bacterial proteomes and a red line for eukaryotic proteomes.

unstructured proteins and partially unfolded proteins for 19 archaean, 159 bacterial and 17 eukaryotic proteomes. Using DISOPRED[21] method long (> 30 residues) disordered segments are found to occur in 2% of arhaean, 4.2% of eubacterial and 33% of eukaryotic proteins for six archaean, 13 bacterial and five eukaryotic proteomes. The higher percentage of unstructured proteins in eukaryotes may be attributed to the increased prevalence of signaling and regulatory processes in eukaryotes.[21]

4. Conclusions

It is worthwhile to underline that from our analysis we can trace the amino acid composition of different taxa of life: the more complex organization of living the more random distribution of amino acids. The differences between proteomes, if one were to consider the correlation between the amino acid compositions and theoretical frequencies (frequencies corresponding to equal usage of all codons) for each proteome, are larger than the differences between taxa.

In this study we have estimated the percentage of totally and partially unstructured proteins in various proteomes. According to our estimates, 2%, 3% and 12% of the proteins in archaean, bacterial and eukaryotic proteomes are totally disordered, and long (> 41 residues) disordered segments are found to occur in 16% of arhaean, 20% of eubacterial and 43% of eukaryotic proteins for correspondingly 19 archaean, 159 bacterial and 17 eukaryotic proteomes.

Moreover, the analysis of amino acid composition of proteins is important for sequence alignments, functional annotation and a phylogenetic assay.

Acknowledgments

This work was supported by the program MCB RAS, by the Russian Foundation for Basic Research, by an International Research Scholar's Award to A.V. Finkelstein from the Howard Hughes Medical Institute (grant 55000567), the INTAS-grant (05-1000004-7747) and by the Russian Science Support Foundation.

References

1. Karlin S, Mrazek J, Gentles AJ, Genome comparisons and analysis, *Curr Opin Struct Biol* **13**:344–352, 2003.
2. Pe'er, Felder CE, Man O, Silman I, *et al.*, Proteomic signatures: amino acid and oligopeptide compositions differentiate among phyla, *Proteins* **54**:20–40, 2004.
3. Jordan IK, Kondrashov FA, Adzhubei IA, *et al.*, A universal trend of amino acid gain and loss in protein evolution, *Nature* **433**:633–638, 2005.
4. Bairoch A, Apweiler R, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, *Nucleic Acids Res* **28**:45–48, 2000.
5. Wright PE, Dyson HJ, Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm, *J Mol Biol* **293**:321–331, 1999.
6. Dyson HJ, Wright PE, Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance, *Adv Protein Chem* **62**:311–340, 2002.
7. Wootton JC, Non-globular domains in protein sequences: automated segmentation using complexity measures, *Comput Chem* **18**:269–285, 1994.
8. Dunker AK, Garner E, Guilliot S, *et al.*, Protein disorder and the evolution of molecular recognition: theory, predictions and observations, *Pac Symp Biocomput* **3**:473–484, 1998.
9. Romero P, Obradovic Z, Kissinger CR, *et al.*, Thousands of proteins likely to have long disordered regions, *Pac Symp Biocomput* 437–448, 1998.
10. Romero P, Obradovic Z, Dunker AK, Folding minimal sequences: the lower bound for sequence complexity of globular proteins, *FEBS Lett* **462**:363–367, 1999.
11. Galzitskaya OV, Surin AK, Nakamura H, Optimal region of average side-chain entropy for fast protein folding, *Protein Sci* **9**:580–586, 2000.
12. Vucetic S, Brown CJ, Dunker AK, Obradovic Z, Flavors of protein disorder, *Proteins* **52**:573–584, 2003.
13. Obradovic Z, Peng K, Vucetic S, *et al.*, Predicting intrinsic disorder from amino acid sequence, *Proteins* **53**:566–572, 2003.
14. Radivojac P, Obradovic Z, Smith DK, *et al.*, Protein flexibility and intrinsic disorder, *Protein Sci* **13**:71–80, 2004.
15. Uversky VN, Gillespie JR, Fink AL, Why are "natively unfolded" proteins unstructured under physiologic conditions?, *Proteins* **41**:415–427, 2000.
16. Uversky VN, What does it mean to be natively unfolded? *Eur J Biochem* **269**:2–12, 2002.
17. Garbuzynskiy SO, Lobanov MY, Galzitskaya OV, To be folded or to be unfolded? *Protein Sci* **13**:2871–2877, 2004.

18. Hudson DJ, Statistics. Lectures on elementary statistics and probability, Geneva, 1964.
19. Crick FHC, *Cold Spring Harbor Symp Quant Biol* **31**:1–9, 1996.
20. Takeuchi F, Futamura Y, Yoshikura H, Yamamoto K, Statistics of trinucleotides in coding sequences and evolution, *J Theor Biol* **222**:139–149, 2003.
21. Dosztanyi Z, Csizmok V, Tompa P, Simon I, The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins, *J Mol Biol* **347**:827–839, 2005.
22. Ward JJ, Sodhi JS, McGuffin LJ, *et al.*, Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, *J Mol Biol* **337**:635–645, 2004.
23. Linding R, Jensen LJ, Diella F, *et al.*, Protein disorder prediction: implications for structural proteomics, *Structure* **11**:1453–1459, 2003.
24. Galzitskaya OV, Garbuzynskiy SO, Lobanov MY, Prediction of natively unfolded regions in protein chain, *Mol Biol* (Russia), **2**:341–348, 2006.

Natalya S. Bogatyreva received diploma degree in biology (2000) and in mathematics (2001) from Samara State University, Russia. Now she is graduate student in the Laboratory of Protein Physics, Institute of Protein Research, Russian Academy of Sciences. Awarded for Outstanding graduate student, 2004. Area of expertise: investigation of protein folding and misfolding, protein physics, molecular biology, biochemistry, bioinformatics, structure prediction.

Alexei V. Finkelstein graduated from the Moscow Physical-Technical Institute (1970, Honorary Diploma). Ph.D. in Biophysics (Moscow Phys.-Tech., 1976); D.Sc. in Physical & Mathematical Sciences (Moscow University, 1991). Head of the Protein Physics Laboratory (Institute of Protein Research, Russian Academy of Sciences); Full Professor in Biophysics at the Moscow University, Soros Professor. Author of about 170 scientific papers and a book "Protein Physics" (in Russian, 2002 & 2005, and in English, 2002). Awarded the State Prize of Russia (1999) and various Russian and international research grants and awards; the Howard Hughes Medical Institute International Research Scholar (since 1995). Member of the Editorial Boards of "Journal of Computational Biol." (USA) and "Molecular Biology" (Russia). Member and of the AAAS, Protein Society, etc. Area of expertise: investigation of protein folding, protein physics, molecular physics, molecular biology, biochemistry, bioinformatics, protein engineering.

Oxana V. Galzitskaya graduated from the Moscow Physical-Technical Institute (1990, Honorary Diploma). Ph.D. in Biophysics (Institute of Theoretical and Experimental Biophysics, Pushchino, 1996). Senior researcher of the Laboratory of Protein Physics at the Institute of Protein Research, Russian Academy of Sciences. Author of about 40 scientific papers. Awarded for Outstanding Russian Young Scientists, 1997. Member of the Editorial Boards of "Current Protein and Peptide Science". Area of expertise: investigation of protein folding and misfolding, protein physics, molecular biology, biochemistry, bioinformatics, prediction of protein disorder and structure prediction.