Biophysical Reviews (2022) 14:1255–1272
https://doi.org/10.1007/s12551-022-01000-1

REVIEW

# Protein folding problem: enigma, paradox, solution

Alexei V. Finkelstein[1,2,3] · Natalya S. Bogatyreva[1] · Dmitry N. Ivankov[4] · Sergiy O. Garbuzynskiy[1]

Received: 22 July 2022 / Accepted: 19 September 2022 / Published online: 11 October 2022

## Abstract

The ability of protein chains to spontaneously form their three-dimensional structures is a long-standing mystery in molecular biology. The most conceptual aspect of this mystery is how the protein chain can find its native, "working" spatial structure (which, for not too big protein chains, corresponds to the global free energy minimum) in a biologically reasonable time, without exhaustive enumeration of all possible conformations, which would take billions of years. This is the so-called "Levinthal's paradox." In this review, we discuss the key ideas and discoveries leading to the current understanding of protein folding kinetics, including folding landscapes and funnels, free energy barriers at the folding/unfolding pathways, and the solution of Levinthal's paradox. A special role here is played by the "all-or-none" phase transition occurring at protein folding and unfolding and by the point of thermodynamic (and kinetic) equilibrium between the "native" and the "unfolded" phases of the protein chain (where the theory obtains the simplest form). The modern theory provides an understanding of key features of protein folding and, in good agreement with experiments, it (i) outlines the chain length-dependent range of protein folding times, (ii) predicts the observed maximal size of "foldable" proteins and domains. Besides, it predicts the maximal size of proteins and domains that fold under solely thermodynamic (rather than kinetic) control. Complementarily, a theoretical analysis of the number of possible protein folding patterns, performed at the level of formation and assembly of secondary structures, correctly outlines the upper limit of protein folding times.

**Keywords** Protein 3D structure · Protein folding · Levinthal's paradox · "All-or-none" phase transition · Free energy landscape · Folding funnel

## Introduction

Well-defined spatial structure is necessary for functioning of majority of proteins. Protein folding is a process that converts the disordered protein chain into a chain having a definite, unique three-dimensional (3D) structure. However, nowadays the term "protein folding problem" has two meanings: one emphasizing the process, the other the result. The former (sometimes called "the protein folding problem of the first order") implies the answer to the question of *how can* the protein chain choose, in minutes, its unique structure among a giant number of others; the latter (sometimes called "the protein folding problem of the second order") implies the answer to the question of *what* structure will be attained by the protein chain of a certain amino acid sequence.

For a long time, these two problems were considered as one, assuming that once "how" was solved, "what" would be solved right away.

However, now it is clear that these are two different problems because they have been solved by two quite different methods.

The problem of "what" has been very recently solved by bioinformatics with the aid of neural networks (Fariselli et al. 2001) and artificial intelligence (Senior et al. 2019, 2020; Yang et al. 2020; Jumper et al. 2021); see some discussion of these works in Roney and Ovchinnikov (2022); and for a review of early works on protein structure prediction, see Finkelstein and Ptitsyn (2016), lectures 22, 23.

✉ Alexei V. Finkelstein
afinkel@vega.protres.ru

[1] Institute of Protein Research of the Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia

[2] Biotechnology Department of the Lomonosov Moscow State University, 4 Institutskaya Str, 142290 Pushchino, Moscow Region, Russia

[3] Biology Department of the Lomonosov Moscow State University, 1-12 Leninskie Gory, 119991 Moscow, Russia

[4] Center of Life Sciences, Skolkovo Institute of Science and Technology, 121205 Moscow, Russia

The topic of protein structure prediction (or recognition) from its amino acid sequence needs to be described and considered not here but separately. However, here it is appropriate to say that a pronounced success of the best of the latest protein structure prediction programs, AlphaFold2 (Jumper et al. 2021), is based on enormous databases of protein structures (Berman et al. 2003) and sequences (The UniProt Consortium 2021) collected during many decades.

The problem of "how can" the protein chain choose, in minutes, its unique spatial structure among a giant number of others has been solved by physics. The aim of this article is to outline the principal moments of the solution.

The ability of proteins to fold spontaneously puzzled protein science for a long time (see, e.g., (Anfinsen and Scheraga 1975; Jackson 1998; Fersht 2000; Grantcharova et al. 2001; Robson and Vaithilingam 2008; Dill and MacCallum 2012; Wang et al. 2012; Wolynes 2015; Finkelstein and Ptitsyn 2016; Finkelstein 2018)).

As known, in living cells, gene-encoded protein chains are synthesized by special molecular machines, called ribosomes. Most of the protein chains, though not all of them (see (Uversky and Finkelstein 2019) and references therein) have to obtain their unique ("native," working) three-dimensional structures to perform their unique biological functions.

This phenomenon is called "protein folding."

Its importance for protein functioning was recognized in the 1950s (Anfinsen 1959), followed by the finding that protein folding can occur not only in vivo but also in vitro (Anfinsen et al. 1961). Although those early in vitro studies that have shown that proteins can reversibly refold from a denatured, disordered state were focused mainly on small proteins, recent experiments using mass spectrometry-based proteomics (To et al. 2021) have demonstrated that nearly two-thirds of soluble bacterial proteins are refoldable in vitro under physiological conditions. Still, many proteins, especially those with large complicated multi-domain structures, aggregation-prone and non-soluble proteins are not refoldable in vitro under physiological conditions (To et al. 2021); see also Sorokina et al. (2022); besides, some proteins (we will not consider them here) are "intrinsically disordered" (Wright and Dyson 1999; Uversky 2002; Tompa 2005; Uversky and Finkelstein 2019) — they start their work, not yet having a well-folded structure, that they cannot acquire per se either in vivo or (under physiological conditions) in vitro, but usually become well-folded when interacting with other molecules.

Therefore, this review is limited to the folding of single-domain proteins and separate protein domains made of single chains; and here, we virtually do not consider folding of multi-domain proteins (facilitated, of course, by folding of separate domains) and complications of folding associated with interactions of a protein with other proteins (including chaperones), protein aggregation, amyloid formation, etc.

## Experimental studies of protein folding

Since it is rather difficult to trace a change in the structure of a nascent protein chain against the background of many other molecules in a living cell, the investigation of protein folding started with in vitro experiments on the folding of water-soluble molecules of globular proteins, see Finkelstein and Ptitsyn (2016), lectures 19, 20.

However, it makes sense to begin this paper with a short overview of comparatively recent results on the folding that occurs in the course of protein biosynthesis on ribosomes.

The first studies were carried out on large (mostly multi-domain) proteins. They showed that these start to fold before their biosynthesis has been completed: the first synthetized (N-terminal) immunoglobulin domain folds when the whole chain has not been synthesized yet (Isenman et al. 1979); the luciferase protein starts to work immediately upon completion of the chain biosynthesis, so that it has no time to fold after the biosynthesis and should fold cotranslationally (Kolb et al. 1994); and the globin chain can bind to heme when a bit more than a half of the chain has been synthesized by the ribosome (Komar et al. 1997), though it is hard to say whether structuring of this half-made chain occurred before the heme-binding or resulted from it. Anyway, these data suggest that the protein chain folding in vivo starts already on the ribosome ("cotranslationally") and that this cotranslational process may differ from the in vitro folding ("renaturation") of entire protein chains discussed below.

More up-to-date experiments on cotranslational structure acquisition by small, of ≈70 residues, nascent proteins (monitored by $^{15}$ N, $^{13}$C NMR, and FRET) showed that "polypeptides [at a ribosome] remain unstructured during elongation but fold into a compact, native-like structure when the entire sequence is available" (Eichmann et al. 2010); "… folding [occurs] immediately after the emergence of the full domain sequence" (Han et al. 2012); "… cotranslational folding … proceeds through a compact, non-native conformation [i.e., something molten globule-like] … [and] rearranges into a native-like structure immediately after the full domain sequence has emerged from the ribosome" (Holtkamp et al. 2015); thus, the latter case shows that a protein can fold cotranslationally outside the ribosome exit tunnel (and then it meets nearly the same problems as a protein renaturating in vitro.

Further experiments using optical tweezers, single-molecule real-time FRET, cryo-EM, and pulling force-profile analysis allowed a more detailed study of cotranslational folding. It has been shown that a small (of ≈30 residues) protein, zinc-finger domain, can fold deep inside the vestibule of the ribosome exit tunnel (Nilsson et al. 2015; Wruck et al. 2021), and that α -helices, these "one-dimensional" details of the protein structure, can fold sequentially inside and at the vestibule of the ribosomal tunnel. The observed folded or partially folded structures of a nascent α -helical domain of spectrin show that it may fold there via a pathway different from that of the isolated domain (Nilsson et al. 2017), but with the same result. On the other hand, the principal features of the folding pathway of a larger (of ≈100 residues) β-structural Ig domain has been found to remain conserved on and off the ribosome (Tian et al. 2018), while folding of another protein, having a β-barrel shape, demonstrates a switch from the initial dynamic α-helical to β-strand conformation during the co-translational folding (Agirrezabala et al. 2022).

Thus, as shown, there may be no fundamental difference between the in vivo (on the ribosome) and in vitro (out of the ribosome) folding, at least for small proteins, though some details of the on-ribosome and in vitro folding pathways can differ. In both cases, native structures, at least for small proteins, emerge only when the entire sequence of a stable protein domain has been synthesized (in this connection, it should be noted that slightly truncated protein chains lose stability of their native folds, do not refold, and remain compact but disordered in vitro (Flanagan et al. 1992)).

The discovery of chaperones, the cell's troubleshooters (Ellis and Hartl 1999), re-aroused suggestions that the protein folding processes in vivo and in vitro may be quite different because chaperones may have a foldase/unfoldase activity (see, e.g., Libich et al. (2015) and references therein). However, the analysis of data presented in Libich et al. (2015) reveals that the most studied chaperone (GroEL) does not speed up the overall folding process (Marchenko et al. 2015): GroEL accelerates transitions between the unfolded and folded GroEL-bound states of the target protein (Libich et al. 2015; Thirumalai et al. 2020), but not its overall folding. Moreover, when the concentration of the target protein is low so that it does not aggregate, a redundant concentration of GroEL slows down the folding of this protein (Marchenkov et al. 2004). This corroborates the conclusion that GroEL serves as an auxiliary transient trap that simply binds the excess of unfolded protein chains, thus preventing them from irreversible aggregation (Marchenkov et al. 2004; Marchenko et al. 2009).

One can conclude that the self-organization of structures of separate proteins (which in the case of in vitro folding of water-soluble globular proteins unassisted by other biomolecules) captures the main peculiarities of the protein folding phenomenon. This means that all the information necessary to build up the 3D structure of a protein is inscribed in its amino acid sequence (this was Anfinsen's "thermodynamic hypothesis").

Thus, the studies of self-organization have shown that an unfolded protein chain can spontaneously, "by itself," fold into its unique native 3D structure (Anfinsen et al. 1961; Anfinsen 1973). In Anfinsen's experiment, the enzyme ribonuclease A stayed unfolded in the presence of urea and a thiol reagent, and with these agents removed, it spontaneously refolded, recovering its structure (as shown by correct restoration of all four S–S bonds) and function. However, as it has been recently found by David Eisenberg (2018), "essentially the same experiment had been performed earlier by a medical student [Lisa A. Steiner, later MIT professor] at Yale, but neither [she nor] her research supervisor nor her department chair thought it particularly significant, and her work was not published." "Why did this transformative result lay hidden in her thesis?" asked Eisenberg, and answered: "She had the answer to a hugely important question, but that question had not yet been posed" because then (in the mid-1950s) it had not yet been elucidated "how biological information passes from the genome to proteins"…

## The protein folding problem

In the course of self-organization, the protein chain has to find its native (and seemingly, according to Anfinsen's "thermodynamic hypothesis," the most stable) fold among zillions of other alternatives (Fig. 1) within only minutes or seconds given by a cell life for its folding.

The number of alternatives is vast indeed (Levinthal 1968, 1969): it is at least $2^{100}$ but more likely $3^{100}$ or even $10^{100}$ (or $100^{100}$) for a 100-residue chain, because at least 2 ("right" and "wrong") but more likely 3 (α, β, "coil") or ≈10 (Privalov's (1979) experimental estimate), or even 100

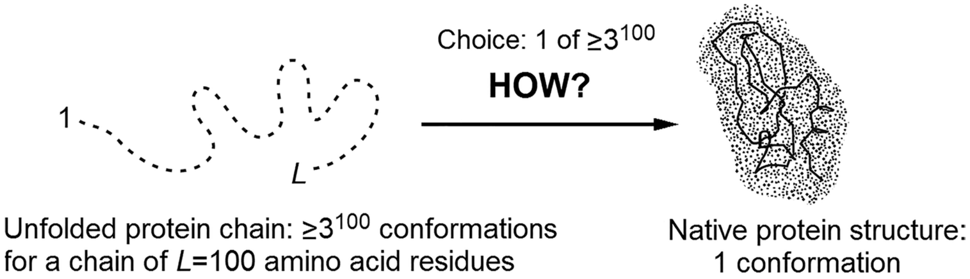
Choice: 1 of ≥3^100
HOW?
1
L
Unfolded protein chain: ≥3^100 conformations for a chain of L=100 amino acid residues
Native protein structure: 1 conformation

**Fig. 1** The Levinthal's choice problem. The choice of the native structure can be determined either by the somehow restricted folding process (Levinthal's "kinetic hypothesis") or by the enhanced native fold stability (Anfinsen's "thermodynamic hypothesis")

(Levinthal 1969) conformations are possible for each amino acid residue.

Since the chain cannot pass from one conformation to another faster than within a picosecond (the time of a thermal vibration), the exhaustive search would take at least ~ $2^{100}$ ps (but more likely $3^{100}$, or even $10^{100}$, or $100^{100}$), that is, ~ $10^{10}$ (or $10^{25}$, or even $10^{80}$, or $10^{180}$) years (Levinthal 1969). And it looks like the sampling should be exhaustive because the protein "feels" that it has attained the stable structure only when hitting it precisely, since even a 1 Å deviation can strongly increase the chain energy in the closely packed globule.

The main protein folding puzzle is why the native protein structure is found within minutes rather than within “"Levinthal's" ~ $10^{10}$ or more years (that is, within ~ $10^{18}$ or more minutes)! This reduction of the folding process by 1 000 000 000 000 000 000 (!) times (compared to iterating over all structures) must be always kept in mind, without distracting to dead-end considerations that promise, say, 1000- or even 1 000 000-fold acceleration of the process.

How can the protein chain choose, in minutes, its native structure among a giant number of others, asked Levinthal (1968; 1969) who first noticed this paradox, and answered: It seems that the protein folding follows some specifically restricted fast pathway, and the native fold is simply the end of this pathway, no matter if it is the most stable chain fold or not (this was Levinthal's "kinetic hypothesis"). In other words, Levinthal suggested that the native protein structure is determined by kinetics rather than stability and corresponds to the easily accessible local free energy minimum rather than the global one.

However, both numerous experiments that demonstrate reversibility of protein folding and unfolding in vitro and computer experiments with folding and unfolding of lattice models of protein chains strongly suggest that the chains fold to their most stable structure, i.e., that the "native protein structure" is the lowest-energy one, and the protein folding (at least for not very long chains) is under thermodynamic rather than kinetic control (Šali et al. 1994; Abkevich et al. 1994).

Nevertheless, most of the proposed and widely discussed hypotheses on protein folding were based on the "kinetic control" (rather than "thermodynamic control") assumption.

In particular, before Levinthal, Phillips (1966) proposed that the protein folding nucleus is formed near the first-synthesized N-end of the nascent protein chain and the remaining chain wraps around it; but it has been shown later that the successful in vitro folding of many single-domain proteins and protein domains does not begin from the N-terminus (Goldenberg and Creighton 1983; Grantcharova et al. 1998; Lappalainen et al. 2008).

Wetlaufer (1973) hypothesized the formation of the folding nucleus by adjacent residues of the protein chain, but in vitro experiments have shown that this is not always so (Fulton et al. 1999; Wensley et al. 2009).

Ptitsyn (1973) proposed a model of hierarchical folding, i.e., a stepwise involvement of different interactions and the formation of different folding intermediate states. However, many not very long protein chains fold without visible folding intermediates (Fersht 1999).

More recently, various "folding funnel" models (Leopold et al. 1992; Wolynes et al. 1995; Dill and Chan 1997; Bicout and Szabo 2000; Dill et al. 2008; Wang et al. 2012) became very popular for illustrating and describing the reason for the speedy folding processes. This issue will be considered below in more detail.

The difficulty of the "kinetics vs stability" problem is that it hardly can be solved by a direct experiment. Indeed, suppose that a protein has some structure that is more stable than the native one (later we will demonstrate one of extremely few examples of this kind; it has been found for a rather long protein chain). How can we find the most stable but kinetically unattainable structure if the protein chain does not do so itself? Shall we wait for ~ $10^{10}$ (or even ~ $10^{180}$) years?

On the other hand, the question as to whether the protein structure is controlled by kinetics or stability arises again and again in solving practical problems of protein physics, engineering, and design. For example, when predicting the protein structure from its sequence, should we look for the most stable structure or the most rapidly attained one? When designing a de novo protein, should we maximize the stability of the desired fold or create a rapid pathway to this fold?

However, is there a contradiction between "the most stable" and the "rapidly folding" structure? Maybe, the stable structure *automatically* forms a focus for the "rapid" folding pathways, and therefore it is *automatically* capable of fast-folding?

## The major thermodynamic peculiarities of protein folding

Before considering these questions, i.e., before considering the *kinetic* aspects of protein folding, let us recall some basic experimental facts concerning protein *thermodynamics* (as usual, we shall only consider single-domain water-soluble globular proteins formed by chains of ~ 100 residues; and we will consider only those experiments in which individual proteins interact only with the solvent). These facts will help us understand what chains and folding conditions we have to consider. The facts are as follows:

1. Nearly all observations show that native states of single-domain water-soluble globular proteins behave as the lowest-energy folds (Tanford 1968; Privalov 1979;

Fersht 1999), i.e., they stay in this fold forever and also come to the same fold after de- and renaturation cycle induced by the change of a solvent. However, it should be mentioned that there is at least one exception: a large (≈ 400 residues) protein, serpin, at first obtains the “native” (that is, “working”) structure, works for half an hour, and then acquires another, non-working but more stable structure (Tsutsui et al. 2012).

2. The denatured state of proteins, at least that of small proteins treated with a strong denaturant, is usually an unfolded random coil (while the temperature-denatured state can be a compact molten globule) (Tanford 1968; Ptitsyn 1995).
3. Protein unfolding is reversible (Anfinsen 1973); moreover, the denatured and native states of a protein can be in a kinetic equilibrium (Creighton 1978); there is an “all-or-none” transition between these two states (Privalov 1979). The latter means that, close to the point of the folding-unfolding equilibrium, only two states of the protein molecule, native and unfolded, are present in a visible quantity, while all others, “semi-native” or misfolded states are virtually absent. (Notes: (i) the “all-or-none” transition makes the protein function reliable: like a light bulb, the protein either works or not; (ii) very important: the physical theory shows that such a transition requires the amino acid sequence that provides a large “energy gap” between the most stable protein structure and the bulk of misfolded and unfolded ones (Shakhnovich and Gutin 1990; Gutin and Shakhnovich 1993; Šali et al. 1994; Galzitskaya and Finkelstein 1995; Shakhnovich 2006; Finkelstein and Ptitsyn 2016)).
4. Even under normal physiological conditions, only a few kilocalories per mole (Privalov 1979) separate the native (i.e., the lowest-energy) state of a protein from its unfolded (i.e., the high-entropy) state (and at mid-transition, these two states have equal stabilities, of course).

(For the below theoretical analysis, it is essential to note that (i) as is customary in the literature on this subject, the term “entropy” as applied to protein folding only means conformational entropy of the chain without solvent entropy; (ii) accordingly, the term “energy” actually implies “free energy of interactions” (often called the “mean force potential”), so that hydrophobic and other solvent-mediated forces, with all their solvent entropy (Tanford 1968), come within “energy”. This terminology is commonly used (and will be used in this paper) to concentrate attention on the main problem of sampling the protein chain conformations.)

The above-mentioned “all-or-none” transition means that native ($N$) and denatured ($U$) states are separated by a rather high free-energy barrier. It is the height of this barrier that limits the rate of this transition, and just this height is to be estimated to solve Levinthal’s paradox.

## The major kinetic peculiarities of protein folding

The “kinetic control” hypothesis initiated very intensive studies of protein folding intermediates.

It was clear almost from the very beginning that the stable intermediates are *not* obligatory for folding, since the protein can also fold and unfold near the mid-point of equilibrium between the native and denatured states (Fig. 2) (Segava and Sugihara 1984; Fersht 1999), where the transition is of the “all-or-none” type (Privalov 1979), which excludes any stable intermediates.

The obtained basic experimental facts on folding *kinetics* of globular proteins are as follows:

1. The protein “"folding unit” is either a whole compact globular protein or a domain (compact sub-globule), if the protein includes several such sub-globules. This has been shown by two groups of evidence: (i) isolated domains, separated from the remaining protein body, are usually capable of folding into the correct structure (Petsko and Ringe 2004); (ii) single-domain proteins usually cannot fold when as few as 10 of their C- (or N-) terminal amino acid residues are deleted (Flanagan et al. 1992; Neira and Fersht 1999a,b).

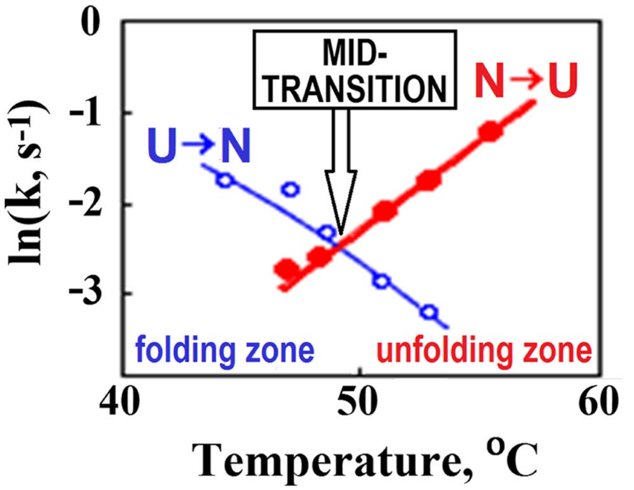

**Fig. 2** The rate ($k$) of lysozyme re- and denaturation vs temperature. The mid-transition point corresponds to ambient conditions where the rates of renaturation ($U \rightarrow N$ transition) and denaturation ($N \rightarrow U$ transition) are equal (i.e., where the blue and red lines intersect) so that the $U$ and $N$ states have equal free energies ($F_U = F_N$)). The plot is adapted from Segava and Sugihara (1984). Note that the folding at physiological temperatures of ≈40 °C is only about fivefold faster than that at the mid-transition point (≈ 50 °C). The similar in value but opposite in sign slopes of the $U \rightarrow N$ and $N \rightarrow U$ lines indicate that the transition state energy $E.^\#$ is close to the value intermediate between those for the native and denatured states (since, according to Arrhenius, $k_B T^2 \frac{d\ln k_{U \rightarrow N}}{dT} = E^\# - E_U$ and $k_B T^2 \frac{d\ln k_{N \rightarrow U}}{dT} = E^\# - E_N$ (Pauling 1970))

2. Folding of some proteins proceeds as a two-state ("all-or-none") process without any accumulating intermediates (when only two states, the native fold and the coil are observable (Matouschek et al. 1990; Fersht 1999)), whereas the folding of other single-domain proteins, mostly larger ones (and especially when the folding occurs far from the equilibrium mid-point) exhibit multi-state kinetics where molten and/or pre-molten globules serve as the folding intermediates (Dolgikh et al. 1984; Ptitsyn 1995; Fersht 1999).
3. When the folding process proceeds via the folding intermediates, the rate-limiting step immediately precedes the native state formation and corresponds to transition from the molten globule (often rather dense) to the native structure (Dolgikh et al. 1984).

## Understanding of the protein folding times

To begin with, it is not out of place considering whether the "Levinthal's paradox" is a paradox indeed. Bryngelson and Wolynes (1989) mentioned that this "paradox" is based on an absolutely flat (and therefore unrealistic) "golf course" model of the protein potential energy landscape (Fig. 3a), and somewhat later Leopold et al. (1992), following the line of Go and Abe (1981), considered more realistic (tilted and biased to the protein native structure) energy landscapes and introduced the "folding funnels" (Fig. 3b), which seemingly (but not indeed, see below) eliminate the "Levinthal's paradox".

Various "folding funnel" models became popular for explaining and illustrating protein folding (Wolynes et al. 1995; Karplus 1997; Nölting 2010; Wolynes 2015). In the funnel, the lowest-energy structure (formed by a set of the most powerful interactions) is the center surrounded by higher-energy structures containing only a part of these powerful interactions. The "energy funnels" may appear not perfectly smooth due to some "frustrations" (Bryngelson and Wolynes 1987), i.e., contradictions between optimal interactions for different links of a heteropolymer forming the protein globule, but a stable protein structure is distinguished by minimal frustrations (that is, most of its elements enhance the native fold stability) (Bryngelson and Wolynes 1987, 1989; Bryngelson et al. 1995; Finkelstein et al. 1995).

In principle, the "energy funnel" can channel the protein chain movement towards the single lowest-energy structure, thereby automatically turning this most stable structure of the chain into the "rapid" folding pathways, which seems (but… – see below) to be able to prevent the "Levinthal's" sampling of the vast majority of chain conformations.

However, this would be so provided there were *only* energy and no entropy, which (if the temperature is > 0 K) opposes the chain movement towards the single structure, even though corresponding to the global energy minimum.

But the protein folding occurs in liquid water, at temperatures $\gtrsim$ 273 K, where the entropy term is large; moreover, at the folding in proximity to the mid-transition conditions (Fig. 2), the entropy term nearly compensates the folding energy.

Mid-transition conditions are the best to analyze Levinthal's paradox (though under the "strongly folding" conditions the folding can be, say, 10- (Segava and Sugihara 1984) or even 1000-fold faster (Kiefhaber 1995; Fersht 1999) than at the mid-transition – but these 10 or 1000 times are incomparable with the puzzling 1 000 000 000 000 000 000-fold acceleration of the folding process compared to iterating over all structures).

In conditions corresponding to the mid-transition, the protein chain has two equally stable low-free-energy thermodynamic states (phases): "denatured" and "native." The latter includes the native structure (corresponding to the global free-energy minimum) and small fluctuations around it. The

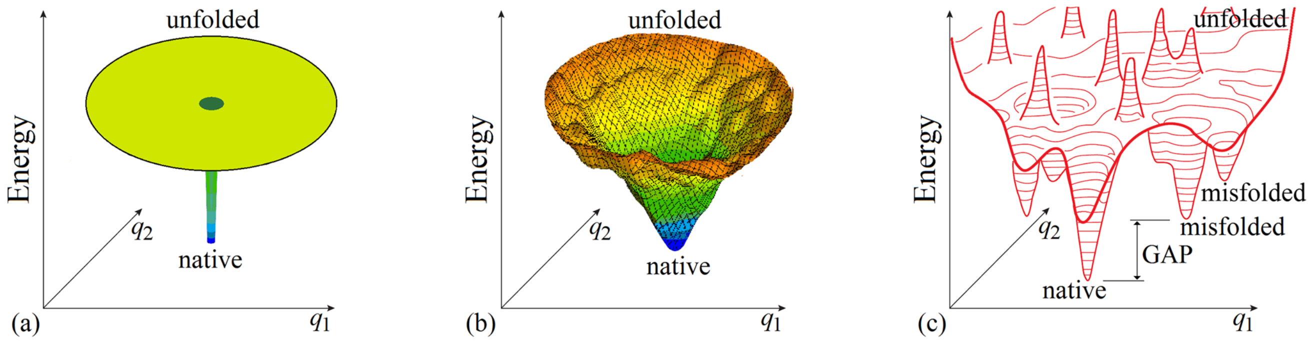

**Fig. 3** Schematic illustration of basic models of the energy landscapes of protein chains. (a) The "Levinthal's golf course model." (b) The "energy funnel" model; the funnel is centered in the lowest-energy ("native") structure. (c) The potential energy landscape of a protein chain in more detail with bumps and wells, the deepest of which ("native") is by many $k_B T_{\text{melt}}$ (where $k_B$ is Boltzmann's constant, and $T_{\text{melt}}$ is protein melting temperature) deeper than the others: the resulting energy GAP between the global and other energy minima is necessary to provide the "all-or-none" type of decay of the stable protein structure (Shakhnovich and Gutin 1990). Only two coordinates ($q_1$ and $q_2$) can be shown in the figures, while the protein chain conformation is determined by hundreds of coordinates

denatured state includes a multitude of the random coil-like conformations, molten-globule-like, "semi-native," and "misfolded" structures. The physical theory (Shakhnovich and Gutin 1990; Gutin and Shakhnovich 1993; Šali et al. 1994; Finkelstein and Ptitsyn 2016) shows that the co-existence of these two phases requires an amino acid sequence that provides a large "energy gap" (Fig. 3c) between the most stable (native) fold and the misfolded structures. It is this energy gap (present in 1 of approximately $10^{11}$ of random polypeptide chains, see Finkelstein and Ptitsyn (2016), lectures 16, 18 and appendix D, and Keefe and Szostak (2001)) that makes the protein fold unique and stable and keeps all misfolded structures very unstable. This allows neglecting misfolded structures when considering protein folding in conditions corresponding to the mid-transition (Finkelstein and Ptitsyn 2016).

The denatured and native states (phases) are separated by a free-energy barrier that provides the all-or-none phase transition between them (Privalov 1979), thus making the energy landscape acquire the "volcano-like" shape (Rollins and Dill 2014), where the funnel only remains in its center (Fig. 4).

Thus, any pathway from the unfolded state to the native one first goes uphill in free energy, and only then, in the vicinity to the native state, after passing the free-energy barrier (i.e., the crater edge), the "free-energy funnel" starts working and pulls the chain downhill to the native state. Note that if there were only a funnel and no barrier, then even a very large protein would fold not in minutes but in microseconds (since the time of conformational rearrangement of one residue is in the nanosecond time range (Zana 1975)).

However, to have a rapid transition from the coil to the native state, the free-energy barrier created by the volcano must be not too high: according to the conventional transition state theory (Eyring 1935; Pauling 1970; Emanuel and Knorre 1984), the time of overcoming the barrier is estimated as

$$TIME \approx \tau \times \exp(+\Delta F\#/k_{\mathrm{B}}T) \qquad (1)$$

where $\tau$ is the time of a step from the barrier onwards, and $\Delta F^{\#}$ is the height of the free energy barrier on the reaction pathway (that is, the free energy of the "folding nucleus").

It should be noted that protein folding is a multistep process (see Finkelstein and Ptitsyn (2016), lecture 19 and references therein), and that the conventional transition state theory is not very accurate when applied to multistep processes, including the protein folding (which is an intramolecular "all-or-none" phase transition (Privalov and Khechinashvili 1974)) and phase transitions in general (Djikaev and Ruckenstein 2016; Ruckenstein and Berim 2016). However, the error in this case only concerns the estimate of the pre-exponential factor ($\tau$ in Eq. (1)), being mainly the error in the estimate of the number of steps at the top of the barrier (Finkelstein 2015; Ruckenstein and Berim 2016), which is not too large in the case of protein folding. Therefore, the uncertainty in the pre-exponential factor is of secondary importance compared to the main, exponential term in Eq. (1), which accounts for the transition state free energy and can be enormous for a high barrier.

The energy funnel helps the fast-folding but does not guarantee that the whole process will be really fast. It is the height of the barrier (which is before the funnel) that determines the protein folding (and unfolding) rate. The energy funnel per se cannot resolve Levinthal's paradox, because *not any* type of energy funnel provides a low free energy barrier created by the edge of the volcano crater. A strict analysis (Bogatyreva and Finkelstein 2001) of the straightforwardly presented funnel models (Zwanzig et al 1992; Bicout and Szabo 2000) corresponding to the uniform condensation of the chain (previously considered by Shakhnovich and Finkelstein (1989)) shows that close to the mid-transition point, such funnels cannot *simultaneously* explain both major features observed in protein folding: (i) the "all-or-none" type of transition, which requires the free-energy barrier; and (ii) the non-astronomical folding time. By the way, the stepwise folding mechanism (Ptitsyn 1973) also cannot (Finkelstein 2002) *simultaneously* explain both of these major features close to the mid-transition point, and hence, also cannot resolve Levinthal's paradox.

Resolution of Levinthal's paradox requires funnels of a special type — those provided by a transient separation of folded and unfolded phases within the folding chain

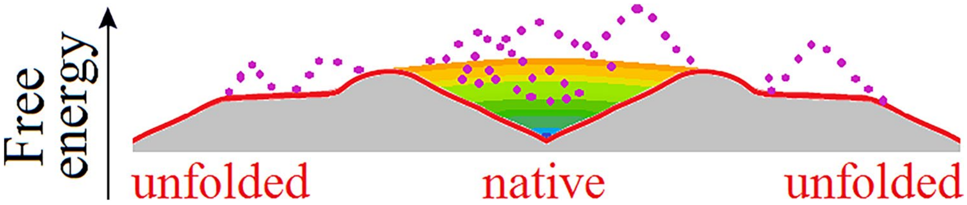

**Fig. 4** This purely illustrative drawing shows how entropy converts the *energy* funnel (Fig. 3b) into a "volcano-shaped" *free-energy* folding landscape with a barrier on any pathway leading from unfolded conformations to the native fold. The smooth free-energy landscape corresponds to compact partly folded structures; the rocks (denoted by dotted lines) present high-energy structures that are non-compact or contain high-energy bumps (see Fig. 3c). A more accurate but less beautiful scheme of the free-energy landscape is shown in Fig. 2 in Galzitskaya and Finkelstein (1999)

(Finkelstein and Badretdinov 1997a, b) (this, as subsequently mentioned in a review by Wolynes (1997), resembles the “capillarity”" theory of nucleation of the first-order phase transitions; the transient separation of the folded and unfolded phases in the course of protein folding was later demonstrated in computer simulations by Shaw et al. (2010)).

It is not as easy to theoretically find a good protein folding pathway. It is much easier to figure out how a good (low-free-energy) unfolding pathway should look like. The compactness of protein globules suggests the existence of surface tension, which results from the free-energy excess at the surface of the globule. Thus, a low-free-energy pathway of the unfolding of the globule to the coil should proceed via the least unstable partly folded structures consisting of two phases (native and unfolded) separated by a relatively small boundary: the globule's cross section that separates the remaining dense, compact part of the globule, and the unfolded loops and tails protruding from it (Fig. 5) (Finkelstein and Badretdinov 1997a, b; Galzitskaya and Finkelstein 1999; Garbuzynskiy et al. 2013).

This good pathway of *unfolding*, when followed in the opposite direction, presents a good pathway of *folding* (Finkelstein et al. 2017) because, according to the well-known in physics *detailed balance* law (Landau and Lifshitz 1980), the direct and reverse reactions, under the same ambient conditions, follow the same pathway and have equal rates when both end-states have equal stability: otherwise, i.e., if the pathways for $A \longrightarrow B$ and $A \longleftarrow B$ reactions were different, the result would be a *permanent* circular flow $A \rightleftarrows B$ (generating, at thermodynamic equilibrium, a perpetual motion machine of the second kind), which contradicts to the second law of thermodynamics.

(Two notes: (i) To resolve Levinthal's paradox, it is not necessary to prove that the above outlined pathway is *the* best possible pathway; it is enough to prove that this pathway resolves the paradox, because any additional pathway will only accelerate the process. Imagine two pools, one full of water and another empty, with water leaking from one to the other through cracks in the wall between them; if the cracks cannot absorb all the water — which is prohibited by the all-or-none kind of transition — each additional crack accelerates filling the empty pool. (ii) The same, of course, applies to additional folding pathways passing through folding intermediates, which are sometimes observed (Aviram et al. 2018) in apparently two-state transitions. (iii) Actually, the pathway itself is of no interest for us here; according to the transition state theory, only the barrier, i.e., the free-energy maximum on the pathway, is important indeed).

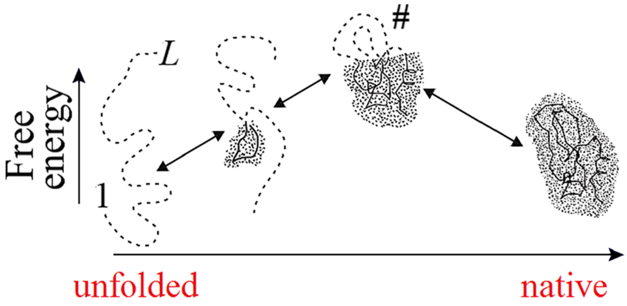

**Fig. 5** Schematic illustration of a sequential folding/unfolding pathway of a globule through compact partly folded intermediate structures. At each step of sequential *un*folding, one residue leaves the native-like part of the globule (shaded) and turns into a coil (shown by a dashed line); the sequential folding follows the same pathway in the opposite direction. The highest-free-energy intermediate structure (i.e., the folding nucleus corresponding to the transition state; marked as #) has the largest (in the pathway) interface of the globular and unfolded phases. Its globular part covers about half of the chain. Adapted from (Finkelstein and Badretdinov 1997a, b)

In a simplified form (for details, see Finkelstein and Badretdinov (1997a, b; 1998; Garbuzynskiy et al. 2013)), the resulting free-energy barrier is estimated as follows.

When the free energies of the folded and unfolded phases are equal (i.e., in the mid-transition ambient conditions), the free energy of a semi-folded protein depends only on the interface between the two phases.

The largest unavoidable interface corresponds to the transition state (structure # in Fig. 5) that looks like a half of the native globule and has $\approx L^{2/3}$ residues at the interface (assuming the most compact spherical shape of the native globule; for an oblate or oblong globule, the largest unavoidable interface can be a little less).

Thus, the transition state free energy is proportional *not* to the number $L$ of the chain residues (as Levinthal's estimate implies), but to $L^{2/3}$ only.

The energy constituent $\Delta E^{\#}$ of the barrier free-energy $\Delta F^{\#}$ results from interactions lost by the interface residues; it is about $(L^{2/3}) \bullet \varepsilon/4$, where $\varepsilon \approx 1.3$ kcal/mol $\approx 2k_{\mathrm{B}}T_{\mathrm{mel}}$ is the average latent heat of protein melting per residue (Privalov 1979) (this $\varepsilon$ is the first empirical parameter used by the theory), and $\approx 1/4$ is, roughly, the fraction of interactions lost by an interface residue (which has lost, roughly, 1 of 6 neighbors in space that it had inside the globule (1 “up,” 1 “down,” and 4 neighbors along the future interface), but 2 of these 6 neighbors in space cannot be lost — they are its neighbors in the chain). Thus,

$$\Delta E^{\#}/k_B T_{\mathrm{melt}} \approx 0.5L^{2/3} \tag{2}$$

The entropy constituent $\Delta S^{\#}$ of the barrier free-energy $\Delta F^{\#}$ is caused by entropy lost by closed loops protruding from the globular into the unfolded phase (note that the partially folded state, denoted as # in Fig. 5, contains two closed loops, and the another partially folded state in Fig. 5 contains no closed loops).

When the shape of the native protein fold and especially the shape of the chain in the transition state are not known, the closed-loops-connected $\Delta S^{\#}$ value (which is $\leq 0$, because it is due to restriction of loop conformations) can only be estimated — from above and from below.

The upper limit of $\Delta S^{\#}$ is zero (when the interface contains no closed loops).

The lower limit of $\Delta S^{\#}$ is about.

$$(\Delta S^{\#})_{\text{lower}} \approx \tfrac{1}{6}(L^{2/3}) \bullet \left[-\tfrac{5}{2}k_{\text{B}} \ln(3L^{1/3})\right] \tag{3}$$

Here, $\frac{1}{6}(L^{2/3})$ is the maximal expected number of loops protruding from the maximal (containing $\approx L^{2/3}$ residues) unavoidable interface. Actually, $\frac{1}{6}(L^{2/3})$ is the average number of loops protruding from the interface containing $L^{2/3}$ residues. The multiplier $\frac{1}{6}$ results from the fact that the chain can have, roughly, 6 directions in each interface residue (4 along the interface, 1 inside the folded part, and only 1 looking outside, thereby initiating a loop). Among many possible cross-section interfaces dividing the globule into two halves, the lowest-free-energy interface should serve for the transition state in the folding/unfolding pathway. Therefore, this "optimal" interface should be covered by no more than $\frac{1}{6}(L^{2/3})$ or possibly a smaller number of closed loops.

The value $3L^{1/3} \equiv (L/2)/(\frac{1}{6}L^{2/3})$ is the average number of residues in a closed loop in the transition state ($L/2$ being the number of unfolded residues in the transition state and $\frac{1}{6}L^{2/3}$ the maximal number of closed loops there). The value $-\frac{5}{2}k_{\text{B}} \ln(3L^{1/3})$ is the entropy lost by a $3L^{1/3}$-residue closed loop at the interface (such a loop cannot cross the interface plane; this restriction changes 3/2, the conventional Flory's (1969) coefficient for the entropy of an unrestricted closed loop, for 5⁄2 (Finkelstein and Badretdinov 1997a, b)). Having $L \sim 100$ (actually, this approximation is good for the whole range of $L = 10$–$1000$), one obtains

$$(\Delta S^{\#})_{\text{lower}} \approx -\frac{5}{12}k_{\text{B}}L^{2/3}\left[\ln(3) + \frac{\ln(L)}{3}\right] \approx -k_{\text{B}}L^{2/3} \tag{3a}$$

In the mid-transition ambient conditions, the corresponding transition state free energy, $\Delta F_0^{\#}$, equals to $\Delta E^{\#} - T_{\text{melt}}\Delta S^{\#}$. The $\Delta F_0^{\#}$ value is not less than $\Delta E^{\#}0$ (when $\Delta S^{\#} = 0$) and not larger than $\Delta E^{\#} - T_{\text{melt}}(\Delta S^{\#})_{\text{lower}}$, that is,

$$[\Delta E^{\#} \approx 0.5L^{2/3}k_{\text{B}}T_{\text{melt}}] \le \Delta F_0^{\#} \le [\Delta E^{\#} - T_{\text{melt}}(\Delta S^{\#})_{\text{lower}} \approx 0.5L^{2/3}k_{\text{B}}T_{\text{melt}} + L^{2/3}k_{\text{B}}T_{\text{melt}}] \tag{4}$$

Thus, when the free-energy difference $\Delta F$ between the native (the most stable) and the unfolded state is equal to zero, the time of both folding and unfolding of the $L$-residue protein chain is estimated as

$$TIME_{\Delta F=0} \approx \tau \times exp\left[+\Delta F_0^{\#}/k_B T_{\text{melt}}\right] \sim \tau \times \exp\left[+(0.5 \div 1.5)L^{2/3}\right] \tag{5}$$

where $\tau \approx 10$ ns is the time of structure growth by one residue (Zana 1975) (this $\tau$ is the second and the last empirical parameter used in the theory (Finkelstein and Badretdinov 1997a, b)).

Here, one thing should be added: A search over folds with different chain knotting can, in principle, create a rate-limiting "quasi-Levinthal" factor since the knotting cannot be changed without globule decay. However, since the computer experiments show that one chain knot involves many tens of residues (Grosberg 1997), this factor for the chain of 100–200 residues can be $2^2$–$2^4$ only, and the search for correct knotting can only be rate-limiting for extremely long ($L> >1000$) chains (Finkelstein and Badretdinov 1998) that cannot fold within a reasonable time (according to Eq. (5)) in any case.

The above Eq. (5) shows that in the mid-transition conditions (where $\Delta F = 0$), a ≈100-residue protein chain should attain its most stable fold within milliseconds or days, but not years.

If the native fold is more stable than the unfolded state (i.e., if $\Delta F < 0$), the folding is faster. Because the folding nucleus covers about half of the chain (more detailed calculations give ≈40% (Garbuzynskiy et al. 2013)), its free energy decreases from $\Delta F_0^{\#}$ (that was at $\Delta F = 0$) to approximately $\Delta F_0^{\#} + 0.4\ \Delta F$ at $\Delta F < 0$, so that

$$TIME_{\Delta F<0} \sim TIME_{\Delta F=0} \times \exp[+0.4\Delta F/k_{\text{B}}T] \tag{6}$$

which can be approximately presented as

$$TIME_{\Delta F<0} \sim 10\text{ns} \times \exp\left[+(0.5 \div 1.5) \times \left(L^{2/3} + 0.4\Delta F/k_B T\right)\right] \tag{6a}$$

(Garbuzynskiy et al. 2013). Because the value $\Delta F \approx$ 40 kJ/mol for a ≈100-residue protein under physiological conditions (Privalov 1979), the folding time of such a protein decreases by about 500-fold, and now ranges from a fraction of a millisecond to tens of minutes.

It should be noted that all the above considerations are focused on the case of the moderate stability of the native fold, which corresponds to the available data on protein folding (occurring near the mid-transition point, see Fig. 2). For the opposite case of a very high native fold stability ($-\Delta F> >k_{\text{B}}T$), another but similar to Eq. (5) scaling law ($\ln(TIME) \sim L^{1/2}$) was obtained by Thirumalai (1995).

Conclusion: one can see that although the protein folding problem is the so-called "NP-hard" problem (Ngo and Marks 1992; Unger and Moult 1993) (which loosely speaking implies an exponentially-long time to be spent to solve it by a folding chain or by a computer), and indeed the time is, in the main term, a stretched-exponential function of the chain length $L$ (see Eqs. (3a), (5), (6a), and the later rigorous mathematical papers (Fu and Wang 2004; Steinhofel et al. 2006)), this does not mean that this time is unreasonably long for a normal-size protein domain of ~ 100 residues.

## Protein folding times: theory and experiment

The observed protein folding times (see Fig. 6) span over 11 orders of magnitude (which is akin to the difference between the lifespan of a mosquito and the age of the universe).

Figure 6 shows the region theoretically allowed in Garbuzynskiy et al. (2013) for the folding times by Eqs. (5)–(6a) (obtained with only two empirical and no adjustable parameters) and describes the observed folding times of all studied before 2013 single-domain globular proteins of any size and stability of their native state.

Figure 6 also shows that a chain of $L \lesssim$ 80–90 residues will find its most stable fold within minutes (or faster) even under "non-biological" mid-transition conditions, where folding is known (Creighton 1978; Fersht 1999) to be the slowest (see also Fig. 2). Thus, native structures of such relatively small proteins are under complete thermodynamic control: they are the most stable among all structures of these chains. In other words, any possible lowest-energy fold can be achieved at a "biologically reasonable" time for these small proteins.

Native structures of larger proteins (of $\gtrsim$ 100 to ≈450 residues) are, in addition, under a kinetic "control of complexity," in a sense that too entangled (due to, e.g., complicated β-sheets) folds of their long chains (having too many intersections with any globule's cross section) cannot be achieved within days or weeks even if they are thermodynamically stable; indeed, globular domains with greatly entangled folds of long protein chains have never been observed (Garbuzynskiy et al. 2013): they seem to be excluded from the repertoire of existing protein structures. Besides, the native fold of at least one protein (serpin) of ≈400 residues is not the most stable but a long-living metastable fold (Tsutsui et al. 2012).

The kinetic control also explains why larger (with $L \gtrsim$ 450) proteins should have far from spherical shape or consist (according to the "divide and rule" principle) of separately folding domains: otherwise, chains of more than 450 residues would fold too slowly. This is a kinetic "size restriction" for domains. In essence, this effect resembles Levinthal's "kinetic control," though at another level and

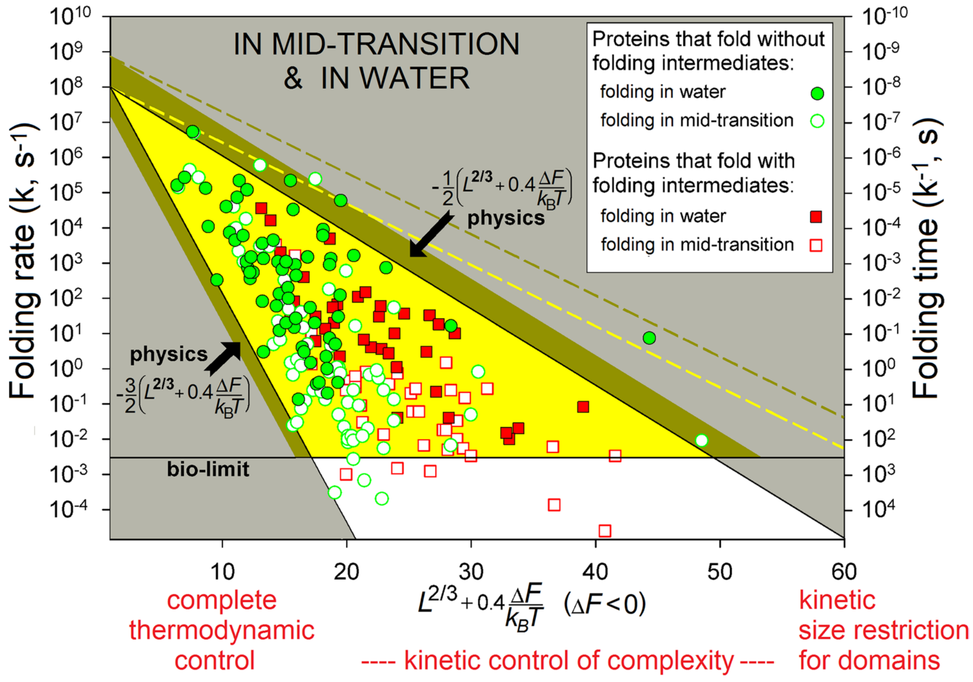

**Fig. 6** Folding rates and times. Experimental in vitro measurements have been made "in water" (under approximately "biological" conditions) and at mid-transition for 107 single-domain proteins (or separate domains) without SS bonds and covalently bound ligands (though the rates for proteins with and without SS bonds are principally the same (Galzitskaya et al. 2001)). The golden-and-white triangle: the region theoretically allowed by physics at the mid-transition. Its golden part corresponds to biologically-reasonable folding times (≤10 min); the bronze belt is the additional area allowed in "biological" conditions. The white zone: the larger folding times (i.e., the lower folding rates) are observed (for some proteins) only under mid-transition (i.e., "non-biological") conditions. The yellow dashed line limits the additional area allowed for oblate (1:2) and oblong (2:1) globules at mid-transition; the bronze dashed line means the same for "biological" conditions. $L$ is the number of amino acid residues in the protein chain. $\Delta F$ is the free energy difference between the native and unfolded states of the chain under the experimental conditions and temperature $T$ close to 300 K. Adapted from (Garbuzynskiy et al. 2013)

only for very large proteins. The above estimates (≈100 and ≈400 residues) are somewhat (by 30–50%) elevated when the native fold free-energy $\Delta F$ is substantially lower than that of the unfolded chain, but essentially they remain nearly the same (Garbuzynskiy et al. 2013).

Equations (5)–(6a) outline the range of folding times depending on the protein size and stability of its native structure under given ambient conditions. To predict the protein folding time more accurately, the shape of its folding nucleus or, for lack of such information, its native fold should be taken into account. So did Plaxco et al. (1998), who introduced a "contact order" (CO, that equals to the average chain separation of the residues that are in contact in the native protein fold, divided by the chain length) as a phenomenological measure of complexity of the native fold (though, CO "works" well only for small proteins that fold without folding intermediates). Later, this CO was added (Ivankov et al. 2003) to the already developed (Finkelstein and Badretdinov 1997a, b) chain length dependence, and the resulting method (Ivankov et al. 2003) showed quite good results, now for all proteins; in particular, it was shown that α-proteins (having low CO due to intra-helical H-bonds) fold faster than other proteins of the same size (Ivankov and Finkelstein 2004), though large α-proteins (with low CO) fold much slower than small β-proteins (with high CO). The subsequent extension of this method (Finkelstein et al. 2013; Ivankov and Finkelstein 2020) gave even more accurate results.

It should be added that no attention was paid in these works to specific 3D structures of folding nuclei; the attention was only paid to their overall features like size, instability and complexity. The reason: although, in some cases, there is evidence that folding nuclei are well-organized and possess specific structural features (see Fersht 1999, 2000; Garbuzynskiy and Kondratova 2008; Shaw et al. 2010)), in other cases, they are poorly organized ("diffused nuclei") (see (Grantcharova et al. 2001; Finkelstein et al. 2007, 2014) and references therein). The latter, together with the observed sensitivity of positions and shapes of the folding nuclei to mutations, led to the conclusion that a "nucleus" is an ensemble of structures rather than a single structure (Galzitskaya and Finkelstein 1999; Garbuzynskiy et al 2004) and that the folding nucleus and folding pathway are much less resistant to amino acid sequence mutations and change of ambient conditions than the native protein structure.

Also, it should be noted that all the above considerations were focused on stability (or rather, instability) of transition states (folding nuclei), and virtually no attention was paid to folding intermediates, because these — in contrast to transition states — do not determine the rate of folding of native protein structures (Fersht 1999, 2000).

## Dependence of the number of compact chain folds (and of the time of iterating over them) on the protein size

The total ("Levinthal's") volume of the protein conformation space estimated at the level of amino acid residues is huge: $\gtrsim 3^{100}$ conformations for a 100-residue chain (see above).

However, should the chain sample all these conformations in search for its most stable fold? No, a vast majority of them are non-compact (that is, high-energy ones) and should not be examined, but the conformation space is covered by local energy minima, each surrounded by a local energy funnel (Fig. 7) providing fast downhill decent to this local minimum. And, actually, the folding protein chain only has to sample various chain folds within these local energy funnels leading to compact protein globules.

To estimate the actual volume of this sampling, one has to estimate the number of low-free-energy local energy minima. This is similar to the idea of enumerating all possible "topomers" that a protein chain can form (Debe et al. 1999; Makarov and Plaxco 2003; Wallin and Chan 2005).

An overview of protein 3D structures shows that interactions occurring in the chains are mainly connected with secondary structures (Levitt and Chothia 1976; Chothia and Finkelstein 1990; Finkelstein and Ptitsyn 2016). Thus, a question arises as to how large the total number of energy minima is if considered at the level of formation and assembly of secondary structures into a compact globule, that is, at the level considered by Ptitsyn (1973) in his model of stepwise protein folding.

We will be interested mostly in proteins that fold under thermodynamic control, that is, those having chains of $L \approx 100$ or less amino acid residues (see above). Such proteins have no more than 10 α- and β-structural elements (Ptitsyn and Finkelstein 1980; Rollins and Dill 2014).

The number of compact globular packings of the chain is by many orders of magnitude smaller than that of conformations of amino acid residues (Finkelstein and Garbuzynskiy 2015): the latter, according to Levinthal's estimate, scales up as something like $100^L$ or $10^L$ or $3^L$ with the number $L$ of residues in the chain, while the former scales up not faster (see below) than $\sim L^N$ with the chain length $L$ and the number $N$ of the secondary structure elements. $N$ is much less than $L$ ($N < L/10$, according to Rollins and Dill (2014)), and this drastic decrease of the power $N$ as compared to $L$ is the main reason for the drastic decrease of the conformation space.

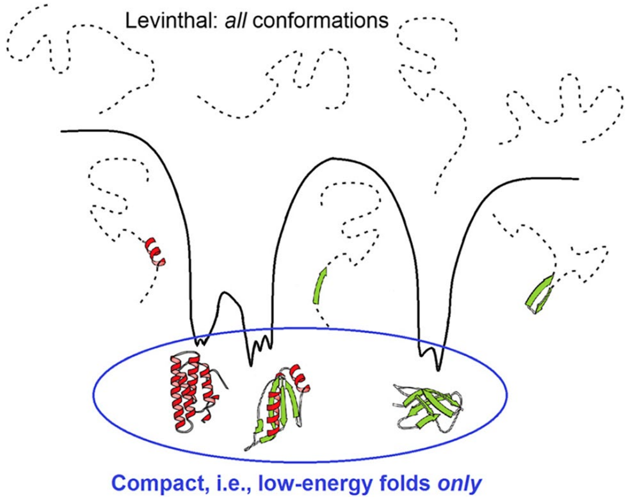
Levinthal: *all* conformations
Compact, i.e., low-energy folds *only*

**Fig. 7** Comparison of a huge search among *all*, mostly disordered, conformations and a much less voluminous search *only* among compact and well-structured globules, thus corresponding to the deep energy minima surrounded by energy funnels. Adapted from (Finkelstein 2017)

The number of compact globular packings of the chain with given secondary structures can be presented (Finkelstein and Garbuzynskiy 2015) as a product of the following multipliers (Fig. 8).

$M_A$, the number of Architectures, i.e., types of dense stacks of given secondary structures. This number is small (cf. (Levitt and Chothia 1976; Murzin and Finkelstein 1988; Chothia and Finkelstein 1990)). It is usually (at $L \lesssim 100$ and $N \lesssim 10$) about 10 or less architectures (Fig. 8a) for a given set of secondary structures, since the architectures are packings of a few secondary structure layers (each containing several secondary structures), and therefore the combinatorics of the layers is very small, as compared to that of much more numerous secondary structure elements, which is described below.

$M_P$, the number of all possible combinations of positions of $N$ structural elements within the given protein architecture that cannot exceed $N! \equiv N \times (N-1) \times \ldots \times 2 \times 1$ (Fig. 8b).

$M_T$, the number of all possible topologies, i.e., all combinations of directions of these structural elements that cannot exceed $2^N$ (Fig. 8c).

The above means that the number of compact packings of $N$ secondary structure elements ("topomers") is about $M_A \times M_P \times M_T \approx 10 \times N! \times 2^N$. Using Stirling's approximation ($N! \approx (N/e)^N$), we have

$$NUMBER\ of\ topomers \approx M_A \times M_P \times M_T \approx \left[10 \times \left(\frac{2}{e}\right)^N\right] \times N^N \approx N^N \tag{7}$$

in the main term at $N >> 1$.

Each of these topomers contains $M_{S \times T} \sim (L/N)^N$ local energy minima connected with shifts and turns of secondary structure elements within a topomer (Fig. 8d).

$M_{S \times T}$ is this number of possible shifts and turns of structural elements within the dense globule. Here, transverse shifts and tilts are prohibited by the dense packing, while longitudinal shifts and rotations of structural elements are coupled (this is shown in (Fig. 8d) using a β-sheet as the

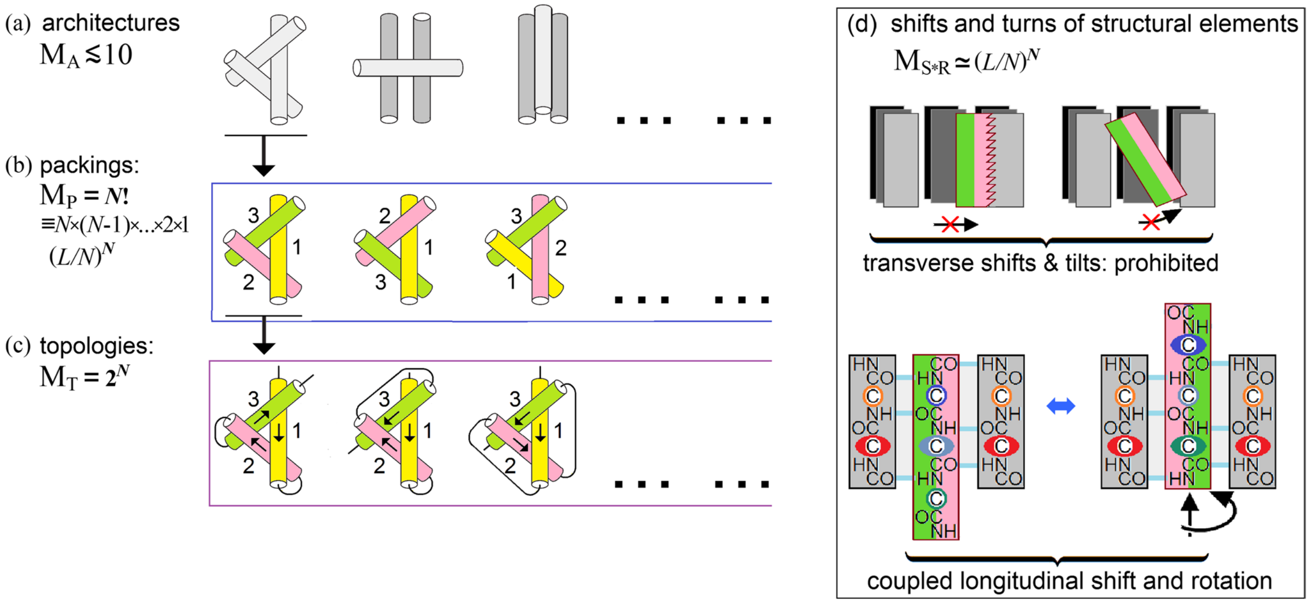
(a) architectures
$M_A \lesssim 10$
(b) packings:
$M_P = N!$
$\equiv N \times (N-1) \times \ldots \times 2 \times 1$
$(L/N)^N$
(c) topologies:
$M_T = 2^N$
3
2
1
(d) shifts and turns of structural elements
$M_{S*R} \simeq (L/N)^N$
transverse shifts & tilts: prohibited
HN
CO
C
NH
OC
coupled longitudinal shift and rotation

**Fig. 8** Scheme for estimating the volume of the conformational space at the level of secondary structure assembly and packing. Explanations are given in the text. Adapted from Supplement to (Finkelstein and Garbuzynskiy 2015)

best illustrative example, but this is also true for α-helices — remember their "knobs in the holes" close packings by Crick (1953)). As a result, each of $N$ $\alpha$- and β-element can have about $L/N$ (that is, about the number of chain residues per an element) possible shifts/turns in the globule formed by $N$ secondary structures in the $L$-residue chain.

So, the

$$\begin{aligned} \mathit{NUMBER\ of\ energy\ minima\ to\ be\ sampled} &\approx \left(M_{\mathrm{A}} \times M_{\mathrm{P}} \times M_{\mathrm{T}}\right) \times M_{S\times T} \\ &\approx N^N \times (L/N)^N = L^N \end{aligned} \quad (8)$$

in the main term (if $L > > N > > 1$) (Finkelstein and Garbuzynskiy 2015).

This number can be somewhat reduced by the symmetry of the globule, by shortness of some loops, by the impossibility to have α-helices inside β-sheets, etc., but this is not important in estimating the upper limit of the number of conformations (Finkelstein and Garbuzynskiy 2015).

As to the question of how the chain knows where and what secondary structures to form, the answer is that most of the secondary structures are determined by local amino acid sequences (Ptitsyn and Finkel'shtein 1970; Ptitsyn 1973; Lim 1974a, b; Chou and Fasman 1974; Schulz et al. 1974; Ptitsyn and Finkelstein 1983; Finkelstein et al. 1990; Jones 1999; etc.).

Because in a chain of $L$ ≈ 20 residues one ($N = 1$) α-helix forms within ≈0.2 μs (Mukherjee et al. 2008), and a β-hairpin of $N = 2$ β-strands forms within ≈6 μs (Muñoz et al. 1997), the time necessary for iterating over ~ $L^N$ of possible assemblies of the secondary structures can be estimated (cf. Equation (6a)) as.

$$\mathit{TIME\ for\ iterating} \sim 10\ \mathrm{ns} \times L^N \quad (8a)$$

In a compact globule, the length of a secondary structure element should be proportional to the globule's diameter, i.e., to ~ $L^{1/3}$. More specifically (taking into account volumes of amino acid residues and their length along the chain and α and β structures), a diameter of a globule of $L$ residues is ≈5 $L^{1/3}$ Å, and thus, on the average, α helix consists of ≈3 $L^{1/3}$ residues, while a β-strand, as well as a loop, comprises ≈1.5 $L^{1/3}$ residues. Thus, an α-helical globule (consisting of α-helices connected by loops) contains ≈$L/[L^{1/3}(3 + 1.5)] = L^{2/3}/4.5$ helices, and a β-structural globule (consisting of β-strands connected by loops) contains ≈$L/[L^{1/3}(1.5 + 1.5)] = L^{2/3}/3$ β-strands (Finkelstein and Garbuzynskiy 2015). This means that

$$\begin{aligned} \mathit{NUMBER\ of\ structural\ elements}\ N &\approx \frac{\ln(L)}{4.5}\ \mathit{for\ \alpha\text{-}\ proteins} \\ &— \frac{\ln(L)}{3}\ \mathit{for\ \beta\text{-}proteins} \end{aligned} \quad (9)$$

Thus, the value $L^N$ of possible secondary structure assemblies is expected to come within the range

$$\begin{aligned} L^{L^{2/3}/4.5} &\equiv \exp(\ln(L)4.5 \times L^{\frac{2}{3}})\ \mathit{for\ \alpha\text{-}proteins}\ —L^{\frac{L^{\frac{2}{3}}}{3}} \\ &\equiv \exp(\frac{\ln(L)}{3} \times L2/3)\ \mathit{for\ \beta\text{-}proteins} \end{aligned} \quad (10)$$

Since $\ln(L = 50 \div 150) = 4 \div 5$, the outlined range of possible secondary structure assemblies, $L^N$, can be estimated, for domains of $L \approx 100$ residues, as

$$L^N \approx \exp(L^{2/3}) — \exp(1.5L^{2/3}) \quad (11)$$

So, the number of the secondary structure assemblies scales with the chain length $L$ approximately as the upper boundary of the range of folding times outlined by Eq. (5) (Finkelstein and Badretdinov 1997a, b), and

$$\begin{aligned} \mathit{TIME\ for\ iterating} &\sim 10\ \mathrm{ns} \times L^N \approx 10\ \mathrm{ns} \times \exp(L^{2/3}) \\ &—10\ \mathrm{ns} \times \exp(1.5L^{2/3}) \end{aligned} \quad (12)$$

coincides with the upper boundary of the range of folding times given by Eq. (5).

It is not out of place mentioning that the scaling of $L^N$ given by Eq. (10) looks exactly like those obtained by Fu and Wang (2004) and Steinhofel et al. (2006) from mathematical consideration of the folding problem complexity for a chain consisting of only two kinds of links ("hydrophobic" and "hydrophilic" ones) rather than from physical reasons.

## Conclusion

The point of this article is not to explain *how* proteins fold (this needs experimental studies of many proteins of various kinds); the point is to explain *why* a protein chain *is able* to choose, in minutes, its unique most stable 3D structure among an enormous number of alternatives.

Throughout the article, we have only considered folding (and unfolding) of a single protein chain that does not interact with anything but a solvent.

Our review is mostly theoretical; it aims to clarify a physical theory behind our understanding of protein folding. The reason for this theoretical accent is that the famous Levinthal's paradox, which concentrates the essence of protein folding enigma, is itself, actually, a theoretical concept, and hence its "ultimate" resolution is also expected to be theoretical. Otherwise, this paradox is doomed to remain unsolved and not understood. This paradox cannot be solved by a direct experiment (which would need enormous time and an experimental investigation of folding of all possible polypeptide sequences), and even this would only give a result, but not its understanding. Besides but not the least: solving the Levinthal's paradox, the presented theory generates experimentally testable predictions that turn out to be correct (see, e.g., Fig. 6 of the review, where 212 out of 214 experimental points fall into the theoretically predicted region).

**Acknowledgements** The authors are grateful to O. B. Ptitsyn, A. M. Gutin, E. I. Shakhnovich, A. Ya. Badretdinov, O. V. Galzitskaya for many years of fruitful collaboration and discussions and to E. V. Serebrova, N. Y. Marchenko, and O. A. Eremeeva for editing the manuscript.

**Funding** This work was supported by the Russian Science Foundation (Grant numbers 14–24-00157 and 21–14-00268).

## Declarations

**Ethical approval** This article does not contain any studies with human participants or animals performed by any of the authors.

**Consent to participate** Verbal informed consent was obtained from all the individual participants included in the study.

**Consent for publication** Not applicable.

**Competing interests** The authors declare no competing interests.

## References

Abkevich VI, Gutin AM, Shakhnovich EI (1994) Specific nucleus as a transition state for protein folding: evidence from the lattice model. Biochemistry 33:10026–10031. https://doi.org/10.1021/bi00199a029

Agirrezabala X, Samatova E, Macher M, Liutkute M, Maiti M, Gil-Carton D, Novacek J, Valle M, Rodnina MV (2022) A switch from α-helical to β-strand conformation during co-translational protein folding. EMBO J 41(4):109175. https://doi.org/10.15252/embj.2021109175

Anfinsen CB (1959) The molecular basis of evolution, chapters 5, 6. John Wiley, New York. https://profiles.nlm.nih.gov/101584571X121

Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181:223–230. https://doi.org/10.1126/science.181.4096.223

Anfinsen CB, Scheraga HA (1975) Experimental and theoretical aspects of protein folding. Adv Protein Chem 29:205–300. https://doi.org/10.1016/s0065-3233(08)60413-1

Anfinsen CB, Haber E, Sela M, White FH Jr (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci USA 47:1309–1314. https://doi.org/10.1073/pnas.47.9.1309

Aviram HY, Pirchi M, Barak Y, Riven I, Haran G (2018) Two states or not two states: single-molecule folding studies of protein L. J Chem Phys 148:123303. https://doi.org/10.1063/1.4997584

Berman HM, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. Nat Struct Biol 10(12):980. http://www.wwpdb.org/; https://doi.org/10.1038/nsb1203-980

Bicout DJ, Szabo A (2000) Entropic barriers, transition states, funnels, and exponential protein folding kinetics: a simple model. Protein Sci 9:452–465. https://doi.org/10.1110/ps.9.3.452

Bogatyreva NS, Finkelstein AV (2001) Cunning simplicity of protein folding landscapes. Protein Eng 14:521–523. https://doi.org/10.1093/protein/14.8.521

Bryngelson JD, Wolynes PG (1987) Spin glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci USA 84:7524–7528. https://doi.org/10.1073/pnas.84.21.7524

Bryngelson JD, Wolynes PG (1989) Intermediates and barrier crossing in a random energy model (with applications to protein folding). J Phys Chem 93:6902–6915. https://doi.org/10.1021/J100356A007

Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins 21:167–195. https://doi.org/10.1002/prot.340210302

Chothia C, Finkelstein AV (1990) The classification and origins of protein folding patterns. Ann Rew Biochem 59:1007–1039. https://doi.org/10.1146/annurev.bi.59.070190.005043

Chou PY, Fasman GD (1974) Prediction of protein conformation. Biochemistry 13:222–245. https://doi.org/10.1021/bi00699a002

Creighton TE (1978) Experimental studies of protein folding and unfolding. Prog Biophys Mol Biol 33:231–297. https://doi.org/10.1016/0079-6107(79)90030-0

Crick FHC (1953) The packing of α-helices: simple coiled coils. Acta Crystallogr 6:689–697. https://doi.org/10.1107/S0365110X53001964

Debe DA, Carlson MJ, Goddard WA 3rd (1999) The topomer-sampling model of protein folding. Proc Natl Acad Sci USA 96:2596–2601. https://doi.org/10.1073/pnas.96.6.2596

Dill KA, Chan HS (1997) From Levinthal to pathways to funnels. Nat Struct Biol 4:10–19. https://doi.org/10.1038/nsb0197-10

Dill KA, MacCallum JL (2012) The protein-folding problem, 50 years on. Science 338:1042–1046. https://doi.org/10.1126/science.1219021

Dill KA, Ozkan SB, Shell MS, Weikl TR (2008) The protein folding problem. Annu Rev Biophys 37:289–316. https://doi.org/10.1146/annurev.biophys.37.092707.153558

Djikaev Y, Ruckenstein E (2016) Model for the nucleation mechanism of protein folding. Ruckenstein E; Berim G (2016) Kinetic theory of nucleation. CRC Press, Boca Raton, pp 231–250

Dolgikh DA, Kolomiets AP, Bolotina IA, Ptitsyn OB (1984) “Molten globule” state accumulates in carbonic anhydrase folding. FEBS Lett 164:88–92. https://doi.org/10.1016/0014-5793(84)80020-4

Eichmann C, Preissler S, Riek R, Deuerling E (2010) Cotranslational structure acquisition of nascent polypeptides monitored by NMR spectroscopy. Proc Natl Acad Sci USA 107:9111–9116. https://doi.org/10.1073/pnas.0914300107

Eisenberg DS (2018) How hard it is seeing what is in front of your eyes. Cell 174:8–11. https://doi.org/10.1016/j.cell.2018.06.027

Ellis RJ, Hartl FU (1999) Principles of protein folding in the cellular environment. Curr Opin Struct Biol 9:102–110. https://doi.org/10.1016/s0959-440x(99)80013-x

Emanuel NM, Knorre DG (1984) The course in chemical kinetics, 4th edn. (in Russian), chapters III (§ 2), V (§§ 2, 5). Vysshaja Shkola, Moscow; (1973) Chemical Kinetics. Wiley, New York

Eyring H (1935) The activated complex in chemical reactions. J Chem Phys 3:107–115. https://doi.org/10.1063/1.1749604

Fariselli P, Olmea O, Valencia A, Casadio R (2001) Prediction of contact maps with neural networks and correlated mutations. Protein Eng 14:835–843. https://doi.org/10.1093/protein/14.11.835

Fersht AR (1999) Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding, chapters 2, 15, 18, 19. W. H. Freeman & Co, New York. https://doi.org/10.1142/10574

Fersht AR (2000) 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 97:1525–1529. https://doi.org/10.1073/pnas.97.4.1525

Finkelstein AV (2002) Cunning simplicity of a hierarchical folding. J Biomol Struct Dyn 20:311–313. https://doi.org/10.1080/07391102.2002.10506846

Finkelstein AV (2015) Time to overcome the high, long and bumpy free-energy barrier in a multi-stage process: the generalized steady-state approach. J Phys Chem B 119:158–163. https://doi.org/10.1021/jp5109703

Finkelstein AV (2017) Some additional remarks to the solution of the protein folding puzzle: Reply to comments on "There and back again: two views on the protein folding puzzle." Phys Life Rev 21:77–79. https://doi.org/10.1016/j.plrev.2017.06.025

Finkelstein AV (2018) 50+ years of protein folding. Biochem Mosc 83:S3–S18. https://doi.org/10.1134/S000629791814002X

Finkelstein AV, Garbuzynskiy SO (2015) Reduction of the search space for the folding of proteins at the level of formation and assembly of secondary structures: a new view on solution of Levinthal's paradox. ChemPhysChem 16:3373–3378. https://doi.org/10.1002/cphc.201500700

Finkelstein AV, Garbuzynskiy SO (2016) Solution of Levinthal's paradox is possible at the level of the formation and assembly of protein secondary structures. Biophysics (moscow) 61:1–5. https://doi.org/10.1134/S0006350916010085

Finkelstein AV, Ya Badretdinov A (1997a) Physical reason for fast folding of the stable spatial structure of proteins: a solution of the Levinthal paradox. Mol Biol (moscow) 31:391–398

Finkelstein AV, Ya Badretdinov A (1997b) Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. Fold Des 2:115–121. https://doi.org/10.1016/s1359-0278(97)00016-3

Finkelstein AV, Ya Badretdinov A (1998) Influence of chain knotting on the rate of folding. Addendum to rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. Fold Des 3:67–68. https://doi.org/10.1016/S1359-0278(98)00009-1

Finkelstein AV, Ya Badretdinov A, Ptitsyn OB (1990) Short alpha-helix stability. Nature 345:300–300. https://doi.org/10.1038/345300b0

Finkelstein AV, Ya Badretdinov A, Gutin AM (1995) Why do protein architectures have a Boltzmann-like statistics? Proteins 23:142–150. https://doi.org/10.1002/prot.340230204

Finkelstein AV, Ivankov DN, Garbuzynskiy SO, Galzitskaya OV (2007) Understanding the folding rates and folding nuclei of globular proteins. Curr Prot Pept Sci 8:521–536. https://doi.org/10.2174/138920307783018695

Finkelstein AV, Bogatyreva NS, Garbuzynskiy SO (2013) Restrictions to protein folding determined by the protein size. FEBS Lett 587:1884–1890. https://doi.org/10.1016/j.febslet.2013.04.041

Finkelstein AV, Badretdin AJ, Galzitskaya OV, Ivankov DN, Bogatyreva NS, Garbuzynskiy SO (2017) There and back again: two views on the protein folding puzzle. Phys Life Rev 21:56–71. https://doi.org/10.1016/j.plrev.2017.01.025

Finkelstein AV, Ptitsyn OB (2016) Protein Physics, 2nd edn., lectures 15–23, appendix D. Academic Press, An imprint of Elsevier Science, Amsterdam – Boston – Heidelberg – London – New York – Oxford – Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo. https://doi.org/10.13140/RG.2.1.1319.8320

Finkelstein AV, Ivankov DN, Garbuzynskiy SO, Galzitskaya OV (2014) Understanding the folding rates and folding nuclei of globular proteins. In eBook Series: Dunn BM (ed) Frontiers in protein and peptide sciences. V.1, chapter 5, 91–138. https://doi.org/10.2174/97816080586241140100008

Flanagan JM, Kataoka M, Shortle D, Engelman DM (1992) Truncated staphylococcal nuclease is compact but disordered. Proc Natl Acad Sci USA 89:748–752. https://doi.org/10.1073/pnas.89.2.748

Flory PJ (1969) Statistical mechanics of chain molecules, chapter 3. Interscience Publishers, New York. https://doi.org/10.1002/bip.1969.360080514

Fu B, Wang W (2004) A 20n1-1/d•log(n) time algorithm for d-dimensional protein folding in the HP-model. Lect Notes Comput Sci 3142:630–644. https://doi.org/10.1007/978-3-540-27836-8_54

Fulton KF, Main ERG, Dagett V, Jackson SE (1999) Mapping the interactions present in the transition state for unfolding/folding of FKBP12. J Mol Biol 291:445–461. https://doi.org/10.1006/jmbi.1999.2942

Galzitskaya OV, Finkelstein AV (1995) Folding of chains with random and edited sequences: similarities and differences. Protein Eng 8:883–892. https://doi.org/10.1093/protein/8.9.883

Galzitskaya OV, Finkelstein AV (1999) A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. Proc Natl Acad Sci USA 96:11299–11304. https://doi.org/10.1073/pnas.96.20.11299

Galzitskaya OV, Ivankov DN, Finkelstein AV (2001) Folding nuclei in proteins. FEBS Lett 489:113–118. https://doi.org/10.1016/s0014-5793(01)02092-0

Galzitskaya OV, Garbuzynskiy SO, Ivankov DN, Finkelstein AV (2003) Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. Proteins 51:162–166. https://doi.org/10.1002/prot.10343

Garbuzynskiy SO, Kondratova MS (2008) Structural features of protein folding nuclei. FEBS Lett 582:768–772. https://doi.org/10.1016/j.febslet.2008.01.049

Garbuzynskiy SO, Finkelstein AV, Galzitskaya OV (2004) Outlining folding nuclei in globular proteins. J Mol Biol 336:509–525. https://doi.org/10.1016/j.jmb.2003.12.018

Garbuzynskiy SO, Ivankov DN, Bogatyreva NS, Finkelstein AV (2013) Golden triangle for folding rates of globular proteins. Proc Natl Acad Sci USA 110:147–150. https://doi.org/10.1073/pnas.1210180110

Go N, Abe H (1981) Noninteracting local-structure model of folding and unfolding transition in globular proteins. I Formulation Biopolymers 20:991–1011. https://doi.org/10.1002/bip.1981.360200511

Goldenberg DP, Creighton TE (1983) Circular and circularly permuted forms of bovine pancreatic trypsin inhibitor. J Mol Biol 165:407–413. https://doi.org/10.1016/s0022-2836(83)80265-4

Grantcharova VP, Riddle DS, Santiago JV, Baker D (1998) Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. Nat Struct Biol 5:714–720. https://doi.org/10.1038/1412

Grantcharova V, Alm E, Baker D, Horwich AL (2001) Mechanism of protein folding. Curr Opin Struct Biol 11:70–82. https://doi.org/10.1016/s0959-440x(00)00176-7

Grosberg AY (1997) Disordred polymers. Uspekhi Fiz. Nauk (Moscow, in Russian) 167:129–166. https://doi.org/10.3367/UFNr.0167.199702b.0129

Gutin AM, Shakhnovich EI (1993) Ground state of random copolymers and the discrete random energy model. J Chem Phys 98:8174–8177

Han Y, David A, Liu B, Magadan JG, Bennink JR, Yewdell JW, Qian SB (2012) Monitoring cotranslational protein folding in mammalian cells at codon resolution. Proc Natl Acad Sci USA 109:12467–12472. https://doi.org/10.1073/pnas.1208138109

Holtkamp W, Kokic G, Jäger M, Mittelstaet J, Komar AA, Rodnina MV (2015) Cotranslational protein folding on the ribosome monitored in real time. Science 350:1104–1107. https://doi.org/10.1126/science.aad0344

Isenman DE, Lancet D, Pecht I (1979) Folding pathways of immunoglobulin domains. The folding kinetics of the $C_{\gamma}3$ domain of human IgG1. Biochemistry 18:3327–3336. https://doi.org/10.1021/bi00582a020

Ivankov DN, Finkelstein AV (2004) Prediction of protein folding rates from the amino-acid sequence-predicted secondary structure. Proc Natl Acad Sci USA 101:8942–8944. https://doi.org/10.1073/pnas.0402659101

Ivankov DN, Finkelstein AV (2020) Solution of the Levinthal's paradox and a physical theory of protein folding times. Biomolecules 10(2):E250. https://doi.org/10.3390/biom10020250

Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco K, Baker D, Finkelstein AV (2003) Contact order revisited: influence of protein size on the folding rate. Protein Sci 12:2057–2062. https://doi.org/10.1110/ps.0302503

Jackson SE (1998) How do small single-domain proteins fold? Fold Des 3:R81–R91. https://doi.org/10.1016/S1359-0278(98)00033-9

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202; current version of the program: http://bioinf.cs.ucl.ac.uk/psipred/; https://doi.org/10.1006/jmbi.1999.3091

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589. https://doi.org/10.1038/s41586-021-03819-2

Karplus M (1997) The Levinthal paradox: yesterday and today. Fold Des 2(Suppl. 1):S69–S75. https://doi.org/10.1016/s1359-0278(97)00067-9

Keefe AD, Szostak JW (2001) Functional proteins from a random-sequence library. Nature 410:715–718. https://doi.org/10.1038/35070613

Kiefhaber T (1995) Kinetic traps in lysozyme folding. Proc Natl Acad Sci USA 92:9029–9033. https://doi.org/10.1073/pnas.92.20.9029

Kolb VA, Makeev EV, Spirin AS (1994) Folding of firefly luciferase during translation in a cell-free system. EMBO J 13:3631–3637. https://doi.org/10.1002/j.1460-2075.1994.tb06670.x

Komar AA, Kommer A, Krasheninnikov IA, Spirin AS (1997) Cotranslational folding of globin. J Biol Chem 272:10646–10651. https://doi.org/10.1074/jbc.272.16.10646

Landau LD, Lifshitz EM (1980) Statistical physics, §§ 7, 8, 150. In: A course of theoretical physics, 3rd edn., volume 5, part 1. Elsevier Science, Amsterdam. https://doi.org/10.1016/C2009-0-24487-4

Lappalainen I, Hurley MG, Clarke J (2008) Plasticity within the obligatory folding nucleus of an immunoglobulin-like domain. J Mol Biol 375:547–559. https://doi.org/10.1016/j.jmb.2007.09.088

Leopold PE, Montal M, Onuchic JN (1992) Protein folding funnels: a kinetic approach to the sequence-structure relationship. Proc Natl Acad Sci USA 89:8721–8725. https://doi.org/10.1073/pnas.89.18.8721

Levinthal C (1968) Are there pathways for protein folding? J Chim Phys Chim Biol 65:44–45. https://doi.org/10.1051/jcp/1968650044

Levinthal C (1969) How to fold graciously. In: Debrunner P, Tsibris JCM, Munck E (eds) Mössbauer spectroscopy in biological systems: proceedings of a meeting held at Allerton House, Monticello, Illinois. University of Illinois Press, 22–24.

Levitt M, Chothia C (1976) Structural patterns in globular proteins. Nature 261:552–558. https://doi.org/10.1038/261552a0

Libich DS, Tugarinov V, Clore GM (2015) Intrinsic unfoldase/foldase activity of the chaperonin GroEL directly demonstrated using multinuclear relaxation-based NMR. Proc Natl Acad Sci USA 112:8817–8823. https://doi.org/10.1073/pnas.1510083112

Lim VI (1974a) Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. J Mol Biol 88:857–872. https://doi.org/10.1016/0022-2836(74)90404-5

Lim VI (1974b) Algorithm for prediction of α-helices and β-structural regions in globular proteins. J Mol Biol 88:873–894. https://doi.org/10.1016/0022-2836(74)90405-7

Makarov DE, Plaxco KW (2003) The topomer search model: a simple, quantitative theory of two-state protein folding kinetics. Protein Sci 12:17–26. https://doi.org/10.1110/ps.0220003

Marchenko NY, Garbuzynskiy SO, Semisotnov GV (2009) Molecular chaperones under normal and pathological conditions. In: Zabolotny DI (ed) Molecular pathology of proteins. Nova Science Publishers, New York, pp 57–89

Marchenko NY, Marchenkov VV, Semisotnov GV, Finkelstein AV (2015) Strict experimental evidence that apo-chaperonin GroEL does not accelerate protein folding, although it does accelerate one of its steps. Proc Natl Acad Sci USA 112:E6831-6832. https://doi.org/10.1073/pnas.1517712112

Marchenkov VV, Sokolovskiĭ IV, Kotova NV, Galzitskaya OV, Bochkareva ES, Girshovich AS, Semisotnov GV (2004) The interaction of the GroEL chaperone with early kinetic intermediates of renaturing proteins inhibits the formation of their native structure. Biofizika (moscow, in Russian) 49:987–994

Matouschek A, Kellis JT, Serrano L, Fersht AR (1990) Transient folding intermediates characterized by protein engineering. Nature 346:440–445. https://doi.org/10.1038/346440a0

Mukherjee S, Chowdhury P, Bunagan MR, Gai F (2008) Folding kinetics of a naturally occurring helical peptide: implication of the folding speed limit of helical proteins. J Phys Chem B 112:9146–9150. https://doi.org/10.1021/jp801721p

Muñoz V, Thompson PA, Hofrichter J, Eaton WA (1997) Folding dynamics and mechanism of beta-hairpin formation. Nature 390:196–199. https://doi.org/10.1038/36626

Murzin AG, Finkelstein AV (1988) General architecture of α-helical globule. J Mol Biol 204:749–770. https://doi.org/10.1016/0022-2836(88)90366-x

Neira JL, Fersht AR (1999a) Exploring the folding funnel of a polypeptide chain by biophysical studies on protein fragments. J Mol Biol 285:1309–1333. https://doi.org/10.1006/jmbi.1998.2249

Neira J, Fersht AR (1999b) Acquisition of native-like interactions in C-terminal fragments of barnase. J Mol Biol 287:421–432. https://doi.org/10.1006/jmbi.1999.2602

Ngo JT, Marks J (1992) Computational complexity of a problem in molecular structure prediction. Protein Eng 5:313–321. https://doi.org/10.1093/protein/5.4.313

Nilsson O, Hedman R, Marino J, Wickles S, Bischoff L, Johansson M, Müller-Lucks A, Trovato F, Puglisi JD, O'Brien EP, Beckmann R, Von Heijne G (2015) Cotranslational protein folding inside the ribosome exit tunnel. Cell Rep 12(10):1533–1540. https://doi.org/10.1016/j.celrep.2015.07.065

Nilsson O, Nickson A, Hollins JJ, Wickles S, Steward A, Beckmann R, Von Heijne G, Clarke J (2017) Cotranslational folding of spectrin domains via partially structured states. Nat Struct Mol Biol 24:221–225. https://doi.org/10.1038/nsmb.3355

Nölting B (2010) Protein folding kinetics: biophysical methods, 2nd edn., chapters 10, 11, 12. Springer, Berlin, Heidelberg, New York. https://doi.org/10.1007/b138868

Pauling L (1970) General Chemistry, 3rd edition, chapter 16. W.H. Freeman & Co, San Francisco

Petsko GA, Ringe D (2004) Protein structure and function, chapter 1. New Science Press Ltd, London

Phillips DC (1966) The three-dimensional structure of an enzyme molecule. Sci Am 215:78–90. https://doi.org/10.1038/scientificamerican1166-78

Plaxco KW, Simons KT, Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 277:985–994. https://doi.org/10.1006/jmbi.1998.1645

Privalov PL (1979) Stability of proteins: small globular proteins. Adv Protein Chem 33:167–241. https://doi.org/10.1016/s0065-3233(08)60460-x

Privalov PL, Khechinashvili NN (1974) A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. J Mol Biol 86:665–684. https://doi.org/10.1016/0022-2836(74)90188-0

Ptitsyn OB (1973) Stages in the mechanism of self-organization of protein molecules. Dokl Akad Nauk SSSR (moscow, in Russian) 210:1213–1215

Ptitsyn OB (1995) Molten globule and protein folding. Adv Protein Chem 47:83–229. https://doi.org/10.1016/s0065-3233(08)60546-x

Ptitsyn OB, Finkel'shtein AV (1970) Relation of the secondary structure of globular proteins to their primary structure. Biofizika (moscow, in Russian) 15(5):757–768

Ptitsyn OB, Finkelstein AV (1980) Similarities of protein topologies: evolutionary divergence, functional convergence or principles of folding? Quart Rev Biophys 13:339–386. https://doi.org/10.1017/s0033583500001724

Ptitsyn OB, Finkelstein AV (1983) Theory of protein secondary structure and algorithm of its prediction. Biopolymers 22:15–25. https://doi.org/10.1002/bip.360220105

Robson B, Vaithilingam A (2008) Protein folding revisited. Prog Mol Biol Transl Sci 84:161–202. https://doi.org/10.1016/S0079-6603(08)00405-4

Rollins GC, Dill KA (2014) General mechanism of two-state protein folding kinetics. J Am Chem Soc 136:11420–11427. https://doi.org/10.1021/ja5049434

Roney JP, Ovchinnikov S (2022) State-of-the-art estimation of protein model accuracy using AlphaFold. BioRxiv Preprint. https://doi.org/10.1101/2022.03.11.484043

Ruckenstein E, Berim G (2016) Kinetic theory of nucleation, parts 1, 3. CRC Press, Taylor & Francis Group, Boca Raton. https://doi.org/10.1201/b21644

Šali A, Shakhnovich E, Karplus M (1994) Kinetics of protein folding. A lattice model study of the requirements for folding to the native state. J Mol Biol 235:1614–1636. https://doi.org/10.1006/jmbi.1994.1110

Schulz GE, Barry CD, Friedman J, Chou PY, Fasman GD, Finkelstein AV, Lim VI, Ptitsyn OB, Kabat EA, Wu TT, Levitt M, Robson B, Nagano K (1974) Comparison of predicted and experimentally determined secondary structure of adenyl kinase. Nature 250:140–142. https://doi.org/10.1038/250140a0

Segava S, Sugihara M (1984) Characterization of the transition state of lysozyme unfolding. I. Effect of protein-solvent interactions on the transition state. Biochemistry 23:2473–2488. https://doi.org/10.1002/bip.360231122

Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D (2019) Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). Proteins 87:1141–1148. https://doi.org/10.1002/prot.25834

Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D (2020) Improved protein structure prediction using potentials from deep learning. Nature 577:706–710. https://doi.org/10.1038/s41586-019-1923-7

Shakhnovich EI (2006) Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. Chem Rev 106:1559–1588. https://doi.org/10.1021/cr040425u

Shakhnovich EI, Finkelstein AV (1989) Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is the first order phase transition. Biopolymers 28:1667–1680. https://doi.org/10.1002/bip.360281003

Shakhnovich EI, Gutin AM (1990) Implications of thermodynamics of protein folding for evolution of primary sequences. Nature 346:773–775. https://doi.org/10.1038/346773a0

Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shah Y, Wriggers W (2010) Atom-level characterization of structural dynamics of proteins. Science 330:341–346. https://doi.org/10.1126/science.1187409

Sorokina I, Mushegian AR, Koonin EV (2022) Is protein folding a thermodynamically unfavorable, active, energy-dependent process? Int J Mol Sci 23:521. https://doi.org/10.3390/ijms23010521

Steinhofel K, Skaliotis A, Albrecht AA (2006) Landscape analysis for protein folding simulation in the H-P model. Lect Notes Comput Sci 4175:252–261. https://doi.org/10.1007/11851561_24

Tanford C (1968) Protein denaturation. Adv Protein Chem 23:121–282. https://doi.org/10.1016/s0065-3233(08)60401-5

The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 49(D1):D480–D489. https://doi.org/10.1093/nar/gkaa1100

Thirumalai D (1995) From minimal models to real proteins: time scales for protein folding kinetics. J Phys I (orsay, Fr) 5:1457–1469. https://hal.archives-ouvertes.fr/jpa-00247149

Thirumalai D, Lorimer GH, Hyeon C (2020) Iterative annealing mechanism explains the functions of the GroEL and RNA chaperones. Protein Sci 29:360–377. https://doi.org/10.1002/pro.3795

Tian P, Steward A, Kudva R, Su T, Shilling PJ, Nickson AA, Hollins JJ, Beckmann R, Von Heijne G, Clarke J, Best RB (2018) Folding pathway of an Ig domain is conserved on and off the ribosome. Proc Natl Acad Sci USA 115(48):E11284–E11293. https://doi.org/10.1073/pnas.1810523115

To P, Whitehead B, Tarbox HE, Fried SD (2021) Nonrefoldability is pervasive across the E coli proteome. J Am Chem Soc 143(30):11435–11448. https://doi.org/10.1021/jacs.1c03270

Tompa P (2005) The interplay between structure and function in intrinsically unstructured proteins. FEBS Lett 579:3346–3354. https://doi.org/10.1016/j.febslet.2005.03.072

Tsutsui Y, Cruz RD, Wintrode PL (2012) Folding mechanism of the metastable serpin α1-antitrypsin. Proc Natl Acad Sci USA 109:4467–4472. https://doi.org/10.1073/pnas.1109125109

Unger R, Moult J (1993) Finding the lowest free energy conformation of a protein is an NP-hard problem: proof and implications. Bull Math Biol 55:1183–1198. https://doi.org/10.1007/BF02460703

Uversky VN (2002) Natively unfolded proteins: a point where biology waits for physics. Protein Sci 11:739–756. https://doi.org/10.1110/ps.4210102

Uversky VN, Finkelstein AV (2019) Life in phases: intra- and intermolecular phase transitions in protein solutions. Biomolecules 9:E842. https://doi.org/10.3390/biom9120842

Wallin S, Chan HS (2005) A critical assessment of the topomer search model of protein folding using a continuum explicit-chain model with extensive conformational sampling. Protein Sci 14:1643–1660. https://doi.org/10.1110/ps.041317705

Wang J, Oliveira RJ, Chu X, Whitford PC, Chahine J, Han W, Wang E, Onuchic JN, Leite VB (2012) Topography of funneled landscapes determines the thermodynamics and kinetics of protein folding. Proc Natl Acad Sci USA 109:15763–15768. https://doi.org/10.1073/pnas.1212842109

Wensley BG, Gärtner M, Choo WX, Batey S, Clarke J (2009) Different members of a simple three-helix bundle protein family have very different folding rate constants and fold by different mechanisms. J Mol Biol 390:1074–1085. https://doi.org/10.1016/j.jmb.2009.05.010

Wetlaufer DB (1973) Nucleation, rapid folding, and globular intrachain regions in proteins. Proc Natl Acad Sci USA 70:697–701. https://doi.org/10.1073/pnas.70.3.697

Wolynes PG (1997) Folding funnels and energy landscapes of larger proteins within the capillarity approximation. Proc Natl Acad Sci USA 94:6170–6175. https://doi.org/10.1073/pnas.94.12.6170

Wolynes PG (2015) Evolution, energy landscapes and the paradoxes of protein folding. Biochimie 119:218–230. https://doi.org/10.1016/j.biochi.2014.12.007

Wolynes PG, Onuchic JN, Thirumalai D (1995) Navigating the folding routes. Science 267:1619–1620. https://doi.org/10.1126/science.7886447

Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 293:321–331. https://doi.org/10.1006/jmbi.1999.3110

Wruck F, Tian P, Kudva R, Best RB, Von Heijne G, Tans SJ, Katranidis A (2021) The ribosome modulates folding inside the ribosomal exit tunnel. Commun Biol 4(523):1–8. https://doi.org/10.1038/s42003-021-02055-8

Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D (2020) Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci USA 117:1496–1503. https://doi.org/10.1073/pnas.1914677117

Zana R (1975) On the rate determining step for helix propagation in the helix-coil transition of polypeptides in solution. Biopolymers 14:2425–2428

Zwanzig R, Szabo A, Bagchi B (1992) Levinthal's paradox. Proc Natl Acad Sci USA 89:20–22. https://doi.org/10.1073/pnas.89.1.20

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.