Closed loops of nearly standard size: common basic element of protein structure

Igor N. Berezovsky[a,*], Alexander Y. Grosberg[b], Edward N. Trifonov[a]

[a] *Department of Structural Biology, The Weizmann Institute of Science, P.O. Box 26, Rehovot 76100, Israel*
[b] *Department of Physics, University of Minnesota, 116 Church Street SE, Minneapolis, MN 55455, USA*

Received 13 December 1999

Edited by Vladimir Skulachev

Abstract By screening the crystal protein structure database for close Cα–Cα contacts, a size distribution of the closed loops is generated. The distribution reveals a maximum at 27±5 residues, the same for eukaryotic and prokaryotic proteins. This is apparently a consequence of polymer statistic properties of protein chain trajectory. That is, closure into the loops depends on the flexibility (persistence length) of the chain. The observed preferential loop size is consistent with the theoretical optimal loop closure size. The mapping of the detected unit-size loops on the sequences of major typical folds reveals an almost regular compact consecutive arrangement of the loops. Thus, a novel basic element of protein architecture is discovered; structurally diverse closed loops of the particular size.

Key words: Protein; Polymer statistic; Unit loop contour length; Protein folding; Protein evolution; Typical fold

1. Introduction

Attempts to understand protein structure date back to L. Pauling. He recognized that a regular pattern of hydrogen bonds, independently on the sequence of side groups, can support a regular shape of the main chain which we know now as an α-helix. Later on, it took a long time and many researchers to realize that there could be more than a few complications. While an α-helix can be made stable in solution where it is supported indeed by the hydrogen bonds as prescribed by L. Pauling, in a condensed environment, such as the interior of a protein globule, the stability of α-helix can be altered dramatically, and in a sequence-dependent way.

Modern description of structural patterns in proteins is substantially more sophisticated. It includes elements of secondary structure [1], loops [2,3], modules [4–6], domains [7–9] and their hierarchy [10–15]. Importantly, there is more than one conceptually different way to look at all these structures. An exhaustive analysis of structural–functional relationships between families of structural analogues and homologues is presented in [16,17]. Several hierarchical classifications of protein structural units on the basis of evolutionary relationships [18,19] were elaborated. In general, considering any particular element of protein structure, one faces the following major questions: to which extent has the evolution influenced the abundance of this element in real proteins? Indeed, every structural element that exists is obviously consistent with physics, with the geometry and energetics of polypeptide chains. However, from what is allowed by physics, evolution may have made certain selection, promoting some structural elements and suppressing others. For instance, some structures may have been selected because they potentially provide a better versatility of design (so-called designability [20]). Thus, the question is, how strongly has the evolution pressure affected the natural choice of native conformations of modern proteins? This question is of an obvious fundamental importance for our understanding of protein evolution. One way to address this question is to compare the ensemble of conformations of real proteins with that expected for random chains of a similar degree of compactness. As described below, such comparison reveals the existence of at least one novel basic element of protein structure: closed loops of typical contour length.

How can we find the ensemble of random compact conformations? In other words, what do we know about a statistically typical compact conformation? This question is easy to answer using the following simple argument: because of ergodicity, a homopolymer, when supported as a globule of the given degree of compactness, samples all the conformations in a completely unbiased way. Therefore, in order to address statistics of all compact conformations, we just have to take a typical conformation of a compact homopolymer; in other words, we have to take conformations which dominate conformational entropy of a homopolymer globule.

Homopolymers in compact globules have been studied extensively in polymer physics [21]. One of the simplest known characteristics of these conformations is the so-called loop factor $P_\delta(l)$: with the specified cut-off length, δ, this is the probability to find two monomers (say, Cα atoms) within a distance δ from one another provided that they are some l monomers apart along the chain. Qualitatively, the $P_\delta(l)$ behaves in the following way: for very small l, it is very small because it is difficult to bend the very short part of a polymer; for a long l, loop factor is small again, because formation of long loops is entropically suppressed. Finally, $P_\delta(l)$ levels off at large l, corresponding to chain passing through the globule. There is an 'optimal' length at which $P_\delta(l)$ reaches a maximum. The problem of loop factor was extensively studied in polymer physics in conjunction with loop closure experiments on DNA. In this context, it was shown both theoretically [22,23] and experimentally [23,24] that the ring-closure probability as a function of the chain length reaches its maximum at 3–4 persistence lengths (1.5–2.0 Kuhn segments). Although the circularization experiments for polypeptide chains are not available, every chain conformation in the homopolymer globule would provide a spectrum of closed loops which can

*Corresponding author. Fax: (972)-8-9342653.
E-mail: igor.berezovsky@weizmann.ac.il

PII: S0014-5793(00)01091-7

be viewed as an outcome of the ring-closure experiment. Thus, our plan is to compare the theoretical expectations for such a gedanken ring-closure experiment in a homopolymer globule with the statistical data on loops in native conformations from the protein data bank (PDB).

We analyzed a representative set of crystallized protein structures with low sequence identity [25] looking for the closed loops. Surprisingly, the distribution of the loops with close Cα–Cα contacts (less than $\delta = 10$ Å) showed a distinct maximum at 25–30 residues. These unit-size loops are structurally diverse, being built of various combinations of secondary structure elements, however, they have one common property: the same typical contour length.

2. Materials and methods

Protein structures of 200 or more contiguous residues and less than 25% sequence identity (PDB_SELECT) were extracted from the PDB. The total number of analyzed structures is 302 (96 eukaryotic, 151 prokaryotic, 28 fungal, 18 viral and nine archaebacterial). The list of PDB identifiers is presented below:

1SMNA, 1SQC, 1TAHA, 1TDE, 1TML, 1UAE, 1UXY, 1XIKA, 1XYZA, 1YSTH, 1YUB, 1ZID, 1ZIN, 2ABK, 2AT2A, 2AYH, 2BBKH, 2DORA, 2DPG, 2DRI, 2FRVA, 2LIV, 2NACA, 2OMF, 2PIA, 2PLC, 2POLA, 2POR, 2THIA, 3DAAA, 3GSAA, 3MINB, 3PBGA, 3PTE, 3SEB, 3SIL, 3TDT, 4PGAA, 4XIS, 6MHTA, 7AHLA, 8ABP, 1AOL, 1AYM2, 1AYM3, 1BEV1, 1CKNA, 1CRXA, 1HAVA, 1HEIA, 1MML, 1NOYA, 1SMVC, 1SVB, 1TME1, 1VPSB, 2BBVA, 2BPA1, 2EIAA, 2VIUA, 1AK0, 1A2Z, 1AORA, 1FTRA, 1JUK, 1MROA, 1MROB, 1MROC, 1THJ, 1XGS, 16PK, 1A02N, 1A26, 1A28A, 1A4MA, 1A4SA, 1A8E, 1A9S, 1ABRB, 1ADOA, 1ADS, 1AFRA, 1AKZ, 1AN9A, 1AO6A, 1AOZA, 1AQ0A, 1ATLA, 1AXN, 1B0M, 1BG0, 1BG2, 1BGP, 1BP1, 1BQUB, 1BX9, 1BYB, 1BYQA, 1C3D, 1CD1A, 1CFB, 1CMKE, 1CNE, 1CNV, 1CSBB, 1CSH, 1CYDA, 1D2NA, 1DFJI, 1DHR, 1EBPA, 1EFVA, 1EFVB, 1FC1A, 1FSSA, 1FT1A, 1FT1B, 1FZAB, 1GNHA, 1GOTB, 1HKBA, 1HSBA, 1HXN, 1IMDA, 1INP, 1IRK, 1ITBB, 1JKW, 1JMCA, 1LAM, 1LXTA, 1MLDA, 1MRJ, 1NAR, 1NLS, 1OCCC, 1OSPO, 1PPN, 1PTY, 1QCRC, 1RGS, 1RLAA, 1RPT, 1SMD, 1SMEA, 1TADA, 1TFB, 1THV, 1UBY, 1UROA, 1VID, 1VIN, 1VKXB, 1WAB, 1WER, 1XVAA, 1YASA, 1YVEI, 2BAA, 2CTC, 2LBD, 2MASA, 2PGD, 3GRS, 3MDDA, 6GSVA, 1ARV, 1ASYA, 1AUA, 1BOL, 1BQ3, 1BXO, 1CPO, 1CSN, 1GOH, 1IDK, 1IPSA, 1OYC, 1PLR, 1PMI, 1PYP, 1RMG, 1RYP1, 1RYP2, 1RYPF, 1RYPI, 1TCA, 1TIB, 1TKAA, 1YSC, 2CYP, 2VAOA, 6CEL, 1A7TA, 1A8D, 1AD2, 1AGJA, 1AH7, 1AHJB, 1AI7A, 1AIJ, 1AIJS, 1AJ2, 1AK1, 1AKO, 1AL3, 1ALO, 1AMP, 1AN8, 1ANF, 1AOQA, 1AR1A, 1ARB, 1ARZC, 1ATG, 1AURA, 1AVMA, 1AXWA, 1B2NA, 1BC5A, 1BD0A, 1BF8, 1BFD, 1BJK, 1BRT, 1BTMA, 1BU7A, 1CBY, 1CEM, 1CFR, 1CG2A, 1CHKA, 1CHMA, 1CLC, 1CP2A, 1DAD, 1DEAA, 1DKZA, 1DLC, 1DOSA, 1DXY, 1ECEA, 1ECPA, 1ECRA, 1EDG, 1EDT, 1ELYA, 1ESC, 1EUU, 1EZM, 1FCDA, 1FDZB, 1FGJA, 1FRVB, 1FTS, 1FUIA, 1FURA, 1FVPA, 1GDOA, 1GSA, 1GUQB, 1HDEA, 1HRDA, 1ISO, 1IXH, 1JETA, 1JFRA, 1KIT, 1KNYA, 1KVU, 1LBU, 1LRV, 1LXA, 1MOQ, 1MSK, 1MTYB, 1MTYD, 1NBAB, 1NIF, 1NOX, 1OFGA, 1ONRA, 1OPR, 1OTP, 1PCL, 1PEA, 1PGS, 1PHC, 1PNKB, 1POT, 1PRCC, 1PUD, 1PYAB, 1PYSA, 1QAPA, 1QBA, 1QNF, 1REQB, 1RVAA, 1SBP, 1SKYE, 1SLY.

We define closed loops as continuous subtrajectories of the folded chains with a small distance between their ends. The analysis of the loops was performed in two steps. First, we calculated the histograms which display the occurrence of loops of various sizes at a given distance between their ends. Second, we mapped the detected unit-size loops (27 ± 5 residues) along the sequences. We performed the mapping procedure for the representatives of nine major fold families [26] and membrane protein porin [27]. The mapping procedure selects sequentially the tightest loops of 22–32 residues (closest end-to-end distances, here and below, the distance between Cα atoms is meant). At each step, the sequence region corresponding to the mapped loop is excluded from further calculations. As a result of two occasionally overlapping loops, the tightest one is selected for the mapping. A marginal overlapping of 1–5 residues is, however, allowed. The routine starts with the end-to-end distance less than 4 Å and is normally exhausted at the distance 10 Å between the loop ends.

3. Results and discussion

Fig. 1 presents the contour length distributions of the detected loops (in number of amino acid residues), with distances between their ends less than 5, 7 and 10 Å, respectively. The histograms show prominent maxima in the range of loop sizes 22–32 amino acid residues, essentially independent of the distance δ within 5–10 Å. The loop size analysis, thus, reveals the novel property of the protein structure, previously overlooked: the closed loops of preferred size.

It should be noted that the observed loops are not necessarily independent, in particular, some of the large loops may include smaller loops. In order to find out where along the sequences the unit-size loops are positioned, we analyzed the set of nine representatives of standard folds [26] and membrane protein porin (2omf.pdb) [27]. Fig. 2 demonstrates positional distributions of the loops of 22–32 amino acid residues along 10 sequences, corresponding to major folds and membrane protein. The sequences are covered from 40% (in 1ubq.pdb) to 85% (2rhe.pdb) by the unit-size loops (total involvement for all 10 sequences, 66%). Thus, the typical size

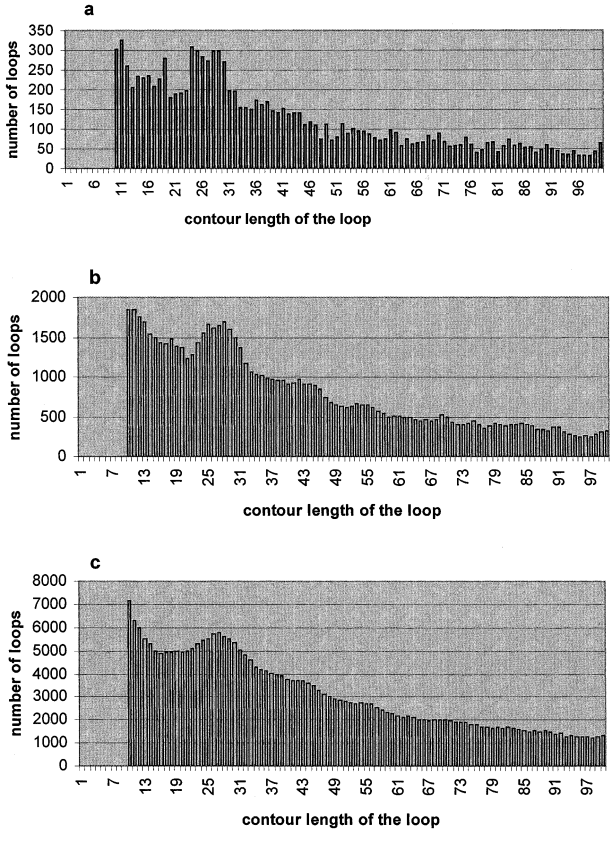

Fig. 1. Histograms for the number of loops in the representative PDB set depend on the contour loop length: a: end-to-end distances less than 5 Å; b: 7 Å; c: 10 Å.

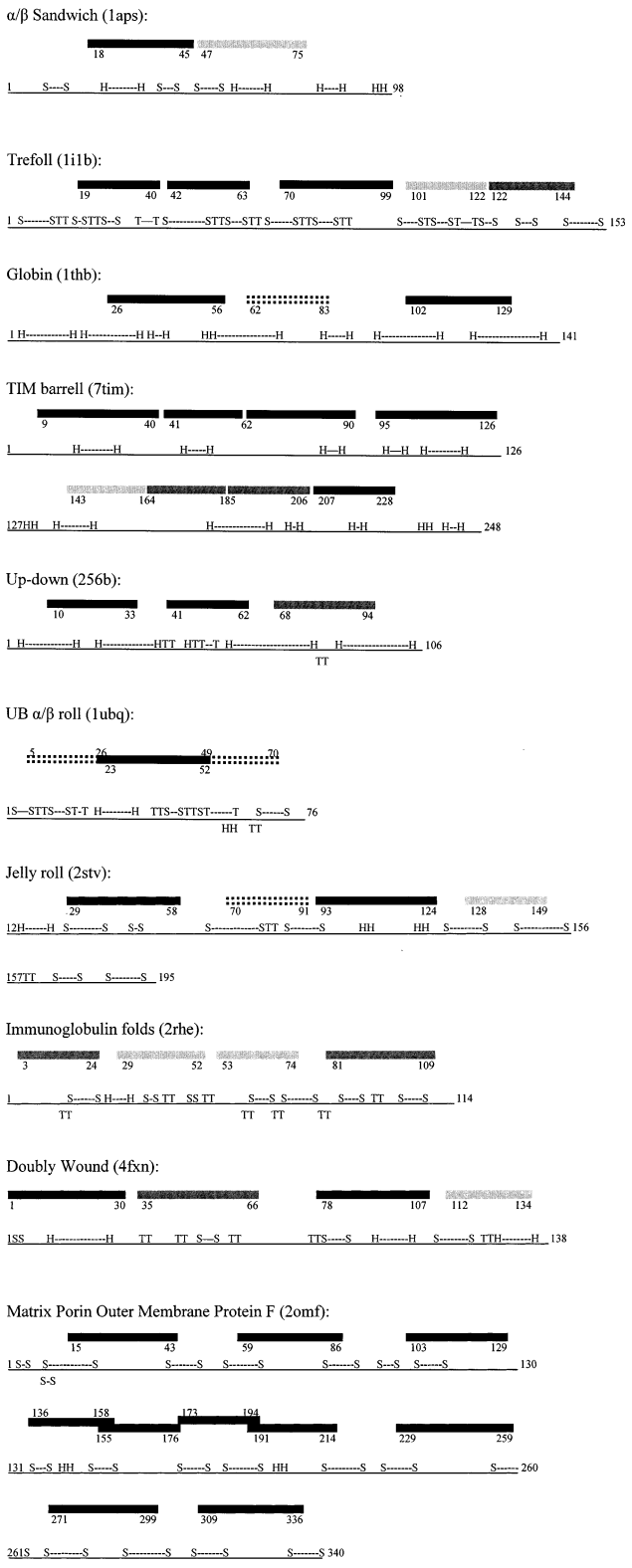

Fig. 2. Results of the mapping procedure for representatives of nine major folds and membrane protein porin: black, loops with end-to-end distance between 4 and 6 Å; dark gray, loops with end-to-end distance between 6 and 8 Å; light gray, loops with end-to-end distance between 8 and 10 Å; dotted, relaxed loops with end-to-end distances 10.156–10.308 Å. Secondary structure elements (H: α-helices, S: β-strands, T: turns) are indicated according to PDB assignment.

loops, indeed, constitute major standard elements of protein structure.

In the individual maps, few gaps are observed. This can be understood because, as suggested by the loop size distribution in Fig. 1, the loops of other sizes exist as well. Indeed, many of the gaps in the sequence maps in Fig. 2 could be filled by such larger or smaller than standard loops (data not shown). In addition, relaxed loops, those with end-to-end distances even slightly beyond the 10 Å cut-off, may contribute to the gaps as well. For example, in UB α/β roll fold (1ubq.pdb), loops 5–26 and 49–70 with the end-to-end distances 10.156 and 10.006 Å, respectively, fill the gaps to 86% of total length. Similarly, the loops 70–91 (10.209 Å, in jelly-roll fold, 2stv.pdb) and 62–83 (10.308 Å, in globin fold, 1thb.pdb) could be added as well to the maps. The standard and non-standard loops, obviously, leave uncovered only a few percent of the sequence length: extended stretches which cannot be qualified as loops. We may conclude, thus, that the globular and membrane proteins are largely made of the closed loops, preferentially of standard size, irrespective of the type of the fold.

Fig. 2 also demonstrates that the closed loops are structurally heterogeneous. They may consist of virtually any combination of helices, turns, β-strands within their contours, while keeping the contour length largely uniform. For example, standard α/β sandwich fold (1aps.pdb) has two closed loops with α-helix, β-strand and non-structured regions therein. Doubly wound fold (4fxn.pdb) is divided into four unit loops. The loop 1–30 contains α-helix 10–27 and β-strand 1–5; the loop 35–66: β-strand 49–55 and turns 34–37, 43–46, 56–59; the third loop 78–107: α-helix 93–104, β-strand 80–88 and turn 77–80; and the last one 112–134: α-helix 124–136, β-strand 109–119 and turn 121–124. The loop size uniformity renders proteins a topological regularity, essentially independent on how the elements of secondary structure are arranged. In this sense, the ends of the loops serve as sort of punctuation marks organizing the overall protein structure.

Let us now return to the question formulated in Section 1 and compare the loop factor averaged over the native conformations of the PDB, as presented in Fig. 1, with the theoretical expectation for a thermodynamically equilibrium (or, in other words, perfectly molten) homopolymer globule. Does the observed preferential loop size in proteins correspond to the statistical loop closure size expected in thermodynamic equilibrium? The dominant loop size for an equilibrium globule depends on the chain flexibility, that is on its persistence length. The chains shorter than a couple of persistence lengths are too rigid to effectively form loops. By the same token, for chains in excess of several persistence lengths, loop closure gets again improbable, because chain ends are wondering in space essentially independently from each other. Thus, in a scaling sense, the maximum of loop factor is determined by the persistence length scale. Accurate calculations, performed by Shimada and Yamakawa in the context of DNA ring-closure problem [22,23], demonstrated that the circularization maximum lies in the range of $3 \div 4$ persistence lengths, or $1.5 \div 2$ statistical segments. Of course, this result was obtained for Gaussian coils, while we are dealing here with very dense globules. Nevertheless, our considerations, including the reference to the result by Shimada and Yamakawa, remain valid, although for the very subtle reason. According to the so-called Flory theorem, chain in the globule shields itself from inter-

actions in such a way that its statistics, including the loop factor, remain Gaussian on all scales smaller than the globule overall size (see [28,21] for the detailed discussion of this delicate and important point).

What is the persistence length for polypeptide chains of interest here? Experimental data as well as theoretical calculations for homopolymers with various amino acid compositions show that the values of limiting characteristic ratio ($C_\infty = (\langle r \rangle_0 / nl^2)_{n\to\infty}$ in Flory's notations, [29]), i.e. statistical segment (twice persistent length), vary from 8 to 12 residues [30,31], with an exception for Pro ($C_\infty = 116$, [32]) and Gly ($C_\infty = 2$, [31]). For the copolymer Gly-Ala, this value ranges between nine and two, depending on the Gly content [33]. Since the content of Gly and Pro in real proteins is low, the expected statistical segment of a mixed polypeptide chain would be within the range of 8–12 residues. One, thus, estimates the optimal loop size for the statistically typical conformation about 10–25 amino acid residues. Inclusion of comparatively rigid sections of chains involved in α-helices and β-sheets should effectively increase the loop size to about twice this value (taking the average content of structured sections equal 50%), i.e. 20–50 residues.

Thus, the estimated loop length for the thermodynamically equilibrium, or entropically dominant, or statistically typical loops of about 20–50 is fairly close to the value 27 ± 5 residues observed in our data for the native protein conformations.

We can make further comparisons by looking quantitatively at the decay of the loop factor with l. For statistically typical conformations (and disregarding the overall normalization factor), we expect that the loop factor behaves as $P_\delta(l) \approx (\delta/a) \cdot l^{-3/2}$, with a about one monomer size. This should hold in the range of l, roughly, between circularization maximum and the globule size. Based on the data presented in Fig. 1 (and other similar data for several other values of δ, not shown), we found that both the $\approx \delta^3$ dependence on δ and $l^{-3/2}$ dependence on l are followed reasonably accurately by the loop factors averaged over the ensemble of native protein conformation.

For the conclusion, it is tempting to speculate that in the early stages of evolution the initially short polypeptide chains would acquire some evolutionary advantage by closing into loops of 20 or so residues. The loop stability, for example, could have been the advantage. Further evolution of the proteins, perhaps, involved fusion of the genes encoding the loops, with formation of larger proteins consisting of several loops. In modern proteins, the fused loops, apparently, are still close to their original sizes, selected in the earlier evolutionary stages. Secondary structure elements, requirements of stability of the protein globule, may influence the original size, so that the typical contour length of the closed loops observed today could be the result of many selection pressures acting in different directions and producing results which may be difficult to see using simple statistical tests. The punctuation of the polypeptide chain by the closed loops suggests a straightforward scheme of folding of the proteins. The 'stitches' by the closed loops could nucleate the folding. The existence of the major class of the nearly standard size closed loops should be taken into account in any protein folding scheme.

Acknowledgements: The authors are thankful to J. Schuchhardt and E. Yakobson for discussion. I.B. is supported by the Ministry of Absorption.

References

[1] Lim, V.I. (1974) J. Mol. Biol. 88, 857–872.
[2] Kwasigroch, J.-M., Chomilier, J. and Mornon, J.-P. (1996) J. Mol. Biol. 259, 655–672.
[3] Martin, A.C.R., Toda, K., Stirk, H.J. and Thornton, J.M. (1995) Protein Eng. 8, 1093–1101.
[4] Patthy, L. (1994) Curr. Opin. Struct. Biol. 4, 383–392.
[5] Doolittle, R.F. (1995) Annu. Rev. Biochem. 64, 287–314.
[6] Bork, P. (1992) Curr. Opin. Struct. Biol. 2, 413–421.
[7] Janin, J. and Wodak, S. (1983) Prog. Biophys. Mol. Biol. 42, 21–78.
[8] Holm, L. and Sander, C. (1994) Proteins 19, 256–268.
[9] Jones, S., Stewart, M., Michie, A., Swindells, M.B., Orengo, C. and Thornton, J.M. (1998) Protein Sci. 7, 233–242.
[10] Crippen, G.M. (1978) J. Mol. Biol. 126, 315–332.
[11] Rose, G.D. (1979) J. Mol. Biol. 134, 447–470.
[12] Berezovskii, I.N., Esipova, N.G. and Tumanyan, V.G. (1997) Biophysics (Moscow) 42, 557–565.
[13] Berezovsky, I.N., Tumanyan, V.G. and Esipova, N.G. (1997) FEBS Lett. 418, 43–46.
[14] Berezovskii, I.N., Esipova, N.G. and Tumanyan, V.G. (1998) Biofizika (Moscow) 43, 392–402.
[15] Berezovsky, I.N., Namiot, V.A., Tumanyan, V.G. and Esipova, N.G. (1999) J. Biomol. Struct. Dyn. 17, 133–155.
[16] Russell, R.B. and Barton, G.J. (1994) J. Mol. Biol. 244, 332–350.
[17] Russell, R.B., Saqi, M.A.S., Sayle, R.A., Bates, P.A. and Sternberg, M.J.E. (1997) J. Mol. Biol. 269, 423–439.
[18] Murzin, A., Brenner, S.E., Hubbard, T.J.P. and Chothia, C. (1995) J. Mol. Biol. 247, 536–540.
[19] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997) Structure 5, 1093–1108.
[20] Li, H., Helling, R., Tang, C. and Wingreen, N. (1996) Science 273, 666–668.
[21] Grosberg, A. and Khokhlov, A. (1994) Statistical Physics of Macromolecules, AIP Press (particularly sections 20–22).
[22] Shimada, J. and Yamakawa, H. (1984) Macromolecules 17, 689–698.
[23] Yamakawa, H. and Stokmayer, W.H. (1972) J. Chem. Phys. 57, 2843–2854.
[24] Shore, D., Langowski, J. and Baldwin, R.L. (1981) Proc. Natl. Acad. Sci. USA 78, 4833–4837.
[25] Hobohm, U. and Sander, C. (1994) Protein Sci. 3, 522–524.
[26] Orengo, C.A., Jones, D.T. and Thornton, J.M. (1994) Nature 372, 631–634.
[27] Cowan, S.W., Schirmer, T., Rummel, G., Steiert, M., Ghosh, R., Pauptit, R.A., Jansonius, J.N. and Rosenbusch, J.P. (1992) Nature 358, 727–733.
[28] de Gennes, P.-G. (1979) Scaling Concepts in Polymer Physics, Cornell Univ. Press.
[29] Flory, P. (1969) Statistical Mechanics of Chain Molecules, Interscience.
[30] Brant, D.A. and Flory, P.J. (1965) J. Am. Chem. Soc. 87, 2788–2791.
[31] Miller, W.G. and Goebel, C.V. (1968) Biochemistry 7, 3925.
[32] Schimmel, P.R. and Flory, P.J. (1967) Proc. Natl. Acad. Sci. USA 58, 52–59.
[33] Miller, W.G., Brant, D.A. and Flory, P.J. (1967) J. Mol. Biol. 23, 67–80.