Romanian Reports in Physics **72**, 604 (2020)

# BIOPHYSICS OF PROTEIN FOLDING. A SHORT REVIEW

CLAUDIA G. CHILOM, AUREL I. POPESCU*

Department of Electricity, Solid State and Biophysics, Faculty of Physics, University of Bucharest, 405 Atomistilor, Măgurele–Ilfov, Romania
* Corresponding author: *prof.aurel.popescu@gmail.com*

*Received March 16, 2020*

*Abstract.* A cell expresses over 10,000 proteins for metabolism driving, cellular morphology maintaining, and signaling towards other cells. All these proteins are properly operating only if they possess a native 3D structure. Protein folding is the dynamical process by which a protein attains its unique functional native 3D conformation. Two factors mainly contribute to protein folding: the physical and chemical properties of its amino acids and the influence of the crowded cellular microenvironment of the nascent protein. Although the number of the intermediate states, starting from linear unfolded nascent chain to the native state, is huge, the folding is very rapid implying thus energetically favoured folding paths. Protein folding not only generates biologically active structures, but also protects the protein from degradation by proteases and reduces the probability of abnormal aggregation as in the case of some diseases (*e.g.*, Alzheimer).

*Key words:* random coil protein, native protein state, folding mechanism, entropy variation, chaperones, protein misfolding

## 1. INTRODUCTION

Proteins are biological macromolecules essential to all cellular processes. Indeed, all metabolic biochemical reactions in the cell (*e.g.*, degradation and biosynthesis) are assisted by proteins. Proteins are also involved in the replication of the deoxyribonucleic acids and in their own biosynthesis. For all these reasons, proteins could be considered as the executive power of the cells.

## 2. PRIMARY STRUCTURE OF PROTEINS

The Human Genome comprises fewer than 21,000 protein-encoding genes [1], while the number of proteins for one organism is of the order of hundred thousand. Due to the fact that there are millions of animal and vegetal species on the Earth, each with its own set of proteins, it appears that the number of different proteins in Nature is huge. In spite of their great number and variety, all proteins are made up from a small number of elementary building blocks: 20 types of amino acids (AAs).

The protein structure is hierarchical, with four levels of structural organization. The primary structure is the lowest level of protein structure and consists of a shorter or longer chain of AAs bound together by strong peptide covalent bonds. Due to the fact that the peptide bonds are very strong, moderate thermal or chemical factors do not affect the primary structure of proteins.

A polypeptide chain presents a backbone formed by the repetition of the atoms, -$NC_{\alpha}C$-, starting from the amino group and ending at the carboxyl group, to which *n* side AA residues are attached. The number, *n*, of AAs in the natural proteins can vary from 50 to about 4,000 [2].

## 3. LOCAL STRUCTURE OF PROTEINS

Even in the process of their synthesis, protein chains do not remain linear due to a multitude of physical interactions (*e.g.*, ionic interactions, intra-chain hydrogen bond formation, van der Waals interactions). Thus, the protein thread will adopt different secondary structures elements. This process is known as local folding and can be easily highlighted by physico-chemical methods, such as, for example, the spectrofluorimetric ones [3].

There are mainly four types of regular secondary structures: α-helix; β-strand (occasionally generating β-sheets); β-turn and random coil structure.

In the case of α-helix, the atoms of the protein backbone (*i.e.*, N, $C^{\alpha}$, and C) are situated on an ideal helix, with a pitch of 5.4 Å, a diameter of about 12 Å and with 3.6 amino acids per pitch. The α-helix structure is stabilized by intrachain hydrogen bonds [4].

The β-strand is more extended than the α-helix, the distance between the two $C^{\alpha}$ being of the order of 3.5 Å. The β-strands of the protein chains are usually aggregated, by many inter-strand hydrogen bridges, in to two or more strands, forming β-pleated sheets. There are two classes of β-sheets: parallel β-sheets ($\beta_p$) and antiparallel β-sheets ($\beta_a$). From the energetic point of view, the $\beta_a$-sheets are more stable than the $\beta_p$-sheets. The β-sheets structures are stabilized by hydrogen bonding between polypeptide strands [4].

The β-turn structures consist of a small number of AAs and look like kinks of the protein thread, being also known as hairpin turns which connect α-helices and/or β-strands.

One finds also, loops of the protein chain not organized as a secondary structure, called random coils. These structures confer flexibility to the protein chain, unlike α-helices and β-strands that provide rigidity to the chain.

## 4. LONG-RANGE PROTEIN STRUCTURE (GLOBAL ASSEMBLY)

Due to interactions between AAs of the protein chains through disulfide bridges, as well as, due to chain interactions with water molecules, the elements of

secondary structure fold into a 3D compact structure, attaining a global minimum in the protein energetic landscape. However, in this structure, it is possible that some elements of the secondary structure, situated close to the amino end, to reach into close vicinity of the elements, near the carboxyl end of the protein chain, justifying the qualification of this type of folding as a long-range folding.

In general, protein 3D structures are very compact, many proteins adopting a more or less globular form, with the hydrophobic AAs buried inside and the hydrophilic AAs exposed to the aqueous exterior, that is, to the "surface" of the globule. This level of 3D structure of a protein is known as the tertiary structure in which the protein is fully functional. For this reason, this state is called the native state of the protein in which is very sensitive to denaturant agents (*e.g.*, radiation, heat, pH and urea); an example is that of the serum albumin proteins, whose 3D structure is greatly affected by physico-chemical disturbing factors [5, 6].

## 5. LONG MULTIMERIC ORGANIZATION PROTEINS

A lot of proteins attain only the tertiary structure (*e.g.*, myoglobin, ribonuclease) but, there are many others that are functional only if they are associated with other molecules of the same or different kind, already folded in tertiary structures (monomers), to form a functional multimer known as a quaternary structure.

The association of protein monomers in multimers is usually the result of weak but numerous interactions (*e.g.*, hydrogen bonds, hydrophobic interactions) and also of stronger ionic interactions. The quaternary structure is also very sensitive to the action of different denaturing agents. There are three classes of 3D protein structures: globular (*e.g.*, hemoglobin), fibrous (*e.g.*, collagen, etc.) and membrane proteins (*e.g.*, ionic channels).

Protein folding may sometimes go beyond the quaternary level: some proteins, in order to become fully functional, form supramolecular assemblies by associated with lipids in membrane, with nucleic acids in chromosomes and ribosomes, etc.

## 6. PROTEIN FOLDING

Protein folding is defined as the process through which the polypeptide chain is acquiring its final 3D native structure during or after its translation by ribosomes. Levinthal "paradox", states that the huge age of the Universe is not sufficient for an average sized protein to sample all space conformations until it finds its native state. This is in strong contrast with the experimental evidence. Therefore, the process of folding still seems to remain largely unsolved. However, because proteins fold on a timescale of μs to s, this process must be somehow biased and not at random walk [7].

The protein folding can be viewed as a series of steps, each following a complex pathway on energy landscape called a funneled energy landscape. During their search for the native conformation, the proteins will not sample all the possible

states, but are rather "guided" towards an energy global minimum shortcutting many possible intermediate states [8].

The problem of protein folding can be simply formulated as follows: how does the particular AA sequence specify the native state of a protein? According to Anfinsen dogma, the protein folding is mainly determined by its AA sequences [9]. This is true, especially, for short protein chain.

Another question is how is it possible for proteins to fold, within μs, into their native states, given the large number of misfolded but potentially states a protein could attain: about $10^{95}$ for 100 residue protein chain [10].

## 7. THE ROLE OF ENTROPY IN PROTEIN FOLDING

As concerns the biochemical reactions, one can say that the catabolic reactions are accompanied by a generation of disorder at cellular level, described by an increase of entropy ($\Delta S > 0$). These degradation processes are therefore, entropic. On the contrary, the anabolic reactions are inducing order into the cell, thus being considered as antientropic processes ($\Delta S < 0$). The second principle of thermodynamics refers to the direction of system evolution. According to this principle, during an irreversible spontaneous evolution, the variations of Gibbs free energy of a system, must decrease ($\Delta G < 0$).

In order to interpret protein folding, from thermodynamics point of view, we shall use, in this section, the expression of Gibbs free energy variation, at constant temperature, given by the common expression [11]:

$$\Delta G = \Delta H - T\Delta S, \tag{1}$$

where $\Delta H$ is the variation of enthalpy.

The above relation emphasises the combined effect on Gibbs free energy variation, both of entalpy and entropy variation, as we shall see later.

The nascent protein (*i.e.*, leaving the ribosomes) can reach a huge number of 3D conformations which, however, could not be biochemically functional. This number of nonfunctional random coil stereoisomers ($W_{rc}$) can be seen as initial precursors of the unique native state endowed with a specific biochemical activity. From thermodynamics point of view, to this number of random coil stereoisomeric precursors, one can associate the entropy, $S_{rc}$, given by Boltzmann's formula:

$$S_{rc} = k_B \ln W_{rc}, \tag{2}$$

where $k_B$ is the Boltzmann's constant ($k_B = 1.38 \times 10^{-23}$ J/K) and $W$ represents the thermodynamics probability [*i.e.*, the number of system (here, the protein) microstates compatible with a given macrostate].

An average protein has more than 100 AAs and, therefore, the possible number of equiprobable states is much higher. A protein chain will have the highest

energy and approximately zero entropy, in the state where it is fully stretched. On the contrary, it possesses a lower energy as it reaches smaller extensions, as the entropy increases (Fig. 1). Hence, based on second thermodynamics principle, one can conclude that a protein linear chain has the natural tendency firstly to collapse and maximize its entropy, while minimizing its free energy when many possible conformations are possible (position 2 in Fig. 1) for a collapsed molecule than for an extended one (position 1 in Fig. 1). This first step of protein folding (1 → 2) is purely driven by entropy, because a higher number of possible states corresponds to a lower free energy.

As the size of the peptide bond is independent of the nature of the AA it is important to note that the entropic collapse (2 → 3 → 4), when the protein evolves towards its native state, passing through the molten globule states, depends on how many AAs are forming the protein structure.

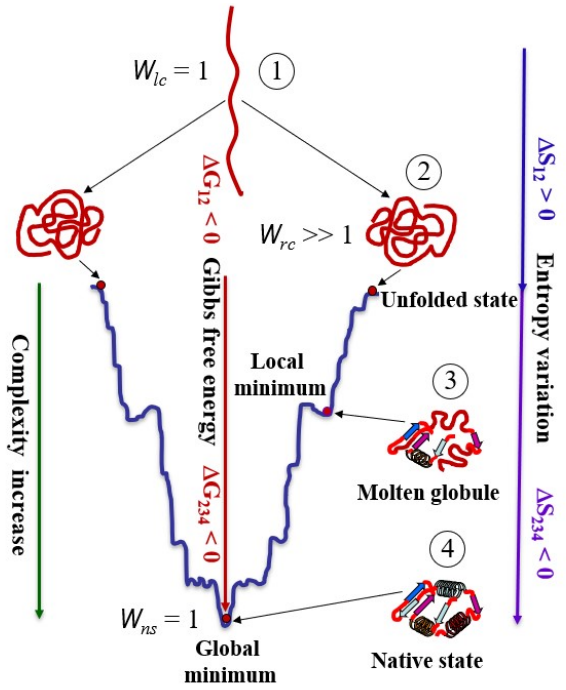


Fig. 1 – Energetic funnel-like landscape of protein folding. Proteins are starting from a stretched linear chain (1) potentially, collapsed in multiple unfolded states (2), then are passing through a smaller number of partially folded states (molten globule) in a local, energetic minimum (3) and, eventually, are reaching the native states (4), a global minimum, in which they are attaining the global Gibbs free energy minimum [modified after [12]: A. I. Popescu, *Biophysics. Current Status and Future Trends* (2016)].

Usually, the native state (*ns*) of a protein has a unique 3D conformation (*i.e.*, $W_{ns} = 1$) so that, its associated entropy is, according to (2) equal to zero. Therefore, the process of protein folding, starting from the nonfuctional random coil stereoisomers

towards the highly ordered protein state, is accompanied by an entropy collapse, $\Delta S_{rc\to ns}$:

$$S_{rc\to ns} = k_B \ln W_{rc} < 0 \tag{3}$$

in contrast with the requirement of the second principle of thermodynamics.

Indeed, taking into account the requirement of the second principle of thermodynamics, it results:

$$\Delta G_{rc\to ns} = \Delta H_{rc\to ns} - T\Delta S_{rc\to ns} < 0. \tag{4}$$

Because $\Delta S_{rc\to ns} < 0$, the second term of (4) is positive. This could mean that only high exotermic folding ($\Delta H_{rc\to ns} < 0$) would have a chance to occur, when $|\Delta H_{rc\to ns}| > T\Delta S_{rc\to ns}$. But this is not the case.

In order to solve this contradiction, it is naturally to postulate the existence, at least, of a competitive process generator of entropy, $\Delta S_{gen} > 0$, taking place concomitently with protein folding, so that to surpass the antientropic folding process:

$$\Delta S_{gen} > |\Delta S_{rc\to ns}|. \tag{5}$$

Therefore, only in this case, the total variation of entropy, $\Delta S_{tot}$, would be positive, being the algebraic sum of the two processes:

$$\Delta S_{tot} = \Delta S_{rc\to ns} + \Delta S_{gen} > 0. \tag{6}$$

Therefore the condition (4) is replaced by the relation (7) which holds, irrespective of the value of $\Delta H_{rc\to ns}$:

$$\Delta G_{tot} = \Delta H_{rc\to ns} - T\Delta S_{tot} < 0. \tag{7}$$

In the next section, we shall mention two processes which could generates an excess of entropy, $\Delta S_{gen} > 0$, so that to drive the entire process of protein folding.

Supposing that, initially, after protein translation by ribosomes, the protein linear chain (*lc*) structure would be unique, $W_{lc} = 1$, it results that its entropy is zero, according to (2). But this is true only theoretically, because, just during the protein translation, the nascent chain has the tendency to fold, attaining potentially a huge number, $W_{rc}$, of nonfunctional random coil stereoisomers. Therefore, these partial transitions of the linear chain is accompanied by a positive variation of entropy, $\Delta S_{lc\to rc} = k_B \ln W_{rc} > 0$ and a decrease of Gibbs free energy, $\Delta G_{lc\to nr} < 0$. Any of the possible random coil chain, attained by the protein molecule, has the tendency to accomplish its natural folding reducing more and more its free energy until to a minimum, reaching the functional native protein state (Fig. 1) characterized by a unique configuration, $W_{ns} = 1$, with zero entropy. Therefore, this last part of

folding is, as we already stated, accompanied by an entropy collapse, $\Delta S_{rc\to ns} = -k_B \ln W_{rc} < 0$ (Fig. 1). Finally, after a few ms, the protein acquires its native structure (with almost zero entropy).

In fact, in the native state, the entropy of the protein is not zero, because here, the spatial structure of protein is not as rigid as that of a crystal near the 0 K. On the contary, at room temperature, the component atoms of the protein are oscillating in different modes. The intricate collective and variable superimposition of their movements have as result a kind of global ”breathing” of the entire molecule, which have the tendency to attain the true energy global minimum, but never reaching it. This multitude of „like-native states” of protein, near the energy global minimum, are all compatible with the adequate biological function of protein, because on the one hand, these compact states are highly specific towards their ligands and, on the other hand, they are sufficiently flexible to dynamically adapt their conformations to that of ligands during their interactions, according to the induced fit model of interaction.

The folding process is involving multiple parallel (hundreds or thousands) noncovalent (van der Waals) interactions, which have the main contribution to the folding, followed by hydrogen bonds and, in the third place, by hydrophobic interactions [13]. This explains why protein folding is progressing very fast (in some cases, in μs) being a biased process within funnel-like energy landscape [14–16]. The rapid folding is explaining, for example, why the bacteria are able to spend their whole short lifespan and then, to divide, in about half an hour.

It is interesting to mention that the net free Gibbs energy, $\Delta G$, of folded protein stabilization, is about (21–42) kJ/mol, that is, about only ten times the thermal energy ($k_BT$) at physiological temperature. Therefore, the most proteins, in their native states, are marginally stable because the Gibbs free energy released during the folding is counterbalanced by the loss of entropy (*i.e.*, loss of conformational flexibility). One can say that a folded protein is a thermodynamic compromise [17].

In the case of large protein chains, due to the crowded cellular medium populated with similar or different biomacromolecules, the folding is assisted by a special kind of proteins, known as chaperones, which protect the protein from the misfolding of the structure. For instance, many chaperones are pertaining to the class of heat shock proteins which are expressed, after cell exposure to high temperatures. Other chaperones are involved in assisting the correct folding of nascent large proteins during their extrusion from the ribosomes.

## 8. THE ROLE OF WATER IN PROTEIN FOLDING

It is known that the protein chains have a variable sequence of AAs endowed with different properties: some AA residues are acidic, other are basic, polar and hydrophylic, while others are apolar and hydrophobic. Therefore, the interactions of proteins with their aqueous microenvironment is very complex. One can say that the polar hydrophylic residues are hydrated, at least, in the initial phase of nascent protein. In the process of folding, some of the hydrophylic AAs are constrained to

be embedded into the core of native state, losing their water molecules which are expelled into the bulk water whose structure is perturbed (Fig. 2). On the contrary, other AAs (*e.g.*, Trp) due to their hydrophobic character, are naturally present deep inside the protein structure. Therefore, it is expected that the majority of hydrophylic AAs will be distributed on the surface of the native state, while the majority of the hydrophobic AA residues will be burried into the protein core.

For this reason, many researchers agree with the idea that the source of excess entropy generation during folding, is due the disorganization of the ordered water molecules surrounding the protein molecule (the so named bound water). These ordered water molecules are disorganized during the protein folding and expelled to the bulk water so that, the associated entropy to water rearrangement accompanying the folding, $\Delta S_{aq} = \Delta S_{gen}$, is positive [18] and, moreover, greater than conformation entropy collapse, according to the equation (6).

If one consider the extended thermodynamic system (protein and aqueous surrounding), the protein folding (an antientropic process), accompanied by the disorganization of the water molecules from its surrounding (an entropic process), there is no contradiction of protein folding with the second principle of thermodynamics (Fig. 2).

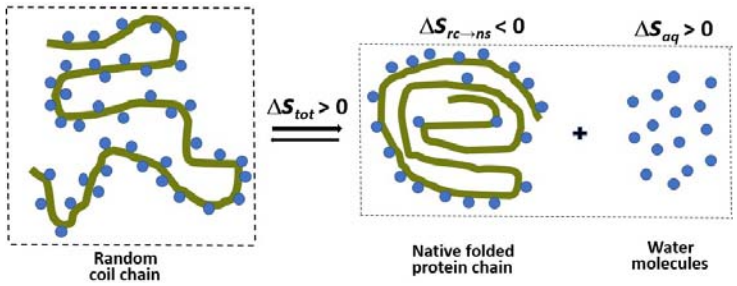


Fig. 2 – The schematic illustration of entropy evolution during the ptotein folding of the complex (protein + hydration water). $\Delta S_{tot}$ = the global variation of entropy; $\Delta S_{rc \to ns}$ = the entropy collapse during transition from the random coil stereoisomeric precursors towards the native state; $\Delta S_{aq}$ = the variation of entropy of water molecules surrounding the protein chain [19].

On the other hand, there are opinions that the excess production of entropy in folding, is not the result of disorganisation of water shell around the protein chain. Indeed, one can observe that, in general, both the core and the surface of the folded proteins present a somehow mosaic like structure of hydrophylic and hydrophobic AAs. Consequently, the bound water molecules, initially around the protein chain, are not clearly distributed in majority on the exterior of the protein after the folding [20]. These authors consider that the entropy excess, $\Delta S_{aq}$, is due to the translational entropy of water (*i.e.*, due to the translational movement of water molecules). If the number of AA residues is sufficiently large, the translational entropy generation can compete with the entropy collapse during folding [20].

### 9. THE PROTEIN MISFOLDING

The protein misfolding is favoured both by the closeness of the partially folded nascent proteins and by the high degree of macromolecular crowding in the cells.

Several diseases (*e.g.*, Alzheimer, Parkinson) are triggered by protein misfolding, where a mutation or an external process forces a given class of proteins towards a wrong conformation state. Protein misfolding and aggregation are considered to be the cause of death for the cells hosting these proteins. Thus, some misfolded membrane-bound proteins (*e.g.*, misfolded prions) are involved in more than 20 types of diseases [21]. The misfolded prion proteins, trigger the properly folded host proteins to convert them into the misfolded forms: the misfolded prions seem to act as templates guiding the proteins to become misfolded. Each misfolded prion triggers a polymerization chain reaction inducing β-sheet misfolded proteins [22] resulting thus, tightly packed structures named amyloid fibrils. These are very stable assemblies and, being accumulated in the infected cells, cause the cell death and, consequently, the injury especially of nervous tissues [23].

The treatments of these diseases should, therefore, prevent the formation of β-sheet amyloid fibrils, that is, the transition from partially folded precursors to the stable aggregated states (the transition 3 in Fig. 3), because the conversion of the already formed amyloid fibrils to the native states (*via* pathway 4) is quite improbable.

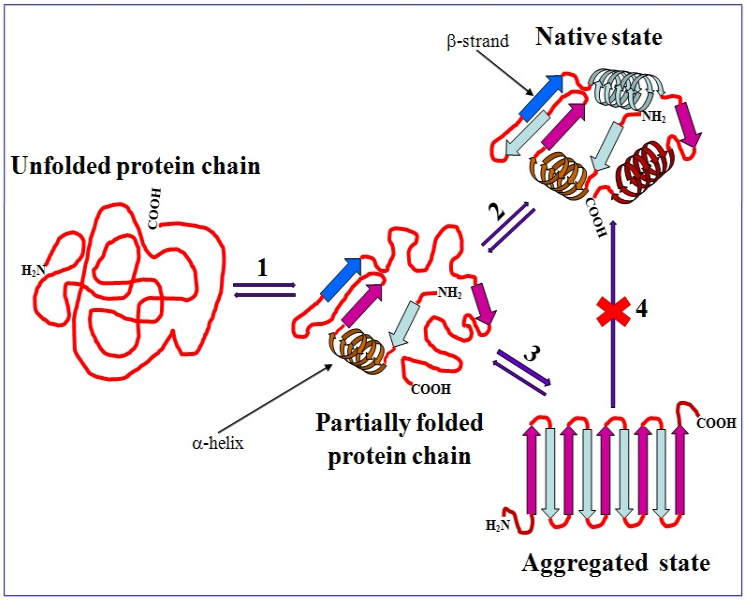


Fig. 3 – A scheme of the equilibria between different conformational states of a protein in a cell. To be noted that the transition from the aggregated state (amyloid) to the native state it is not spontaneous ([12]: A. I. Popescu, *Biophysics. Current Status and Future Trends*, 2016).

## 10. CONCLUSIONS

The protein folding is a *sine qua non* process for attaining the native states endowed with correct and efficient cellular functions.

The protein folding is a biased process shortcutting many possible intermediate states that could slowdown it, dictated by requirement of the second principle of thermodynamics.

A decisive role in protein folding is played by the evolution of Gibbs free energy and entropy variation, as well. While the variation of Gibbs free energy is steadily decreasing, the evolution of entropy is characterized by an initial increase ($\Delta S_{12} > 0$, from the linear nascent protein chain, at ribosomes, to a random coil structure), and a final decrease ($\Delta S_{234} < 0$, from random coils chain, to native state).

The total entropy variation of the global system, (protein and aqueous microenvironment) is positive due to the involvement of water molecule ($\Delta S_{aq} > 0$).

In the case of different neuronal diseases, the process of correct protein folding is perturbed. The understanding of this wrong folding would help to devise means to prevent it and, therefore, to cure these diseases.

Proteins are the results of a competitive natural molecular selection among the multitude of possible variants, during living matter evolution, the actual proteins being successfully adapted to find the fastest way towards their functional native states.

REFERENCES

1. E. Pennisi, Science **337**, 1159–1161 (2012).
2. H. Lodish, A. Berk, P. Matsudaira, C. K. Kaiser, M. Krieger, M. P. Scott, S. L. Zipursky and J. Darnell, *Molecular Cell Biology*, 5th edition, W. H. Freeman Company, New York, 2004
3. C. Chilom, G. Barangă, D. Găzdaru, A. Popescu, Journal of Optoelectronics and Advanced Materials **15** (3–4), 311–316 (2013).
4. J. M. Berg, J. L Tymoczko, L. Stryer, *Biochemistry*, New York, W H Freeman, 2002.
5. C. Chilom, G. Barangă, D. Găzdaru, A. Popescu, Journal of Optoelectronics and Advanced Materials **13** (5), 583–587 (2011).
6. N. Sandu, C. G. Chilom, A. I. Popescu, Romanian Journal of Physics **65** (3–4), 703 (2020).
7. C. Levinthal, Are there pathways for protein folding? J. Chim. Phys. **85**, 44–45 (1968).
8. V. Raicu and A. Popescu, *Integrated Molecular and Cellular Biophysics*, Edition Springer, 2008
9. C. B. Anfinsen, Science **181**, 223–230 (1973).
10. G. D. Rose, P. J. Fleming, J. R. Banavar, A. Maritan, Proc. Natl. Acad. Sci. USA **103**, 16623–16633 (2006).
11. B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, *Molecular Biology of the Cell*, 4th edition, Garland Science, Taylor & Francis Group, New York, 2002.
12. A. I. Popescu, *Biophysics. Current Status and Future Trends*, The Publishing House of the Romanian Academy, 2016.
13. B. Nölting, Protein Folding Kinetics. Biophysical Methods, Second Edition, Springer (2006).
14. K. Dill, H. S. Chan, Nat. Struct. Biol. **4**, 10–19 (1997).
15. T. E. Fisher, A. F. Oberhauser, M. Carrion-Vazquez, P. E. Marszalek, J. M. Fernandez, Trends Biochem. Sci. **24**(10), 379–84 (1999).

16. M. M. Lin and A. H. Zewail, Ann. Phys. (Berlin) **524** (8), 379–391 (2012).
17. G. A. Petsko and D. Ringe, *Protein Structure and Function*, New Science Press Ltd., London, 2004.
18. https://www.ius.edu.ba/sites/default/files/u796/protein_folding_notes.pdf
19. C. Chilom, M. Bărbînţă-Pătraşcu, D. Găzdaru, A. Popescu, Biophysics of Protein Folding. Mechanism and Importance, *Annual Scientific Conference*, June 21–22, Faculty of Physics, University of Bucharest, 2019.
20. Y. Harano and M. Kinoshita, Biophys. J. **89**, 2701–2710 (2005).
21. I. S. Krull and B. K. Nunnally, *Prions and mad cow disease*, New York, N.Y., Marcel Dekker, 2004.
22. A. Aguzzi, Proc. Natl. Acad. Sci. USA **105**, 11–12 (2008).
23. J. Laurén, D. A. Gimbel, H. B. Nygaard, J. W. Gilbert, S. M. Strittmatter, Nature **457**, 1128–32 (2009).