

Protein Folding: Looping From Hydrophobic Nuclei

Igor N. Berezovsky,^{1*} Valery M. Kirzhner,² Alla Kirzhner,² and Edward N. Trifonov^{1,2}

¹Department of Structural Biology, The Weizmann Institute of Science, Rehovot, Israel

²Genome Diversity Center, Institute of Evolution, University of Haifa, Haifa, Israel

ABSTRACT Protein structure can be viewed as a compact linear array of nearly standard size closed loops of 25–30 amino acid residues (Berezovsky et al., *FEBS Letters* 2000; 466: 283–286) irrespective of details of secondary structure. The end-to-end contacts in the loops are likely to be hydrophobic, which is a testable hypothesis. This notion could be verified by direct comparison of the loop maps with Kyte and Doolittle hydrophobicity plots. This analysis reveals that most of the ends of the loops are hydrophobic, indeed. The same conclusion is reached on the basis of positional autocorrelation analysis of protein sequences of 23 fully sequenced bacterial genomes. Hydrophobic residues valine, alanine, glycine, leucine, and isoleucine appear preferentially at the 25–30 residues distance one from another. These observations open a new perspective in the understanding of protein structure and folding: a consecutive looping of the polypeptide chain with the loops ending primarily at hydrophobic nuclei. *Proteins* 2001;45:346–350.

© 2001 Wiley-Liss, Inc.

Key words: protein folding; closed loops; nuclei; hydrophobicity; domains; complete genomes; major folds

A fundamental property of protein structure is preferentially hydrophobic interior and hydrophilic exterior.² Protein folding, therefore, should involve formation of hydrophobic nuclei.^{3,4} The search for the nucleation centers is one of the most promising ways to solve the problem of protein folding.^{5,6} Among various forces stabilizing the assumed nucleation centers, the hydrophobic interactions (more generally, van-der Waals interactions) play a special role because of their enthalpy nature.⁴ Hence, analysis of distribution of the hydrophobic residues in 3-D and along the protein sequence is highly relevant to the problem of the protein folding.

Universal structural elements of the proteins—closed loops of 25–30 amino acid residues—have been recently discovered by exhaustive analysis of crystallized protein structures.¹ In that work, the ends of the closed loops were defined as chain-to-chain contacts with the shortest C α to C α distances, typically less than 10 Å (in almost 70% of cases, less than 7 Å). The typical closed loop size, 25–30 residues, is found to be the same for small and large proteins.⁷ The secondary structure was found to be, essentially, in no relation to the loop nature of the protein.¹

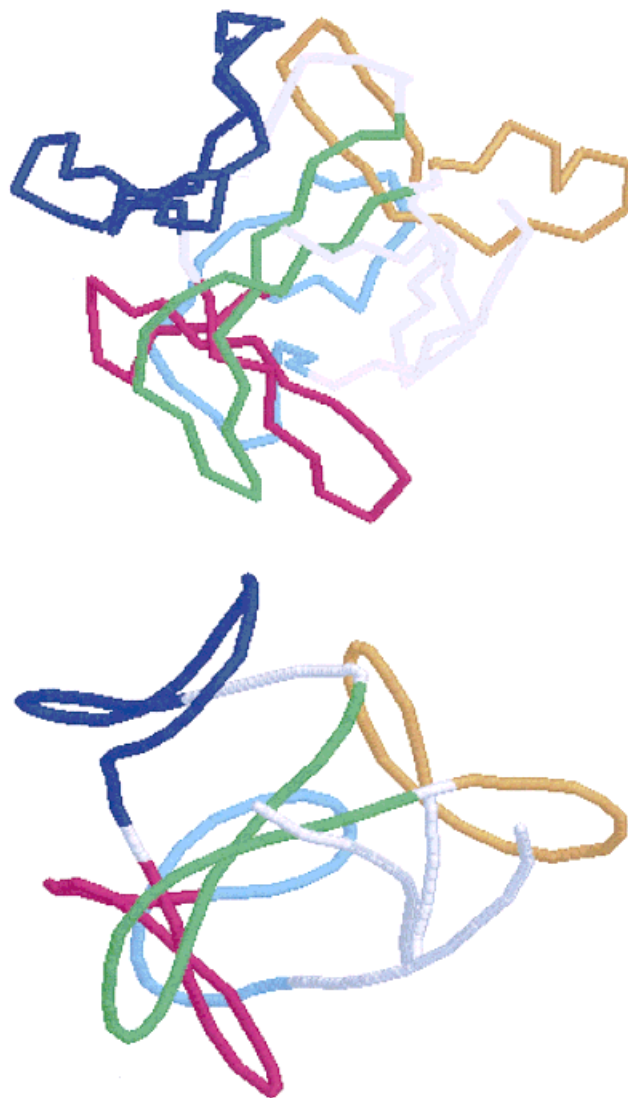


Fig. 1. Loop structure of Interleukin-1beta (Trefoil, 1i1b) in the traditional backbone presentation (**top**) and in smoothed form (**bottom**, see text). The loops of five different colors correspond to the map shown in Figure 2(B).

*Correspondence to: Dr. I.N. Berezovsky, Department of Structural Biology, The Weizmann Institute of Science, PO Box 26, Rehovot 76100, Israel. E-mail: Igor.Berezovsky@weizmann.ac.il

Received 23 February 2001; Accepted 2 July 2001

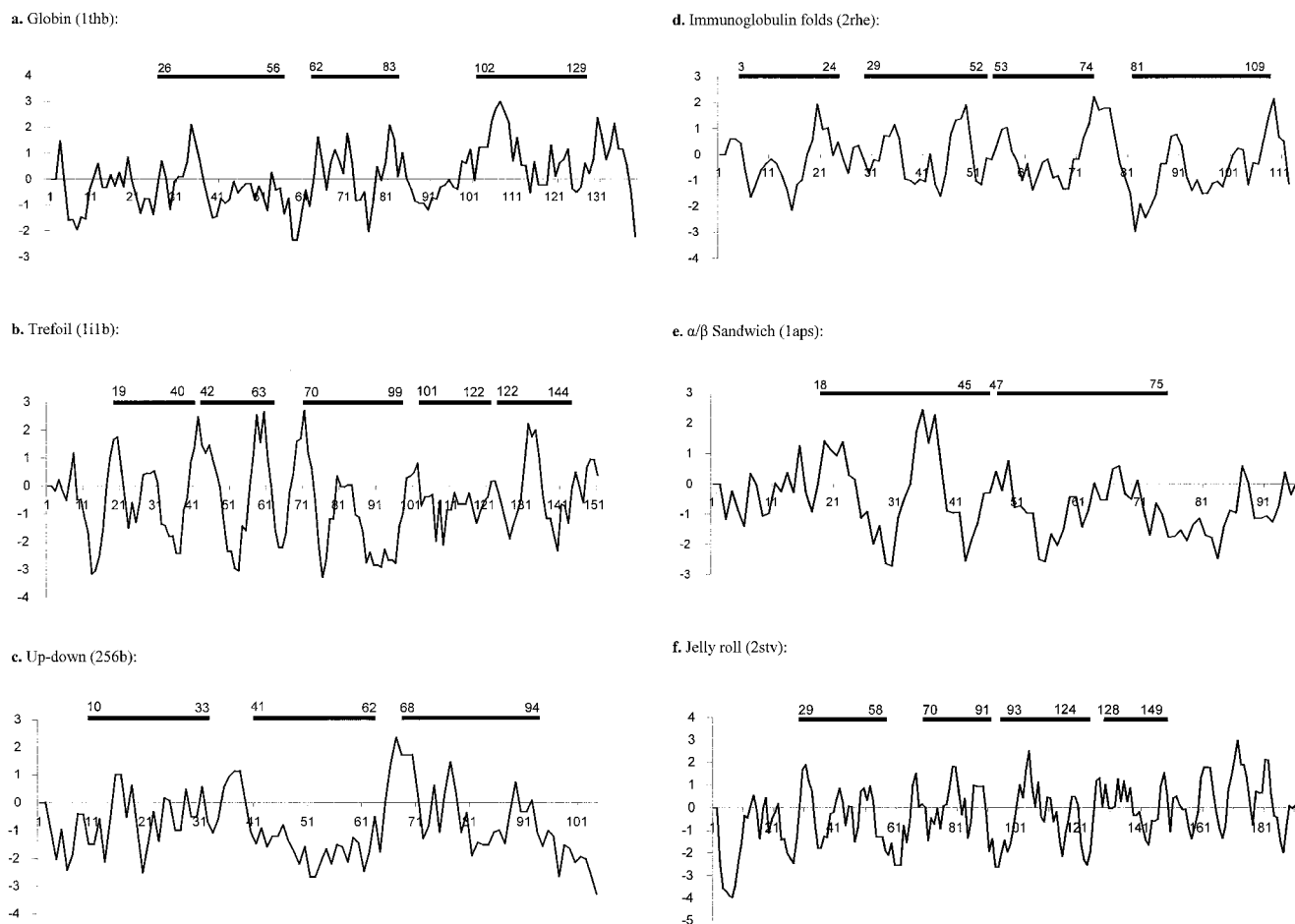


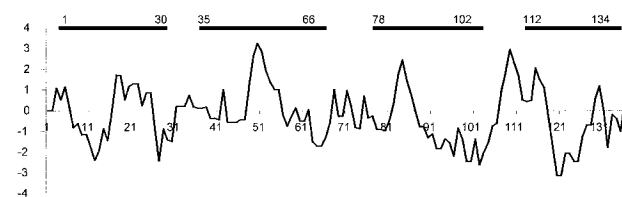
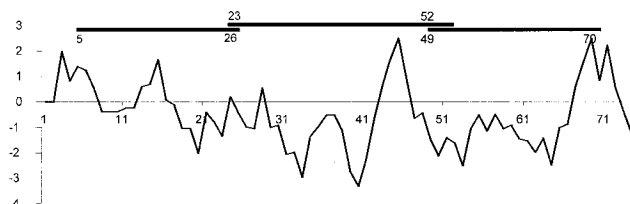
Fig. 2. Superposition of nearly standard size loops (1) on the Kyte and Doolittle hydrophobicity profiles (6) for nine major protein folds (5). **a:** Globin (1thb); **b:** Trefoil (1i1b); **c:** Up-down (256b); **d:** Immunoglobulin folds (2rhe); **e:** α/β Sandwich (1aps); **f:** Jelly roll (2stv); **g:** Doubly Wound (4fxn); **h:** UB α/β roll (1ubq); **i:** TIM barrel (7tim). The Kyte and Doolittle plots are calculated by ProtScale routine of the ExPASy Proteomic Tools. Hydrophobicity units correspond to original definition.⁹ Smoothing window for the hydrophobicity plots is taken to be equal five residues. The sequence coordinates for the protein folds are indicated at the x-axes, and the hydrophobicity values are indicated at the y-axes.

There was no obvious correlation between sequence positions of α and β elements on the one hand and the loop ends on the other. More than a quarter of the loop's ends actually reside in the middle of α and β sections.

In Figure 1, the protein Interleukin-1beta (Trefoil) is presented as a combination of the loops according to the map calculated in a previous study.¹ For the purpose of clarity, at the bottom of Figure 1 the path of the polypeptide chain is smoothed by a sliding window of seven amino acid residues (two periods of α -helices). The consecutive organization of the loops (one right after another) exemplified in Figure 1 is typical for all proteins tested.¹ The closing ends of the loops may be considered as the folding nuclei or parts thereof. One would expect, then, that the hydrophobic interactions would make a substantial contribution to the loop closure. To verify this expectation, we compared sequence locations of the loop ends¹ in nine major folds⁵ with positions of hydrophobic clusters in their sequences. These clusters were identified by the Kyte and Doolittle procedure⁹ as individual peaks in the hydrophobicity

plots. Figure 2 presents pairwise comparisons of the hydrophobicity⁹ and the loop maps. Figure 2(b) corresponds to the Trefoil structure (1i1b) in the Figure 1. Inspection of Figure 2 reveals that majority of the ends of the loops (85%) are found, indeed, to coincide with the peaks of hydrophobicity (loop mapping error bars: ± 3 amino acid residues). Quantitative agreement of the loop ends with the Kyte and Doolittle plots is further demonstrated by Figure 3, where a total of 52 sections of the hydrophobicity plots around the loop ends (left and right) of the nine folds are synchronized and summed up together. Essentially, Figure 3 represents an averaged hydrophobicity plot in the vicinity of the loop end. The position zero corresponds to the loop ends. The averaged hydrophobicity plot shows clear elevation over mean value with the main maximum only two amino-acid residues off the synchronized ends. The ragged shape of the plot is due, apparently, to a small ensemble size of the structures analyzed. While the nearly standard size of the loops can be explained by polymer statistics of the polypeptide

g. Doubly Wound (4fxn):

h. UB α/β roll (1ubq):

i. TIM barrel (7tim):

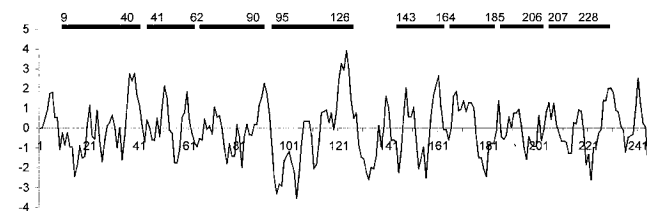


Figure 2. (Continued.)

chains,¹ the hydrophobicity of the ends of the loops cannot. The amino-acid sequence and its local biases could only be a result of evolutionary selection pressures. The results above, therefore, indicate that, apparently, there was such a selection pressure. Then a specific prediction follows: a tendency of hydrophobic amino acids to maintain a specific distance between them—25–30 residues. If, indeed, there is a sequence bias towards the standard distance, then the sequence correlation analysis may indicate such a bias. Occurrence of locally enriched hydrophobic sites would be rather high even in a random sequence, and clusters of four-five hydrophobic amino acid residues would appear along the random sequence about every 16 to 32 residues (taking for the purpose of the estimate 50% content of the hydrophobic residues). Therefore, if the expected distance correlation exists, it may appear only as a weak bias, with, e.g., avoidance of too short distances between the clusters. To ensure detection of such, presumably, weak signals we carried out positional autocorrelation analysis of protein sequences of 23 fully sequenced bacterial genomes (total over 42,000 sequences). The expected result of the calculation observation of a preferred distance of 25–30 residues between hydrophobic amino acids. In Figure 4(A), the positional autocorrelation functions for all 20 amino acids are summed up and calculated for 23 genomes. In addition

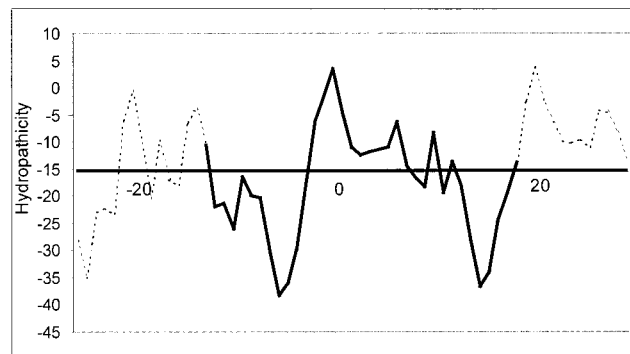


Fig. 3. Sum of synchronized hydropathicity plots around the ends of the mapped loops. To eliminate the sequence end effect, the loop ends closer than 30 amino acid residues to the ends of the sequences were excluded from the calculation.

to strong variations at short distances, peaks at residues 3, 4, and 7, responsible for α -helical components in the proteins,^{10,11} a clear excess of occurrences in the region 24–31 residues is, indeed, observed. Respective plot for randomized (shuffled) sequences of 23 proteomes is shown for comparison (small dots in Fig. 4). In Figure 4(B), the same curves are presented in smoothed form. Valine, alanine, glycine, leucine, and isoleucine, all hydrophobic residues, as expected, are the main contributors to the observed maximum (Table I).

Thus, not only the loop sizes show the preference to the standard 25–30 residues, but the sequence as well maintains this standard distance between, primarily, hydrophobic residues (Fig. 5). Interestingly, the minimum at 15–22 residues is also common for both curves, that, apparently, means that both the loop sizes and the hydrophobic character of their ends had been under the same selection pressure. In our previous work, the only criteria used for the loop mapping were end-to-end distances and closeness to nearly standard loop size. The hydrophobic interactions were not at all taken into account. The above correlations suggest that they should have been considered as a major criterion for the mapping of the loops. As it follows from

TABLE I. Absolute Contributions to the Maximum at 25–30 Residues

Amino acid	Total excess at 25–30	Error bar ^a
Val	4,280	568
Ala	4,160	714
Gly	3,360	615
Leu	2,495	814
Ile	2,015	517
Arg	1,595	441
Pro	1,545	364
Lys	1,445	456
His	510	168
Cys	335	96
Total	24,665	1,993

^aError bars are calculated as square roots of average occurrences.

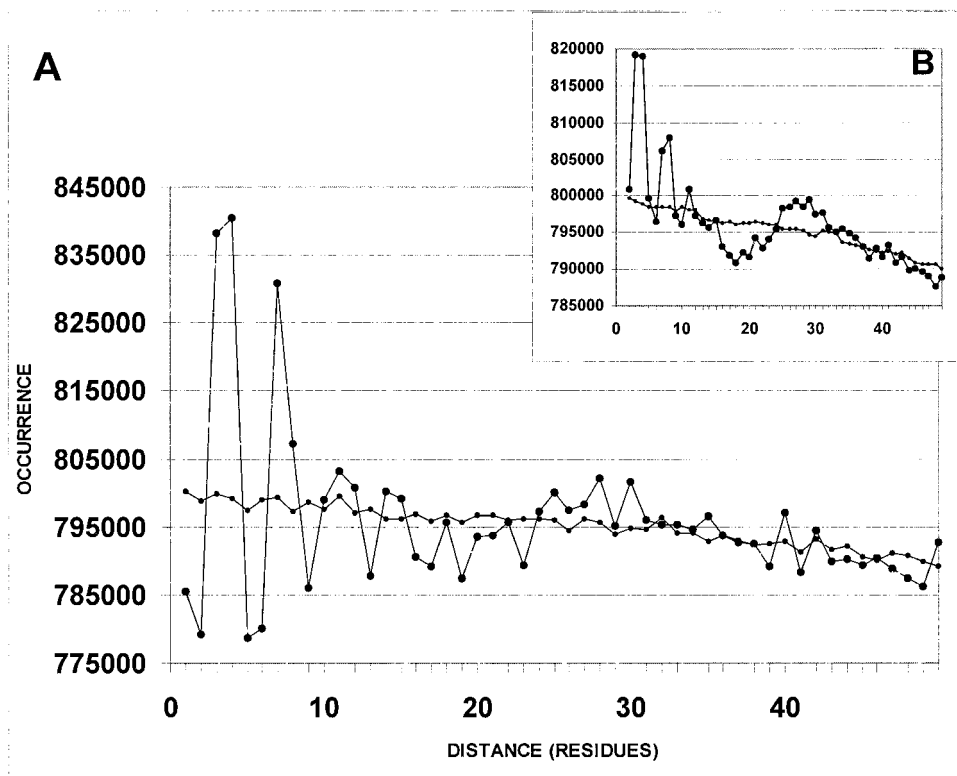


Fig. 4. **A:** Autocorrelation function for all 20 amino acids, calculated for 23 genomes (Archaea: *A. pernix*, *A. fulgidus*, *M. thermoautotrophicum*, *P. horikoshii*; Eubacteria: *A. aeolicus*, *B. burgdorferi*, *C. jejuni*, *C. muridarum*, *D. radiodurans*, *E. coli*, *H. influenzae*, *H. pylori*, *M. tuberculosis*, *M. pneumoniae*, *N. meningitidis*, *R. prowazekii*, *Synechocystis*, *T. maritima*, *T. pallidum*, *U. urealyticum*, *V. cholerae*, *X. fastidiosa*) available as of July 2000 through the Entrez Browser and provided by the National Center for Biotechnology Information. The plot demonstrates a clear excess of occurrences in the region 24–31 residues. The respective plot for randomized (shuffled) sequences of 23 genomes is shown for comparison (small dots). **B:** The same curves in smoothed form. Smoothing is done by running window of three residues. Several sequence shufflings are made both of close-range shuffling and long-range shuffling with the same general result: disappearance of the minimum at about 16–23 residues and maximum at 24–31 residues. The shuffling plot shown is derived from the sequences where every residue is replaced by a randomly chosen residue within the neighboring ten amino acids. The statistical error bar (± 515) around the peak 25–30, calculated as the square root of expected occurrences ($3 \cdot 795,000$), does not exceed the size of the dots on the curve.

comparisons with hydrophobicity plots, not all hydrophobic sites participate in the linear arrangement of the nearly standard loops and their ends. The additional hydrophobic sites correspond, apparently, to secondary loop-to-loop interactions, that is interactions between ends of loops, substantially larger than standard 25–30 residue size. On the other hand, some end-to-end contacts may involve also polar residues. Such sites would not appear as hydrophobic ones, though they may well correspond to strong contacts if van der Waals interactions are accounted for. Additional analysis is required to further elucidate distributions of various amino acids (clusters, neighbor preference, density, etc.).

The correlations we found out demonstrate the fundamental importance of hydrophobic nuclei, on the one hand, and of the loops as protein structure primary building blocks, on the other hand. This immediately suggests a protein folding scenario that has not been considered in earlier works. A new insight that enters

the scene is the loop fold nature of proteins and the existence of standard size closed loops (not in the sense of the traditional definition).^{12–15} The analysis above also reveals an important new structural meaning for the maxima of the Kyte and Doolittle hydrophobicity plots. These plots as well as positional sequence correlation can be used now as a specific guide in the studies of protein folding, indicating likely positions of the closing contacts. An obvious route for the protein folding would be, thus, formation of the standard size loops with the closed ends (nucleation centers) as an initial stage. The resulting looping structure is similar to the theoretically derived flower-like design of internally correlated globules with potential wells.¹⁶ Sequential co-translational formation of these loops may be an essential initial step in the biogenesis of 3D structure of the protein molecule. The final shape of the molecule is further stabilized by secondary interactions such as loop-to-loop contacts, formation of helices, β -sheets.

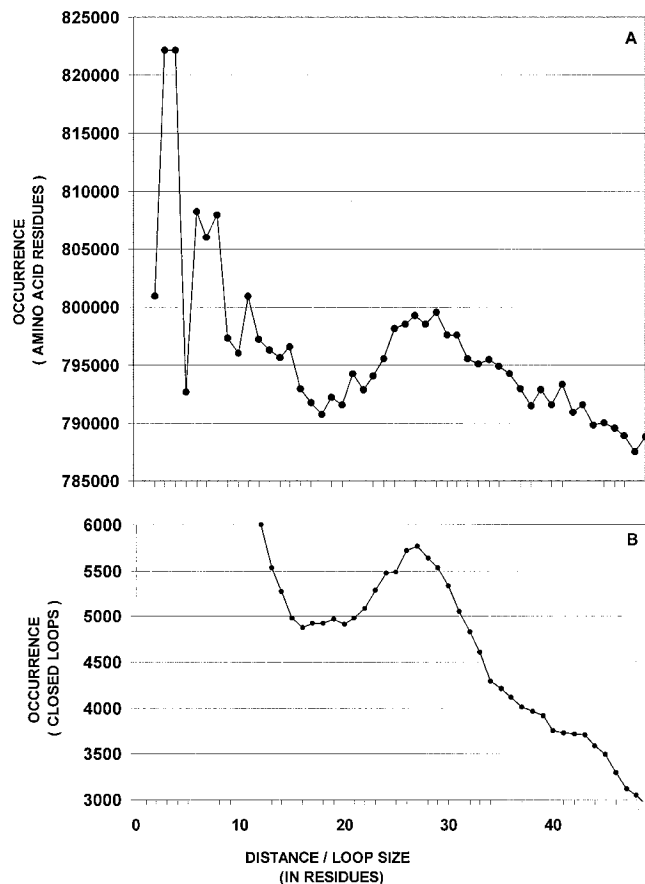


Fig. 5. A comparison of autocorrelation function derived by the analysis of all 23 bacterial genomes (A) and distribution of loop sizes (B) in 302 representative crystallized protein structures (96 eukaryotic, 151 prokaryotic, 28 fungal, 18 viral, and 9 archaeobacterial). Plot B corresponds to figure 1(C) from Berezovsky et al.¹

ACKNOWLEDGMENTS

We are grateful to A.Y. Grosberg for illuminating discussions and E.A. Yakobson for reading the manuscript and

valuable comments. I.N.B. is Post-Doctoral Fellow of the Feinberg Graduate School, Weizmann Institute of Science.

REFERENCES

1. Berezovsky IN, Grosberg AY, Trifonov EN. Closed loops of nearly standard size: common basic element of protein structure. *FEBS Lett* 2000;466:283–286.
2. Bresler SE, Talmud DL. The nature of globular proteins. *Comp Rend Acad Sci URSS* 1944;43:310–314.
3. Epstein HF, Schechter AN, Chen RF, Anfinsen CB. Folding of staphylococcal nuclease: kinetic studies of two processes in acid renaturation. *J Mol Biol* 1971;60:499–508.
4. Berezovsky IN, Namiot VA, Tumanyan VG, Esipova NG. Hierarchy of the interaction energy distribution in the spatial structure of globular proteins and the problem of domain definition. *J Biomol Struct Dyn* 1999;17:133–155.
5. Fersht AR. Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. *Proc Natl Acad Sci USA* 2000;97:1525–1529.
6. Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI Identifying the protein folding nucleus using molecular dynamics. *J Mol Biol* 2000;296:1183–1188.
7. Berezovsky IN. Protein structure: chapters which have been missed. In: Gromiha MM, Selvaraj S, editors. Recent research developments in protein folding, stability and design. Trivandrum: Transworld Research Network. In press.
8. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. *Nature* 1994;372:631–634.
9. Kyte J, Doolittle RF. A simple method for displaying the hydrophobic character of a protein. *J Mol Biol* 1982;157:105–132.
10. Kanehisa MI, Tsong TY. Local hydrophobicity stabilizes secondary structures in proteins. *Biopolymers* 1980;19:1617–1628.
11. Herzog H, Weiss O, Trifonov EN. 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding. *Bioinformatics* 1999;15:187–193.
12. Leszczynski JF, Rose GD. Loops in globular proteins: a novel category of secondary structure. *Science* 1986;234:849–855.
13. Martin ACR, Toda K, Stirk HJ, Thornton JM. Long loops in proteins. *Protein Eng* 1995;8:1093–1101.
14. Kwasigroch JM, Chomilier J, Mornon JP. A global taxonomy of loops in globular proteins. *J Mol Biol* 1996;259:855–872.
15. Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJE. An automated classification of the structure of protein loops. *J Mol Biol* 1997;259:814–830.
16. Grosberg A, Khokhlov A. In: *Statistical physics of macromolecules*. New York: AIP Press; 1994. 350 p.