Volume 325, number 1,2, 23–28 FEBS 12371 June 1993

Minireview

Why are the same protein folds used to perform different functions?

Alexei V. Finkelstein, Alexander M. Gutun* and Azat Ya. Badretdinov

Institute of Protein Research, Russian Academy of Sciences, 142292 Pushchino, Moscow Region, Russian Federation

Received 14 March 1993

A small number of folding patterns describe in outline most of the known protein globules, the same folds being found in non-homologous proteins with different functions. We show that the 'popular' folding patterns are those which, due to some thermodynamic advantages of their structure, can be stabilized by a lot of random sequences. In contrast, the folds which are rarely or never observed in natural globular proteins can be stabilized only by a tiny number of random sequences. The advantageous folds are few, they tolerate various primary structures, and therefore they can and ought to perform different functions. A connection between the inherent 'weak points' of protein folding patterns and positions of active sites are discussed.

Folding pattern; Physical selection; Random sequence; Conformational temperature; Energy; Entropy; Active site

1. INTRODUCTION

Inspection of protein structures shows that a small number of folding patterns describe in outline most of the known domains, the same patterns being found in proteins which have no genetic relationship and perform quite different functions [1–8]. Rossmann folds, TIM barrels, helical bundles and Greek key motifs are examples of the most popular folding patterns. It has been shown [7] that the wide-spread protein architectures are those that have some advantages in thermodynamic stability.

For example, a right-handed connection of parallel β-strands is a standard detail of the most popular folds, while a left-handed one is extremely rare (Fig. 1A). At the same time, a right-handed connection in a sheet with a right-handed twist (this twist is energetically favorable for natural L-amino acids [9]) demands less loop bending [10] and, due to polypeptide chain rigidity, it 'costs' ~ 2 kcal/mol less than a left-handed one [7,8]. The same predominance of less-bent loops is also observed for other standard connections of secondary structure elements [11–14]. Similarly, loop crossing (Fig. 1B) is rare in proteins, and it seems that this 'defect' is prohibited because the crossing either buries and dehydrates a loop peptide group (which costs ~ 5 kcal/mol), or demands additional loop bending to avoid this dehydration which also costs a few kcal/mol [7].

Thus, the most popular protein folds have some obvious thermodynamic advantages, yet it is not clear why these small advantages provide the observed rigid limitations in patterns of protein folding.

First, a defect costs only a few kcal/mol, while different sequences can readily add or subtract ~ 50 kcal/mol to the energy of a fold [15]. Then, why does the small energy of a defect play any selective role, and why can it not be compensated for by an 'appropriate' amino acid sequence?

Second, the arguments based on loop rigidity must mainly concern their entropy, because the main reason for polymer elasticity is that an additional bending decreases the number of possible chain conformations [16,17]. Yet the entropy of a native protein globule which has a unique fold is zero in any case! Why do the 'entropic' arguments against some protein architectures make any sense at all?

2. HOW MANY AMINO ACID SEQUENCES CAN STABILIZE ...?

To clarify the origins of physical selection of protein structures, we consider an *amount* of randomly synthesized sequences which stabilize a folding pattern of this or that kind. Some preliminary considerations of this approach and a few examples can be found in [7,8,18].

The question 'How many sequences can form this or that?' is stimulated by recent development of the general physical theory of structures formed by random heteropolymers [19–23]. Investigations of this kind are related

Correspondence address. A.V. Finkelstein, Institute of Protein Research, Russian Academy of Sciences, 142292 Pushchino, Moscow Region, Russian Federation.

**Present address:* Department of Chemistry, Harvard University, Cambridge, MA 02138, USA.

to globular proteins, since their amino acid sequences (unlike e.g. the periodic chains of fibrous proteins) look like random heteropolymers which are 'edited' only slightly [24]. In this connection it is noteworthy that, according to the theory in [23], a significant part of randomly synthesized heteropolymers can have a thermodynamically dominant fold.

Below we investigate how easily the random polypeptides can form different folding patterns and show that both energetic and entropic defects of the patterns result in a drastic, exponential decrease of the number of pattern-stabilizing sequences, so that at most they can stabilize only a few more-or-less 'perfect' folding patterns, if at all.

3. BOLTZMANN-LIKE STATISTICS OF PROTEIN DETAILS

For small elements of protein structure, the occurrence-to-energy relationship was established two decades ago [25]. It looks like an exponential predominance in the occurrence of low-energy elements over high-energy ones;

$$OCCURRENCE \sim \exp(-ENERGY/RT_*) \quad (1)$$

here R is the gas constant, and T_*, the 'conformational temperature', is equal to room temperature in order of magnitude. This relationship concerns statistics of ϕ, ψ, χ angles [25,26], occurrence of empty cavities [27], of *cys*- and *trans*-prolines [28], distribution of residues between the globular surface and interior [29], between secondary structure regions [30,31], etc.

Although this relationship looks like a conventional Boltzmann statistic of thermodynamic fluctuations, it must have quite a different origin, since the observed protein structures do not fluctuate, in the sense that the links of a protein chain do not wander from surface to interior of the protein, and from one secondary structure to another. Rather, the basis for the observed protein statistic is that any low-energy element exponentially enlarges the number of sequences which ensure protein stability, while any high-energy one reduces this number [18].

A similar approach can clarify the origin of 'physical selection' of protein architectures.

4. ENERGY SPECTRA OF HETEROPOLYMER GLOBULES

A given fold can be stable only for those sequences where the total energy E (one can imagine that

$$E = \sum_p \varepsilon_p$$

summarizes energies of all the residue-to-residue contacts and bends inherent in this fold) is at the very bottom of the energy spectrum, below the energy level of any other fold of the chain.

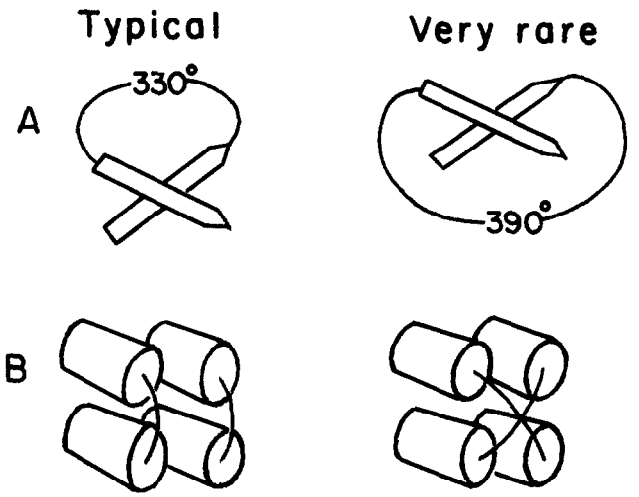

Fig. 1. Two examples of folding patterns typical for globular proteins in comparison with similar ones which are, however, quite rare. β-strands are shown by arrows, α-helices by cylinders, loops by solid lines. (A) Left-handed connection of parallel β-strands is rare; it demands a greater bend of a loop. (B) Crossing of loops is rare; it either dehydrates a peptide group of a loop or demands additional loop bending.

Basic properties of energy spectra of heteropolymer globules [21] can be summarized as follows.

The form of an energy spectrum (Fig. 2a) is governed by overall properties of a chain (such as the content of attracting and repulsing residues). Overall properties of a vast majority of long random sequences are close as a result of statistical averaging. Therefore their energy spectra are also similar.

For any chain, most energy levels occur within an interval of $(\bar{E}^\circ - \sigma, \bar{E}^\circ + \sigma)$. $\bar{E}^\circ$ is the mean value of chain energy averaged over all globular folds, and σ is the root mean square deviation of fold energies from this mean value. Density of energy levels is proportional to

$$\bar{m}_E = M \exp(-(E - \bar{E}^\circ)^2/2\sigma^2) \quad (2)$$

where M is the total number of globular folds. The lowest energy fold of a sequence relates to the energy range

$$E^* = \bar{E}^\circ - \sigma\sqrt{2\ln M} \quad (3)$$

In this region $\bar{m}_E \sim 1$, and when $\bar{m}_E$ is much less than unity, this means that energy levels are absent for nearly all sequences.

Abatement of density of energy levels determines a 'critical temperature' corresponding to the end of energy spectrum

$$T_c = (R\,\partial\ln(\bar{m}_E)/\partial E|_{\bar{m}_E=1})^{-1} = \sigma^2/(R(\bar{E}^\circ - E^*)) = \sigma/\sqrt{2\ln M} \quad (4)$$

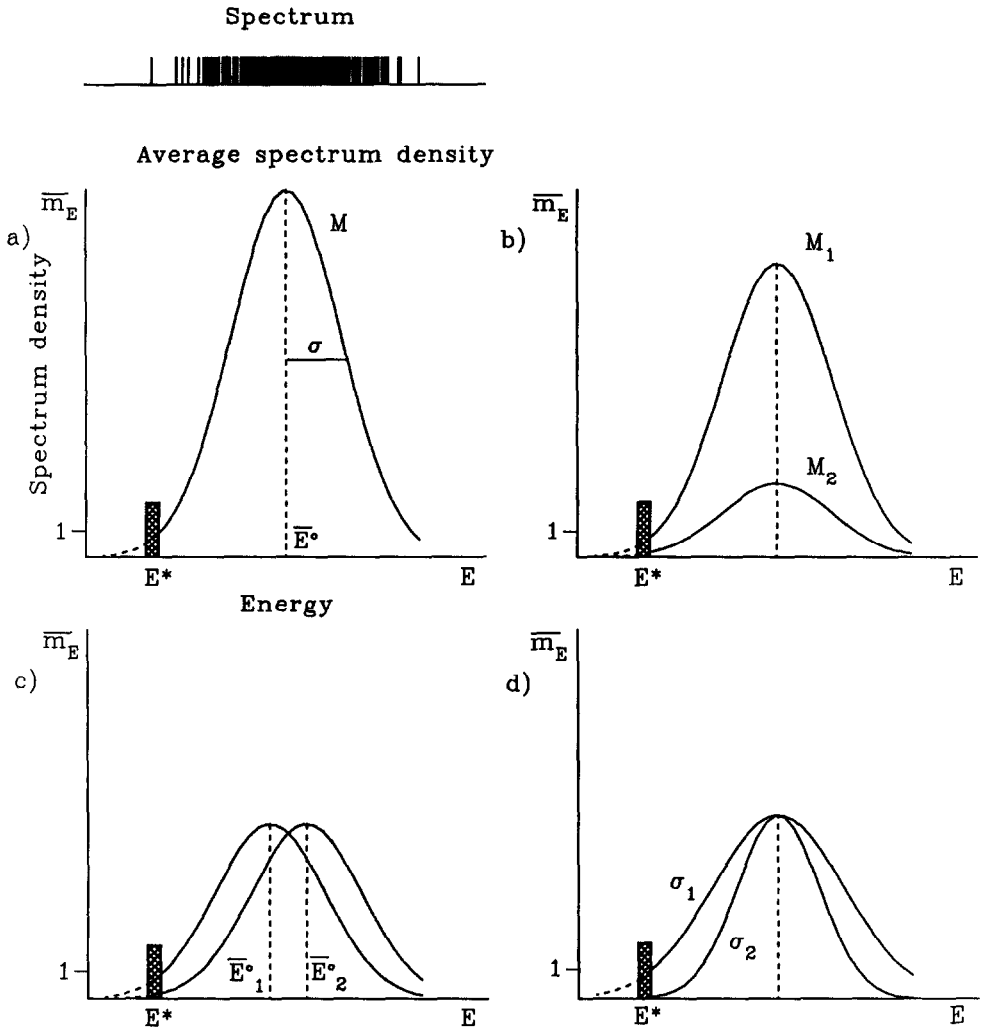

Fig. 2. Typical distribution of energy levels for globular folds of a random heteropolymer. (a) The simplest case: all folds have the same overall properties (compactness, etc.). Each spectrum line corresponds to a fold of a random sequence; a plot of spectrum density averaged over the random sequences. $\bar{E}°$ is the mean energy of the spectrum; σ is a characteristic width of it: most levels occur in the range of $\bar{E}° \pm \sigma$. A typical position of the lowest energy fold is shown by hatching. The broken line corresponds to a region where $\overline{m}_E \ll 1$; this region is accessible only for a small fraction of random sequences. Other plots illustrate a 'physical selection' of folding patterns: a smaller amount of low-energy folds gives a sequence a smaller chance of finding its energy minimum within a handicapped group of folds. Three basic cases are shown. (b) Group 2 has an 'entropic defect': it contains much less folds than another one ($M_2 \ll M_1$); other properties of the folds (mean energies and dispersions) are the same for both groups, $\bar{E}°_1 = \bar{E}°_2$, and $\sigma_1 = \sigma_2$. (c) Group 2 includes folds with an 'energetic defect': $\bar{E}°_2 > \bar{E}°_1$; other properties of the folds of both groups are the same, $M_1 = M_2$, and $\sigma_1 = \sigma_2$. (d) Group 2 includes folds with a smaller variety of interactions: $\sigma_2 < \sigma_1$; other properties of the folds of both groups are the same, $\bar{E}°_1 = \bar{E}°_2$, and $M_1 = M_2$.

Below T_c, the lowest energy folds of the chains are 'frozen out'; above T_c, these folds are not stable thermodynamically. T_c depends on heterogeneity of residues and folds rather than on the chain length because σ^2 is proportional to this heterogeneity, and σ^2, $\ln M$, and $\bar{E}°-E*$ are all proportional to the protein size [19,21].

5. THE ORIGINS OF BOLTZMANN-LIKE STATISTICS OF PROTEIN STRUCTURES

The above results have been obtained for a basic case when all folds have the same overall properties, such as density. To understand a physical selection of protein folding patterns we have to consider a case when the folds are divided into groups with different properties. Let us consider the basic cases.

1. Suppose that all M folds are divided into two groups. The first contains M_1 folds, the second M_2 ones, and $M_2 \ll M_1$ (Fig. 2b). This corresponds, for example, to division of folding patterns into those with right- and left-handed connections of β-strands (Fig. 1A): a left-handed connection includes a smaller number of chain conformations (see above). Let all energetic properties of the folds be the same. Then for each fold, there is an equal probability that it serves as the lowest energy fold

for some random amino acid sequence. However, the number of right-handed folds is much greater than that of their competitors. Proportionally, the lowest energy fold of a random chain has a much greater chance to happen to be a right-handed one.

This can also be explained from another point of view. The expected energy of the lowest energy fold of Group 1 is $\bar{E}^\circ - \sigma\sqrt{2\ln M_1}$ (see eqn. 3), while the lowest energy fold of Group 2 usually has energy of $\bar{E}^\circ - \sigma\sqrt{2\ln M_2}$. As $M_1 \gg M_2$, the best fold of Group 1 usually has a lower energy than the best fold of Group 2. This shows that a smaller variety of folds gives a random sequence a smaller possibility of obtaining a low-energy fold. Thus, an 'entropic' defect (a lack of folds) results in an 'energetic' one. Thus, entropic defects can discriminate protein structures, even though entropy of a native protein globule is zero.

2. Suppose now that folds of Group 2 have some energetic defect which enlarges their mean energy $\bar{E}^\circ_2$ relative to $\bar{E}^\circ_1$, the mean energy of the folds of Group 1 (Fig. 2c). This corresponds, in particular, to division of folding patterns into those with non-crossed and crossed loops (Fig. 1B): the crossed loop loses an H-bond with water (see above). Let both groups have the same number of folds ($M_1 = M_2$) and the same variety of interactions (i.e. $\sigma_1 = \sigma_2$). Then

$$\bar{m}'_{E^*}/\bar{m}''_{E^*} = \exp(-(E^* - \bar{E}^\circ_1)^2/2\sigma^2)/\exp(-(E^* - \bar{E}^\circ_2)^2/2\sigma^2)$$

is the ratio of average numbers of folds of these groups in the energy range related to the lowest energy fold (in this range, $\bar{m}'_{E^*} + \bar{m}''_{E^*} = 1$, see above). Evidently, this ratio can be represented in a simple 'Boltzmann-like' form:

$$\bar{m}'_{E^*}/\bar{m}''_{E^*} = \exp(-(E^\circ_1 - \bar{E}^\circ_2)/RT_*) \quad (5)$$

where

$$T_* = \sigma^2/(R(\bar{E}^\circ - E^*)) \quad (6)$$

and $\bar{E}^\circ$ is equal to $(\bar{E}^\circ_2 + \bar{E}^\circ_1)/2$.

This ratio shows a proportion of random sequences which can stabilize the folds of competing groups.

T_* is a 'conformational temperature', the same as that which governs a Boltzmann-like statistic of small elements of protein globules [18]. Furthermore, this conformational temperature coincides with the critical temperature T_c which limits thermodynamic stability of protein structure (cf. eqns. 4 and 6).

Thus, Boltzmann-like statistics cover not only small details, but also the overall chain folds.

Even the small difference, $\bar{E}^\circ_1 - \bar{E}^\circ_2$ can discriminate folding patterns. One must not compare it with the total energy of protein structure; to discriminate a pattern 'inconvenient' for random sequences, $\bar{E}^\circ_1 - \bar{E}^\circ_2$ has to exceed only RT^*, i.e. only ~ 1 kcal/mol!

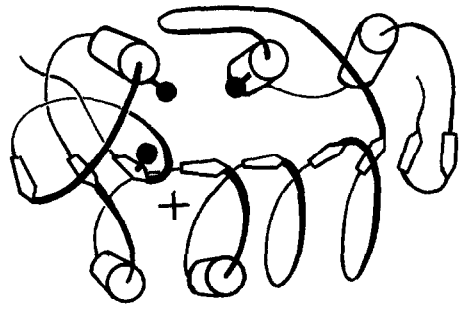

Fig. 3. A scheme of subtilisin folds and the position of its active site. Subtilisin is a rare example of a protein violating both of the structural rules shown in Fig. 1. Note that any loop forming a left-handed connection of parallel β-strands always has to cross at least one of the surrounding right-handed ones; this makes such a connection twice as unfavorable. The catalytic site of subtilisin is located just near the main 'defect' of the fold where a left-handed connection crosses a normal right-handed one. The residues of the catalytic site (His^{62}, Asp^{32}, Ser^{221}) are shown by filled circles. The substrate binding site is shown by a cross and is situated in a crevice formed by diverged loops in a 'switch point' of the parallel β-sheets [32].

Of course, any defect can be compensated for by an appropriate amino acid sequence but, according to the above estimates, the amount of these lucky sequences is low.

This explains why the 'popular' folds of globular proteins are those which have practically no defects like crossed loops, cavities, etc.

3. The last case to be considered is the case when the groups differ in a number or variety of interactions (this can be caused, in particular, by different compactness of competing folds). As a result, the energy spectrum of one group is broader than that of another (Fig. 2d). Let both groups have the same number of folds ($M_1 = M_2$), the same mean energy ($\bar{E}^\circ_1 = \bar{E}^\circ_2$), but different energy dispersions ($\sigma_1 > \sigma_2$).

Then

$$\bar{m}'_{E^*}/\bar{m}''_{E^*} = \exp(-(E^* - \bar{E}^\circ)^2/2\sigma_1^2)/\exp(-(E^* - \bar{E}^\circ)^2/2\sigma_2^2)$$

is the ratio of average numbers of folds of these groups in a region related to the lowest energy fold ($\bar{m}'_{E^*} + \bar{m}''_{E^*} = 1$, as above). When $|\sigma_1 - \sigma_2| \ll \sigma$ (σ^2 is $(\sigma_1^2 + \sigma_2^2)/2$), this ratio has a simple form

$$\bar{m}'_{E^*}/\bar{m}''_{E^*} = \exp((\sigma_1^2 - \sigma_2^2)/2RT_*^2) \quad (7)$$

which shows a predominance of low-energy folds of that group which has a greater dispersion of interactions.

The same can be explained from another point of view. The expected energy of the lowest energy fold of Group 1 is $\bar{E}^\circ - \sigma_1\sqrt{2\ln M}$ (see eqn. 3), while the lowest energy fold of Group 2 usually has the energy of $\bar{E}^\circ - \sigma_1\sqrt{2\ln M}$. When $\sigma_1 > \sigma_2$, the best fold of Group 1 has usually a lower energy than the best fold of Group 2. Thus, a smaller variety of interactions (like a smaller

variety of folds, see above) gives a random sequence a smaller possibility of forming a low-energy fold. In general, this effect discriminates non-compact folds.

Summarizing the above equations, we see that abundance of the folding pattern 'p' is proportional to

$$\exp(-\tilde{F}_p/RT_*)$$

where

$$\tilde{F}_p = \bar{E}^\circ_p - \sigma_p^2/2RT_* - RT_* \ln M_p \quad (8)$$

is the 'selective free energy' of the pattern. It depends on M_p, a number of different folds within the pattern, on $\bar{E}^\circ_p$, the mean energy of the folds, on σ_p, the mean dispersion of their energies, and on the universal conformational temperature T_*.

Boltzmann-like statistics is a general feature of stable folds of random heteropolymers. The conformational temperature T_* emerges from a diversity of residues; the same temperature limits thermodynamic stability of the most stable folds of heteropolymer chains.

The basis of this statistic is that the more sequences provide stability of folds with a given feature, the more often this feature can be observed. As far as chains of globular proteins (unlike, e.g. fibrous ones) resemble random heteropolymers [24], this '*multitude principle*' is valid for them as well.

6. 'DEFECTS' OF FOLDING PATTERNS AND ACTIVE SITES OF PROTEIN GLOBULES

Equation 8 explains why coarse defects of protein architecture are rare. Small defects, however, are quite possible according to the same equation, especially because they can occur in various places of a globule, and one observes their Boltzmann-like distribution. As a result, all known structural rules [1–12] are statistical, they are met by the majority of proteins but not by all of them, and one does not see protein structures which are perfect in all aspects.

Moreover, there is no absolutely perfect folding pattern. Any one has an inherent, sequence-independent 'weak point' (crevice, funnel, etc.). Inspection of protein structures shows that these weak points are often occupied by active sites (Fig. 3).

At first it was noticed that substrates and cofactors bind to C-ends of parallel β-sheets of α/β proteins [32] and occupy the crevices formed by oppositely directed loops surrounding a 'switch point' of parallel β-sheets [6,33]. The position of this point is determined only by topology of the protein folding pattern.

Although a comprehensive review of protein active sites is still to be made, it is possible to present a lot of other examples of active sites occupying the inherent weak points of different folding patterns. Among them are various active sites in a 'funnel' of α/β barrel active sites of proteases and retinol-binding protein, in the 'splayed corner' [12] of orthogonal β-sandwiches, as well as haem-binding and active sites between diverged ends of helices of long α-helical bundles [34].

7. CONCLUDING REMARKS

One can imagine two basic ways in which new proteins could evolve. The first is repetition or fusion; it is clearly imprinted in amino acid sequences: repetition of small motifs in fibrous proteins, repetition or fusion of domains in multi-domain proteins, and of blocks in some membrane proteins. The second way is a choice from random sequences. This seems to be the case for globular proteins. It is imprinted in their quasi-random primary structures, in Boltzmann-like statistics of elements of their 3D structures, and, moreover, in their folding patterns which are just those that can be most readily formed by random sequences.

Acknowledgements: We are grateful to O.B. Ptitsyn, E.I. Shakhnovich, J. Janin and C. Chothia for valuable discussions, and to E.V. Serebrova for editing the text.

REFERENCES

[1] Rao, S.T. and Rossmann, M.G. (1973) J. Mol. Biol. 76, 241–256.
[2] Levitt, M. and Chothia, C. (1976) Nature 261, 552–557.
[3] Richardson, J.S. (1977) Nature 268, 495–500.
[4] Ptitsyn, O.B. and Finkelstein, A.V. (1980) Quart. Rev. Biophys. 13, 339–386.
[5] Richardson, J.S. (1981) Adv. Protein Chem. 34, 167–339.
[6] Branden, C. and Tooze, J. (1991) Introduction to Protein Structure, Garland, New York.
[7] Finkelstein, A.V. and Ptitsyn, O.B. (1987) Progr. Biophys. Mol. Biol. 50, 171–190.
[8] Chothia, C. and Finkelstein, A.V. (1990) Annu. Rev. Biochem. 59, 1007–1039.
[9] Chothia, C. (1973) J. Mol. Biol. 75, 295–302.
[10] Sternberg, H.J.E. and Thornton, J.M. (1976) J. Mol. Biol. 105, 367–382.
[11] Efimov, A.V. (1984) FEBS Lett. 166, 33–38.
[12] Chothia, C. (1984) Annu. Rev. Biochem. 53, 537–572.
[13] Efimov, A.V. (1991) FEBS Lett. 284, 288–292.
[14] Kajava, A.V. (1992) FEBS Lett. 302, 8–10.
[15] Novotny, J., Bruccoleri, R. and Karplus, M. (1984) J. Mol. Biol. 177, 787–818.
[16] Flory, P.J. (1969) Statistical Mechanics of Chain Molecules, Interscience, New York.
[17] Birstein, T.M. and Ptitsyn, O.B. (1966) Conformations of Macromolecules, Interscience, New York.
[18] Gutin, A.M., Badretdinov, A.Ya. and Finkelstein, A.V. (1992) Mol. Biol. (USSR) 26, 118–127.
[19] Bryngelson, J.B. and Wolynes, P.G. (1987) Proc. Natl. Acad. Sci. USA 84, 7524–7528.
[20] Garel, T. and Orland, H. (1988) Europhys. Lett. 6, 307–310.
[21] Shakhnovich, E.I. and Gutin, A.M. (1989) Biophys. Chem. 34, 187–199.
[22] Bryngelson, J.B. and Wolynes, P.G. (1990) Biopolymers 30, 177–188.
[23] Shakhnovich, E.I. and Gutin, A.M. (1990) Nature 346, 773–775.
[24] Ptitsyn, O.B. (1985) J. Mol. Struct. (Theochem.) 123, 45–65.
[25] Pohl, F.M. (1971) Nature New Biology 234, 277–279.
[26] Pohl, F.M. (1980) in: Protein Folding (Jaenicke, R., ed.) pp. 183–196, Elsevier, Amsterdam.

[27] Rashin, A.A., Ionif, M. and Honig, B. (1986) Biochemistry 25, 3619–3625.
[28] MacArthur, M.W. and Thornton, J.M. (1991) J. Mol. Biol., 218, 397–412.
[29] Miller, S., Janin, J., Lesk, A.M. and Chothia, C. (1987) J. Mol. Biol. 196, 641–656.
[30] Finkelstein, A.V., Ptitsyn, O.B. and Kozitsyn, S.A. (1977) Biopolymers 16, 497–524.
[31] Serrano, L., Sancho, J., Hirshberg, M. and Fersht, A.R. (1992) J. Mol. Biol. 227, 544–559.
[32] Hol, W.G.J., Van Duijnen, P.T. and Berendsen, H.J.C. (1978) Nature 273, 443–446.
[33] Branden, C. (1980) Quart. Rev. Biophys. 13, 317–338.
[34] Murzin, A.G. and Finkelstein, A.V. (1988) J. Mol. Biol. 204, 749–769.