Available online at www.sciencedirect.com

ScienceDirect

Physics of Life Reviews 21 (2017) 56–71

PHYSICS of LIFE reviews

www.elsevier.com/locate/plrev

Review

There and back again: Two views on the protein folding puzzle

Alexei V. Finkelstein [a,*], Azat J. Badretdin [b], Oxana V. Galzitskaya [a], Dmitry N. Ivankov [a,c,d], Natalya S. Bogatyreva [a,c,d], Sergiy O. Garbuzynskiy [a]

[a] *Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region 142290, Russian Federation*
[b] *National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA*
[c] *Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, 08003 Barcelona, Spain*
[d] *Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain*

Received 5 September 2016; received in revised form 5 January 2017; accepted 19 January 2017
Available online 27 January 2017
Communicated by E. Shakhnovich

Abstract

The ability of protein chains to spontaneously form their spatial structures is a long-standing puzzle in molecular biology. Experimentally measured folding times of single-domain globular proteins range from microseconds to hours: the difference (10–11 orders of magnitude) is the same as that between the life span of a mosquito and the age of the universe. This review describes physical theories of rates of overcoming the free-energy barrier separating the natively folded (N) and unfolded (U) states of protein chains in both directions: "U-to-N" and "N-to-U". In the theory of protein folding rates a special role is played by the point of thermodynamic (and kinetic) equilibrium between the native and unfolded state of the chain; here, the theory obtains the simplest form. Paradoxically, a theoretical estimate of the folding time is easier to get from consideration of protein unfolding (the "N-to-U" transition) rather than folding, because it is easier to outline a good unfolding pathway of any structure than a good folding pathway that leads to the stable fold, which is yet unknown to the folding protein chain. And since the rates of direct and reverse reactions are equal at the equilibrium point (as follows from the physical "detailed balance" principle), the estimated folding time can be derived from the estimated unfolding time. Theoretical analysis of the "N-to-U" transition outlines the range of protein folding rates in a good agreement with experiment. Theoretical analysis of folding (the "U-to-N" transition), performed at the level of formation and assembly of protein secondary structures, outlines the upper limit of protein folding times (i.e., of the time of search for the most stable fold). Both theories come to essentially the same results; this is not a surprise, because they describe overcoming one and the same free-energy barrier, although the way to the top of this barrier from the side of the unfolded state is very different from the way from the side of the native state; and both theories agree with experiment. In addition, they predict the maximal size of protein domains that fold under solely thermodynamic (rather than kinetic) control and explain the observed maximal size of the "foldable" protein domains.

Keywords: Protein folding; Levinthal's paradox; Folding funnel; Phase separation; Free-energy landscape; Protein secondary structure assembly

[*] Corresponding author.
E-mail address: afinkel@vega.protres.ru (A.V. Finkelstein).

http://dx.doi.org/10.1016/j.plrev.2017.01.025
1571-0645/© 2017 Elsevier B.V. All rights reserved.

1. Introduction

The ability of proteins to fold spontaneously puzzled protein science for a long time (see, e.g., [16,20,50,55,81,98,102]). Our previous review published in PLREV [26] encompassed fundamental experimental facts forming a physical basis of this process and protein physics in general. An updated and extended overview of these facts one can find in a book [30].

It is well known that a protein chain (actually, the chain of a globular protein) can spontaneously fold into its unique native 3D structure [2,3]. In doing so, the protein chain has to find its native (and seemingly the most stable) fold among zillions of others within only minutes or seconds given for its folding.

Indeed, the number of alternatives is vast [62,63]: it is at least 2^{100} but may be 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 [76] (or even 100 [63]) conformations are possible for each 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 $\sim 2^{100}$ picoseconds (or 3^{100} or even 10^{100} or 100^{100}), that is, $\sim 10^{10}$ (or 10^{25} or even 10^{80} or 10^{180}) years. And it looks like the sampling has to be really exhaustive, because the protein can “feel” that it has come to the stable structure only when it hits it precisely, while even a 1 Å deviation can strongly increase the chain energy in the closely packed globule.

Then, how does the protein choose its native structure among zillions of possible others, asked Levinthal [62,63] (who first noticed this paradox), and answered: It seems that the protein folding follows some specific pathway, and the native fold is simply the end of this pathway, no matter if it is the most stable chain fold or not. 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, computer experiments with lattice models of protein chains strongly suggest that the chains fold to their stable structure, i.e., that the “native protein structure” is the lowest-energy one, and protein folding is under thermodynamic rather than kinetic control [1,83].

Nevertheless, most of hypotheses on protein folding are based on the “kinetic control assumption”.

Ahead of Levinthal, Phillips [73] proposed that the protein folding nucleus is formed near the N-end of the nascent protein chain, and the remaining chain wraps around it. However, successful *in vitro* folding of many single-domain proteins and protein domains does not begin from the N-end [48,49,60].

Wetlaufer [100] hypothesized formation of the folding nucleus by adjacent residues of the protein chain. However, *in vitro* experiments show that this is not always so [38,99].

Ptitsyn [77] proposed a model of hierarchical folding, i.e., a stepwise involvement of different interactions and formation of different folding intermediate states.

More recently, various “folding funnel” models [4,15,61,98,103] have become popular for illustrating and describing fast folding processes.

The difficulty of the “kinetics vs. stability” problem is that it hardly can be solved by direct experiment. Indeed, suppose that a protein has some structure that is more stable than the native one. How can we find it if the protein does not do so itself? Shall we wait for $\sim 10^{10}$ (or even $\sim 10^{180}$) years?

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

However, is there a real 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?

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 will consider single-domain proteins only, i.e., chains of ~ 100 residues). These facts will help us to understand what chains and what folding conditions we have to consider. The facts are as follows:

1. The denatured state of proteins, at least that of small proteins treated with a strong denaturant, is often the unfolded random coil [93].

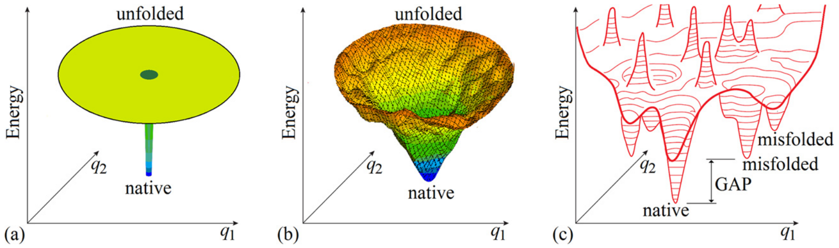

Fig. 1. Schematic illustration of basic models of the energy landscape of protein chains. (a) The "golf course" model of the protein potential energy landscape. (b) The "funnel" model of the protein potential energy landscape. The funnel is centered in the lowest-energy ("native") structure. (c) In more detail: the bumpy potential energy landscape of a protein chain. A wide (of many $k_B T_{melt}$, where k_B is Boltzmann's constant and T_{melt} is protein melting temperature) 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. 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.

2. Protein unfolding is reversible [3]; moreover, the denatured and native states of a protein can be in a kinetic equilibrium [12]; and there is an "all-or-none" transition between them [76]. The latter means that only two states of the protein molecule, native and denatured, are present (close to the mid-point of the folding–unfolding equilibrium) in a visible quantity, while all others, "semi-native" or misfolded, 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) the physical theory shows that such a transition requires the amino acid sequence that provides a large "energy gap" between the most stable structure and the bulk of misfolded ones [39,51,83,87,88].)
3. Even under normal physiological conditions the native (i.e., the lowest-energy) state of a protein is only more stable than its unfolded (i.e., the highest-entropy) state by a few kilocalories per mole [76] (and these two states have equal stability at mid-transition, naturally).

(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 means only 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 [93], come within "energy". This terminology is commonly used to concentrate on the main problem of sampling the protein chain conformations.)

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

However, to begin with, it is not out of place considering whether the "Levinthal's paradox" is a paradox indeed. Bryngelson and Wolynes [7] mentioned that this "paradox" is based on the absolutely flat (and therefore unrealistic) "golf course" model of the protein potential energy surface (Fig. 1a), and somewhat later Leopold et al. [61], following the line of Go and Abe [47], considered more realistic (tilted and biased to the protein's native structure) energy surfaces and introduced the "folding funnels" (Fig. 1b), which seemingly eliminate the "paradox" at all.

It's not as simple as that, though…

The problem of huge sampling does exist even for realistic energy surfaces. It has been mathematically proven that, despite the folding funnels and all that, finding the lowest free-energy conformation of a protein chain is the so-called "NP-hard" problem [70,96], which, loosely speaking, requires an exponentially large time to be solved (by a folding chain or by a man).

Anyhow, various "folding funnel" models became popular for explaining and illustrating protein folding [58,71,102,103]. In the funnel, the lowest-energy structure (formed, thus, by a set of most powerful interactions) is the center surrounded by higher-energy structures containing only a part of these interactions. The "energy funnels" are not perfectly smooth due to some "frustrations", 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 have enhanced stability) [6–8,31]. Anyhow, the "energy funnel" can direct movement towards the lowest-energy structure, which seems to help the protein chains to avoid the "Levinthal's" sampling of all conformations.

However, it can be shown that the energy funnels *per se* do not solve the Levinthal's paradox. Strict analysis [5] of the straightforwardly presented funnel models [4,105] shows that close to the mid-point of the folding–unfolding equilibrium they cannot *simultaneously* explain the both major features observed in protein folding: (i) its non-astronomical time, and (ii) the "all-or-none" transition, i.e., co-existence of native and unfolded protein molecules during the folding process. The latter requires a "volcano-shaped" free-energy folding landscape (see Fig. 5 below), the uphill rim of which creates an enormously high free-energy barrier at the folding pathway in the case of non-nucleated structuration assumed by [4,105] (and earlier considered by [85]; see discussion of the free-energy landscapes below).

By the way, the stepwise mechanism of protein folding [77], taken *per se*, also cannot [21] *simultaneously* explain these two major features observed in protein folding. The key folding-accelerating feature of the "stepwise mechanism" is that the most stable structures formed at the first its step serve as building blocks for the next step of folding, and then the most stable structures obtained at this second step serve as the building blocks for the next step, etc. In principle, this can help to avoid sampling of all the huge conformational space. But such a mechanism implies that the once found structures preserve their form (and do not decay back) until the next step, which means that they must be thermodynamically more stable than their more disordered precursors. The structures formed at the next step also must be thermodynamically more stable than their precursors, etc. Thus, such a mechanism *can* work *only* when the native structure is much more stable than the disordered one, and it *cannot* work when protein folding occurs near the point of thermodynamic equilibrium between the native and disordered states of the protein.

Thus, neither stepwise nor simple funnel mechanisms solve the Levinthal's problem, although they give a hint as to what accelerates protein folding.

The basic solution of the paradox is provided by special nucleation funnels [23,24] considering the separation of the unfolded and native phases within the folding chain (called the "capillarity theory" [101]).

It will be described in the next part of this review.

2. Physical estimate of the height of free-energy barrier between the folded and unfolded states: view at the barrier from the side of the folded state

To solve the "Levinthal's paradox" and to show that the most stable chain fold can be found within a reasonable time, we could, to a first approximation, consider only the rate of the "all-or-none" transition between the coil and the most stable structure. And we may consider this transition only for the crucial case when the most stable fold is as stable as (or only a little more stable than) the coil, with all other states of the chain being unstable, i.e., close to the "all-or-none" transition midpoint. Here the analysis can be made in the simplest form, without accounting for accumulating intermediates. True, the maximum folding rate is achieved when the native fold is considerably more stable than the coil [12,19], and then observable intermediates often arise; but let us consider not the fastest but the simplest case...

Since the "all-or-none" transition requires a large energy gap between the most stable structure and misfolded ones (Fig. 1c), we will assume that the considered amino acid sequence provides such a gap. Our aim is to estimate the rate of the "all-or-none" transition and to prove (if possible) that the most stable structure of a normal size domain (∼100 residues) can fold within minutes or seconds (and sometimes even much faster).

To prove that the most stable chain structure is capable of rapid folding, it is sufficient to prove that *at least one* rapid folding pathway (i.e., passing the low-free-energy barrier) leads to this structure. Additional pathways can only accelerate the folding since the rates of parallel reactions are additive. And we can avoid considering folding of other, non-native structures. They have high energy because of the "energy gap", and, near the point of the "all-or-none" transition between the most stable globule and the unfolded chain, they are unstable even taken together, and therefore, they cannot serve as "folding traps" that absorb folding chains. (One can imagine water leaking from a full pool to an empty one through cracks in the wall between them: when the cracks cannot absorb all the water, each additional crack accelerates filling of the empty pool.)

To be rapid, the pathway must consist of not too many steps, and most importantly, it must not require overcoming of a too high free energy barrier.

An L-residue chain can, in principle, attain its lowest-energy fold in L steps, each adding one fixed residue to the growing structure (Fig. 2). *If* the free energy went downhill along the entire pathway, a 100-residue chain would

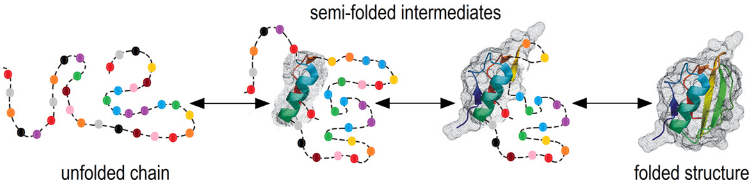

Fig. 2. A scheme [24] of a sequential folding pathway of some globular structure (it is the sequential unfolding pathway of this structure passed in the opposite direction). At each step of sequential folding one residue leaves the coil and takes its final position in the structure. The folded part (shaded) of semi-folded intermediates on the optimal (low-free-energy) pathway is compact (having a small boundary between the folded and unfolded phases). The bold lines and strips show the backbone fixed in the already folded part; fixed side chains are not shown for the sake of simplicity (the volume that they occupy is shaded). The broken line shows the yet unfolded chain.

fold in ∼100–1000 ns, since the growth of a structure (e.g., an α-helix) by one residue is known to take a few nanoseconds [104].

Protein folding takes minutes or seconds or even milliseconds rather than a fraction of a microsecond because of the free energy barrier: most of the folding time is spent on climbing up this barrier and falling back, rather than on moving along the folding pathway.

The key role in this process is played by the transition state [20], i.e., the least stable ("barrier") state on the reaction pathway. According to the conventional transition state theory [17,18,75], the time of the multi-step process of overcoming the barrier is estimated as

$$TIME \sim \tau \times \exp\left(+\Delta F^{\#}/RT\right) \tag{1}$$

where τ is the time of one elementary step, and $\Delta F^{\#}$ is the height of the free energy barrier.

As for $\Delta F^{\#}$, this is our main question: how high is the free energy barrier $F^{\#}$ on the pathway leading to the lowest-energy structure? Formation of this structure decreases both the chain entropy (because of an increase in the chain's ordering) and its energy (due to formation of contacts stabilizing the lowest-energy fold). The former increases and the latter decreases free energy of the chain.

If fold-stabilizing contacts start to arise only when the chain comes very close to its final structure (i.e., if the chain has to lose almost all its entropy *before* the energy starts to decrease), the initial free energy increase would form a very high free energy barrier (proportional to the *total* chain entropy loss). The Levinthal's paradox claiming that the lowest-energy fold cannot be found within any reasonable time since this involves exhaustive sampling of all chain conformations originates exactly from this picture (loss of the entire entropy *before* the energy gain).

However, this paradox can be avoided if there is a folding pathway where the entropy decrease is immediately or nearly immediately compensated for by the energy decrease [47].

Let us consider a *sequential* [100] folding pathway (Fig. 2). More specifically, we will consider a process at each step of which one residue leaves the coil and takes its final position in the lowest-energy 3D structure. True, this pathway may look a bit artificial, but actually the outlined pathway is exactly the pathway that one expects to see watching the movie on unfolding, but in the opposite direction.

According to the well-known in physics *detailed balance* law [59], the direct and reverse reactions follow the same pathway and have equal rates when the both end-states have equal stability. (This law follows from the second law of thermodynamics. It proved by contradiction: if, in thermodynamic equilibrium ambient conditions, the pathway $A \rightarrow 1 \rightarrow B$ is faster than $A \rightarrow 2 \rightarrow B$ for the $A \rightarrow B$ reaction, while the pathway $A \leftarrow 2 \leftarrow B$ is faster than $A \leftarrow 1 \leftarrow B$ for the reciprocal $A \leftarrow B$ reaction under the same conditions, one obtains a *permanent* flow $A \begin{array}{c} \rightarrow 1 \rightarrow \\ \leftarrow 2 \leftarrow \end{array} B$, which contradicts to the second law of thermodynamics.)

Thus, one can use the detailed balance law to find the transition state for folding by finding the optimal transition state for *un*folding! An advantage of analysis of the unfolding pathway is that it is much easier: for any final globular structure, one can easily figure out its sequential unfolding passing through the least unstable semi-unfolded states, i.e., those where the compact globular phase is separated from the unfolded one (Fig. 2) [23,24,40,46].

(In this connection, it is not out of place mentioning that, odd enough, protein unfolding, in contrast to folding, has never been treated as a "puzzle", although it is well known for a long time that these two states, unfolded and folded, can be in kinetic equilibrium! Despite all that, nobody asked a question complementary to Levithal's, that is, how the

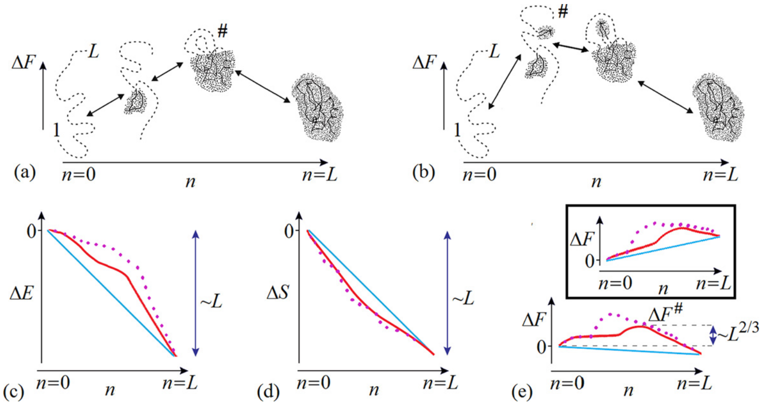

Fig. 3. Schematic illustration of sequential folding/unfolding with compact (a) and non-compact (b) semi-folded intermediates and the change of energy (c), entropy (d) and free energy (e) along these sequential folding/unfolding pathways close to the point of thermodynamic equilibrium between the coil ($n = 0$) and the final structure ($n = L$: all the L chain residues are folded). The full energy and entropy changes, $\Delta E(L)$ and $\Delta S(L)$, are approximately proportional to L. The straight blue lines show the linear (proportional to the number of already folded residues n) parts of $\Delta E(n)$ and $\Delta S(n)$. The non-linear parts of $\Delta E(n)$ and $\Delta S(n)$ result mainly from the surface of the folded part of the molecule (solid lines: for a pathway with compact intermediate structures; dotted lines: for that with non-compact intermediates). The maximal deviations of the $\Delta E(n)$ and $\Delta S(n)$ values from linear dependences are proportional to only $L^{2/3}$. As a result, $\Delta F(n) = \Delta E(n) - T\Delta S(n)$ also deviates from the linear dependence (straight blue line) by a value of only $\sim L^{2/3}$ for compact intermediate structures (while for non-compact intermediates, the deviations are greater). Thus, at the equilibrium point (where $\Delta F(0) = \Delta F(L)$), the maximal on this pathway free energy excess $\Delta F^{\#}$ over the blue free energy baseline (the barrier) is also proportional to only $L^{2/3}$ for compact intermediate structures. The change $\Delta F(n)$ on the pathway to other structures looks similar (see Inset in panel (e)), but these pathways can be neglected, because all these structures are unstable with $\Delta F(n = 0) < \Delta F(L)$ in the presence of the energy gap and the "all-or-none" transition between the unfolded and the most stable globular state of the chain. Adapted from [23,24]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

protein gains a huge energy required for unfolding... This shows that it is easier to imagine how to unfold any protein structure then how to fold it.)

Thus, let us consider the energy change ΔE, the entropy change ΔS and the resultant free energy change $\Delta F = \Delta E - T\Delta S$ along the *sequential* (Fig. 2) folding pathway (reconstructed from the way of sequential *un*folding).

When a piece of the final globule grows sequentially, the interactions that stabilize its final fold are restored sequentially as well. If the folded piece remains compact, as in Figs. 2, 3a, the number of restored interactions grows (and their total energy decreases, see Fig. 3c) approximately in proportion to the number n of residues that have taken their final positions.

Approximately in proportion – but with one significant deviation: At the beginning of folding, the energy decrease is a little slower, since the contact of a newly joined residue with the surface of a small globule is, on average, smaller than its contact with the surface of a large globule. This results in a non-linear *surface* term (the surface being proportional to $\approx n^{2/3}$) in the energy ΔE of the growing globule. Thus, the maximal deviation from the linear energy decrease is proportional to $L^{2/3}$, while the total energy decrease is proportional to the total number L of residues. The deviation is still greater, see Fig. 3c, if the folded parts do not form a compact piece, as in Fig. 3b.

The entropy decrease is also *approximately* proportional to the number n of residues that have taken their final positions (Fig. 3d). At the beginning of folding, though, the entropy decrease can be a little faster owing to disordered but closed loops protruding from the growing globule (Figs. 2 and 4). The maximal number of such loops is proportional to the interface between the folded and unfolded phases, and the free energy of a loop is known [36,56] to have a very slow, logarithmic dependence on its length. This again results in a non-linear *surface* term in the entropy ΔS of the growing globule. The overall entropy decrease is proportional to L again, and the maximal deviation from the linear entropy decrease again is proportional to $L^{2/3}$ (actually, it is proportional to $\sim L^{2/3} \times \ln(L^{1/3})$ at the most, but the multiplier $\ln(L^{1/3})$ is insignificant, about 1–2 when L is 10–1000) [23]; see also the later rigorous mathematical papers [37,91].

Here, it is not out of place mentioning that a separation of the folded and unfolded phases in the transition state of protein folding is very clearly seen in computer simulations of protein folding (see, e.g., [89]).

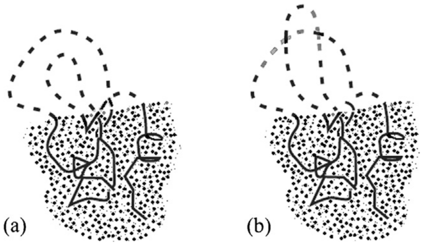

Fig. 4. (a) A compact semi-folded intermediate with protruding unfolded loops. Its growth corresponds to a shift of the boundary between the folded (globular) and unfolded parts. Successful folding requires correct knotting of loops: the structure with incorrect knotting (b) cannot change directly to the correct final structure: first it has to unfold and achieve the correct knotting. However, since a chain of ∼100 residues can only form one or two knots, the search for correct knotting can only slow down the folding two-fold or at most four-fold; thus, the search for correct chain knotting does not limit the folding rate of normal size protein chains. Adapted from [25].

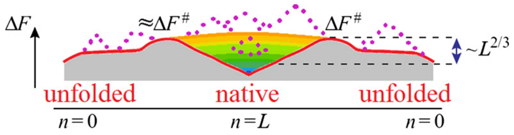

Fig. 5. This purely illustrative drawing shows how entropy converts the *energy* funnel (illustrated in Fig. 1b) into a “volcano-shaped” (as it is called now [82]) *free-energy* folding landscape with free-energy barriers (Fig. 3e) on each pathway leading from an unfolded conformation to the native fold. Any pathway from the unfolded state to the native one first goes uphill, and only then, from the barrier (i.e., crater edge), descends into the “free-energy funnel”. The smooth free energy landscape corresponds to compact semi-folded intermediate structures (shown in Fig. 3a), the rocks (denoted by dotted lines) present a landscape including non-compact semi-folded intermediate structures (shown in Fig. 3b). More accurate but less beautiful scheme of a free-energy landscape is shown in Fig. 2 in [40].

Both linear and surface constituents of ΔS and ΔE enter the free energy $\Delta F = \Delta E - T\Delta S$ of the growing (or unfolding) globule. However, when the final globule is in thermodynamic equilibrium with the coil, the large linear terms *annihilate* each other in the difference $\Delta E - T\Delta S$ (since $\Delta F = 0$ both in the coil (i.e., at $n = 0$) and in the final globule (at $n = L$)), and only the surface terms remain: $\Delta F(n)$ would be *zero* all along the pathway in the absence of surface terms.

Thus, the free energy barrier (Figs. 3e, 5) on a sequential folding pathway with compact semi-folded structures depends only on relatively small globule surface effects, and its height is proportional *not to* L (as Levinthal's estimate implies), but to $L^{2/3}$ only.

In the most simplified form (for details, see [23–25,46]), free energy of the barrier is estimated as follows.

The fastest folding pathway is that having the lowest free energy barrier; the barrier, on a given pathway, corresponds to the intermediate with the highest free energy, that is, the maximal for this pathway interface between the folded and unfolded phases; this interface contains about $L^{2/3}$ residues.

The energy constituent $\Delta E^{\#}$ of the barrier free energy results from interactions lost by the interface residues; it is about

$$L^{2/3} \cdot \varepsilon {}^{1}\!/_{4} \tag{2}$$

where $\varepsilon \approx 1.3$ kcal/mol $\approx 2k_{\rm B}T_{\rm mel}$ is the average heat of protein melting per residue [76] (this is the first empirical parameter used by the theory), and $\approx {}^{1}\!/_{4}$ is the fraction of interactions lost by an interface residue. Thus,

$$\Delta E^{\#/}/k_{\rm B}T_{\rm mel} \approx 0.5L^{2/3} \tag{2a}$$

The entropy constituent $\Delta S^{\#}$ of the barrier free energy is caused by entropy loss in closed loops protruding from the globular into the unfolded phase (see Fig. 4).

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

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

$$(\Delta S^{\#})_{\rm lower} = {}^{1}\!/_{6}L^{2/3} \cdot \left[-{}^{5}\!/_{2}k_{\rm B}\ln\left(3L^{1/3}\right)\right], \tag{3}$$

where ${}^{1}\!/_{6}L^{2/3}$ corresponds to the maximal number of closed loops protruding from the optimal (minimally covered by loops) globule/coil interface (actually, this is the average number for one globule cross section (Fig. 4), since the

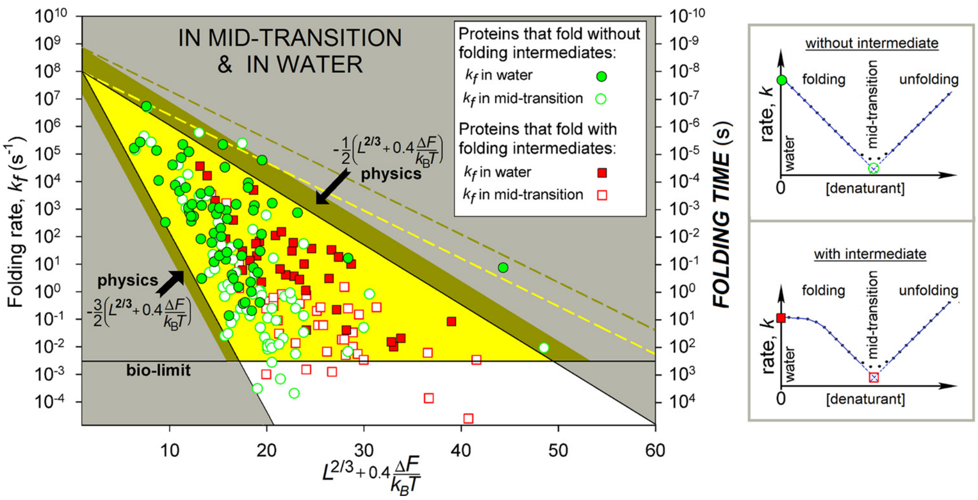

Fig. 6. Main panel: Experimentally measured *in vitro* folding rate constants in water (under approximately “biological” conditions) and at mid-transition for 107 single-domain proteins (or separate domains) without SS bonds and covalently bound ligands (though the rates for proteins with and without SS bonds are principally the same [43]). Triangle: the region allowed by physics; its golden part (with the bronze belt) corresponds to biologically-reasonable folding times ($\leq$10 min); the larger folding times (i.e., the smaller folding rates) are observed (for some proteins) only under mid-transition, i.e., non-biological conditions. Yellow dashed line limits the area allowed only for oblate (1:2) and oblong (2:1) globules at mid-transition; bronze dashed line means the same for “biologically normal” conditions. L is the number of amino acid residues in the protein chain under study. ΔF is the free energy difference between the native and unfolded states of the chain. Adapted from [46]. Supplementary panels: Typical forms of “chevron plots” for the folding/unfolding kinetics of proteins that fold without and with folding intermediates (after [19]).

interface residue can have 6 directions – 4 along the surface, 1 inside, and only 1 outside; and the folding-involved interface must be covered by a minimal, never exceeding the average, number of loops). $3L^{1/3} \equiv (L/2)/({}^{1}\!/_{6}L^{2/3})$ is the average number of residues in such a loop (equal to the number of unfolded residues divided by the number of loops), and $-{}^{5}\!/_{2}k_\mathrm{B}\ln(3L^{1/3})$ is entropy lost by such a closed loop (the interior parts of which do not penetrate inside the globule; this changes the conventional Flory's coefficient, ${}^{3}\!/_{2}$, to ${}^{5}\!/_{2}$). Having $L \sim 100$ (actually, this approximation is good for the whole range of $L = 10$–1000), we obtain

$$\left(\Delta S^{\#}\right)_{\mathrm{lower}} \approx -k_\mathrm{B}L^{2/3} \tag{3a}$$

As a result, the time of both folding and unfolding of the most stable chain structure grows with the number of chain residues L *not* “according to Levinthal” (i.e., *not* as 2^L, or 10^L, or any exponent of L), but, in mid-transition conditions, as

$$TIME \sim \tau \times \exp\left[(1 \pm 0.5)L^{2/3}\right] \tag{4}$$

where $\tau \approx 10$ ns [104] (this is the second and the last empirical parameter used in the theory).

The physical reason for this “non-Levinthal” estimate is that (i) during folding, the entropy decrease is almost immediately and almost completely compensated for by an energy decrease along the sequential folding pathway (and, likewise, the energy increase is almost immediately and almost completely compensated for by an entropy increase along the same sequential *un*folding pathway), and (ii) the free energy results only from surface effects which are relatively weak.

The observed protein folding times span (Fig. 6) 11 orders of magnitude (which is akin to the difference between the life span of a mosquito and the age of the universe). The range of folding times at mid-transition (where $\Delta F = 0$) is from 10 ns $\times \exp(0.5L^{2/3})$ to 10 ns $\times \exp(1.5L^{2/3})$, in accordance with the estimate obtained. Under more physiological conditions (“in water”, where $\Delta F < 0$), $L^{2/3}$ is replaced by $L^{2/3} + 0.4\Delta F/RT$ (see Discussion), but in all other respects the range remains the same.

It is noteworthy that the outlined sequential folding pathways do not require any rearrangement of the dense globular part (which could take a lot of time): all rearrangements occur in the coil.

Anyhow, the obtained eq. (4) illustrated in Fig. 6 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 [12,19] to be the slowest. Native structures of such relatively small proteins are under thermodynamic control: they are the most stable among all structures of these chains. Native structures of larger proteins (of ≈90–400 residues) are, in addition, under a "structural control", in a sense that too entangled folds of their long chains cannot be achieved within days or weeks even if they are thermodynamically stable; and indeed, greatly entangled folds of long protein chains have been never observed [46]: they seem to be excluded from the repertoire of existing protein structures. This also explains why larger proteins should be far from spherical or consist (according to the "divide and rule" principle) of separately folding domains: otherwise, chains of more than 400 residues would fold too slowly. This is a "structural control" again. Its effect, in some sense, resembles that of Levinthal's "kinetic control", though at another level and only for large proteins. The above estimates (80–90 and ≈400 residues) are somewhat elevated when the native fold free energy ΔF is lower than that of the unfolded chain (see below), but essentially they remain the same [46].

One thing is left to be said here:

The "quasi-Levinthal" search over intermediates with different chain knotting (Fig. 4) can, in principle, be a rate-limiting factor, since knotting cannot be changed without a decay of the globular part. However, since the computer experiments show that one knot involves about a hundred residues, the search for correct knotting can only be rate-limiting for extremely long chains [25] which cannot fold within a reasonable time (according to eq. (4)) in any case.

It should be added that above we focused on stability (or rather, instability) of transition states and paid virtually no attention to folding intermediates, because they – in a contrast to transition states – do not determine the rate of folding of native structures [19,20]. We also did not pay attention to structures of folding nuclei, being interested in their size (and, the main, their instability) only. However, there is ample evidence that transition states are well-organized and possess specific structural features in some cases (see [19,20,45,89]), and are poorly organized ("diffused nuclei") in the others (see [34,35,50] and literature therein). The latter, together with the observed sensitivity of positions and shapes of the folding nuclei to mutations, led to a conclusion that a "nucleus" is an ensemble of structures rather than a single structure, 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.

3. Estimating dependence of the sampling volume on protein size: view at the barrier from the side of unfolded state

The above given estimate of the folding time is based on consideration of protein *un*_folding_ rather than *folding*. We have considered *unfolding* because it is easier to outline a good *un*folding pathway of any structure than a good folding pathway leading to the lowest-energy fold, while the free energy barrier at both pathways is the same. In other words, we considered the free energy barrier between the unfolded and folded states (Figs. 4, 5) with the focus on its *un*folding side (connected with energy increase on the pathway from the volcano throat to the crater edge) and did not consider its folding side (connected with entropy loss on the pathway from the unfolded state to the crater edge). Since the rates of direct and reverse reactions are equal under mid-transition conditions (as follows from the physical "detailed balance" principle), here the "*un*folding" and "folding" sides of the barrier are of equal heights, and therefore, examination of only one ("*un*folding") side is sufficient to estimate the barrier height.

However, a complete analysis of folding urges us to look at the barrier from its folding (connected with entropy loss) side, which is most interesting for the audience, and obtain the second view on the protein folding puzzle.

To analyze folding, we have to analyze sampling of conformations of the protein chain.

The total volume of the protein conformation space estimated at the level of amino acid residues by Levinthal [63] is huge indeed: as many as 100^{100} conformations for a 100-residue chain.

However, should the chain sample all these conformations in search for its most stable fold? No: the conformation space is covered by local energy minima, each surrounded by a local energy funnel (Fig. 1b) providing fast downhill decent to this local minimum.

Actually, the folding protein chain has to sample not all its possible conformations, but only various ways of packing the chain in the compact protein globule.

Therefore, to estimate the actual volume of sampling, one has to estimate the number of local energy minima (and also the time taken by jumping from one energy minimum to another). In some sense, this is similar to the idea to enumerate possible "topomers" that a protein chain can form [14,65], but our aim now is not to calculate the protein folding rate, but to estimate its lower limit only (which is very different from the somewhat contradictive [97] theory of the native-like topomer search by simulation).

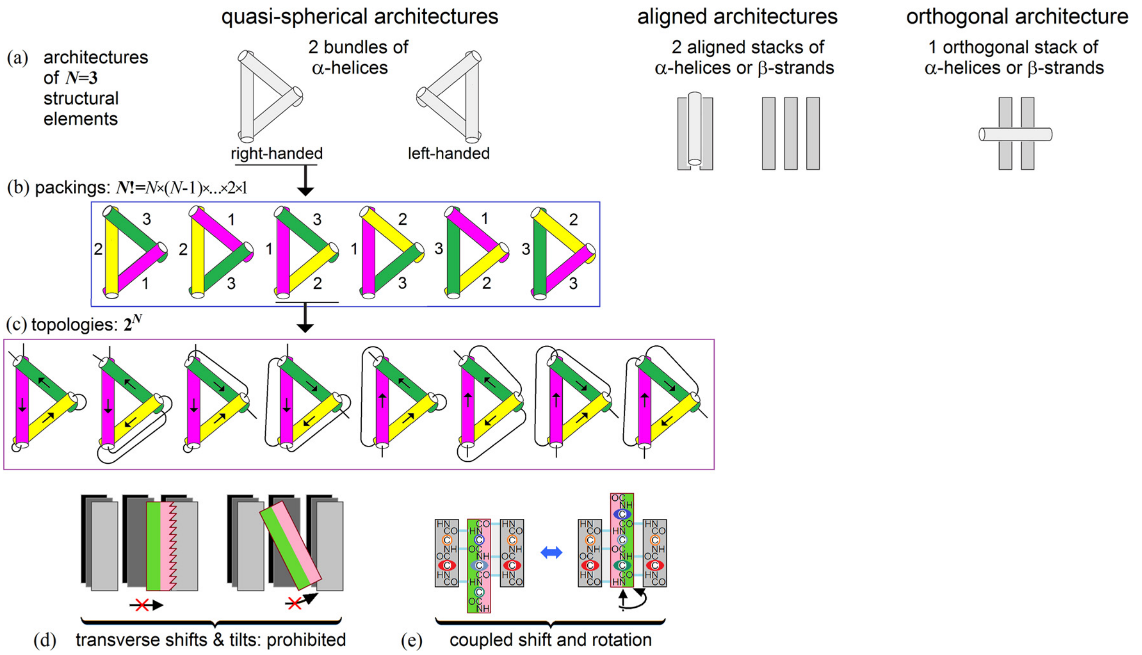
quasi-spherical architectures
2 bundles of α-helices
right-handed
left-handed
aligned architectures
2 aligned stacks of α-helices or β-strands
orthogonal architecture
1 orthogonal stack of α-helices or β-strands
(a) architectures of N=3 structural elements
(b) packings: N!=N×(N-1)×..×2×1
(c) topologies: 2^N
(d) transverse shifts & tilts: prohibited
(e) coupled shift and rotation

Fig. 7. A scheme of estimate of the conformation space volume at the level of secondary structure assembly. Adapted from Supplement to [27].

An overview of protein structures shows that interactions occurring in the chains are mainly connected with secondary structures [9,29,64]. 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 globule, that is, at the level considered by Ptitsyn [77] in his model of stepwise protein folding.

It turns out that this number is by many orders of magnitude smaller than that of conformations of amino acid residues [27]: 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 L and the number N of the secondary structure elements. N is much less than L, and this is the main reason for the drastic decrease of the conformation space.

The estimate L^N was obtained as follows (see Fig. 7).

The number of architectures (i.e., types of dense stacks of secondary structures) is small (cf. [9,64,69]), usually $\sim$10 or less for a given set of secondary structures (Fig. 7a), since the architectures are packings of a few secondary structure layers (each containing several secondary structures), and therefore combinatorics of the layers is very small as compared to combinatorics of much more numerous secondary structure elements, which is described below.

The maximal number of packings, i.e., all combinations of positions of N elements in the given protein architecture is shown in Fig. 7b.

The maximal number of topologies, i.e., all combinations of directions of these elements cannot exceed 2^N (Fig. 7c).

Transverse shifts and tilts of an element within each dense packing are prohibited (Fig. 7d).

Shifts and turns of secondary structure elements within a dense packing are coupled (this is shown in Fig. 7e using a β-sheet as the best illustrative example, but this is also true for α-helices – remember "knobs in the holes" close packings by Crick [13]); as a result, each α or β element can have about L/N (that is, about the element's mean length) possible shift/turns in the globule formed by N secondary structures of the L-residue chain.

All this limits the number of energy minima in the conformational space to $\sim 10 \times (L/N)^N \times 2^N \times N!$ conformations; this (using Stirling's approximation $N! \sim (N/e)^N$) gives

$$\text{NUMBER of energy minima to be sampled} \sim L^N \tag{5}$$

in the main term (if $L \gg N \gg 1$) [27,28].

This number can be reduced by symmetry of the globule; also, no α-helix can take the place of a β-strand without rearrangement of other elements, and *vice versa*, because the β-strand needs a partner to form hydrogen bonds, while the α-helix avoids such a partnership. Further, short or crossing loops between secondary structures can prevent these from taking arbitrary positions and directions in the globule, etc. [79]. However, this reduction is not important to us, because our aim now is to estimate the upper limit of the number of conformations.

Here, a question may arise as to how the chain knows where to form a secondary structure and what secondary structure is to be formed there. The answer is as follows. First of all, our aim now is not to model the folding process, but to estimate the number of conformations at the level of formation and assembly of secondary structures into a globule. Second, most of secondary structures are determined by local amino acid sequences [10,32,57,77,78,80,84], although some of them depend on their environment in the globule. Third, the optimal position of ends of the secondary structures in each ensemble can be rapidly found by descent in energy funnels, independent for each side of packing of the secondary structures. Fourthly, the choice "α or β" for all N secondary structure elements multiplies the estimate given by eq. (5) by 2^N at most, but in fact much less, because intrusion of an α-helix into a β-sheet (or of a β-strand between α-helices) is so energetically unfavorable, that it is never observed in proteins [29,30]. And, at last, the choice of "to be or not to be" for a secondary structure element adds only 1 state to the number L/N of the possible shift/turn states of this element (already taken into account), which is not significant (see Supporting Information to [27]). Thus, the number of energy minima to be sampled can be, rather roughly, estimated as L^N.

In a compact globule of not too small size, the length of a secondary structure element should be proportional to the globule's diameter, i.e., to $\sim L^{1/3}$. More specifically, the globule's volume is about 150 $\text{Å}^3 \times L$ (and thus its diameter is $\approx 5\ \text{Å} \times L^{1/3}$), while the shift per residue is about 1.5 Å in a helix and 3 Å in an extended strand [29]. Therefore, a helix consists of $\approx 3L^{1/3}$ residues, while a β-strand, as well as a loop, comprises $\approx 1.5L^{1/3}$ residues. Thus,

$$\text{NUMBER of “secondary structure + loop” elements } N \approx L^{2/3}/4.5 - L^{2/3}/3, \tag{6}$$

and the value L^N should be expected to come within the range

$$\sim L^{L^{2/3}/4.5} \equiv \exp\left(\left[\ln(L)/4.5\right] \times L^{2/3}\right) - \sim L^{L^{2/3}/3} \equiv \exp\left(\left[\ln(L)/3\right] \times L^{2/3}\right) \tag{7}$$

Analogous scaling of L^N looks like that obtained by [37,91] from mathematical consideration of the problem complexity.

One can see that, since $\ln(L)/4.5 \approx 1$ and $\ln(L)/3 \approx 1.5$ for $L \approx$ 80–90 residues, the above obtained limits are close to the upper limit outlined by eq. (4).

On the other hand, the value of L/N (i.e., the number of residues per secondary structure element plus accompanying loops) is 15 ± 5, according to protein statistics [82]; this, eventually, results in an estimate of L^N, which is numerically more or less close to the above given values.

Taking, from experiments on folding of the smallest proteins [42,66,67], a few microseconds as a rough estimate of the time necessary to sample one conformation and the value $L/N = 15 \pm 5$ from protein statistics, we see that the time theoretically needed to sample the whole conformation space at the level of secondary structure formation and assembly closely approaches (Fig. 8) the upper limit of experimental folding times (that is, the lower limit of experimental folding rates) observed for small ($L \lesssim$ 80–90 residues) proteins. It is also close to the upper limit of the folding time estimate given by eq. (4), earlier obtained from consideration of unfolding and illustrated in Fig. 6; note that folding of these small proteins is, according to eq. (4), under complete thermodynamic control.

The above consideration does not mean, of course, that a folding protein samples the *entire* conformation space at the level of secondary structure formation and packings (though a chain of 80–90 residues or less can do this within minutes (or faster), as Fig. 8 shows for some proteins). It means only that the native fold-leading "energy funnel", working at the level of secondary structures, has to accelerate folding by several orders of magnitude (as Fig. 8 shows for the majority of proteins), rather than by many tens or hundreds of orders, which would have been the case *if* the funnel were to start working from the level of amino acid residues (cf. with the theory of searching for topomers [14,65]). Fig. 8 shows that "funnel-due" acceleration is pronounced for chains of >100 residues, but even then the main work is done by secondary structures.

Bird's-eye view of the obtained estimates (4)–(7) of the number of chain conformations (or rather, of all kinds of chain packing in a compact globule), which have to be enumerated when searching for the most stable protein structure is as follows. This number scales, in the main term, in proportion to the globule's surface, i.e., to the number

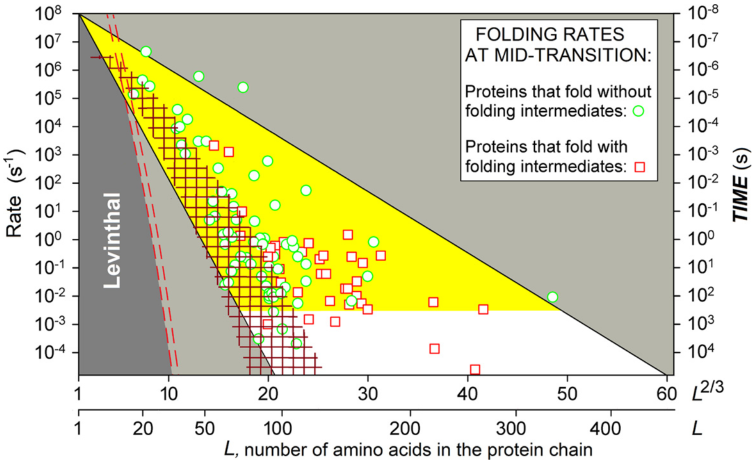

Fig. 8. Sampling rate and folding rate. Folding rates (circles and squares) are shown for proteins experimentally studied at mid-transition (i.e., at equal stability of their folded and unfolded states); the golden/white triangle shows the predicted (from consideration of *unfolding*!) range of these rates (cf. Fig. 6). The netted shading shows a theoretical estimate of the minimal rate of exhaustive sampling, at *folding*, of all possible packings of protein secondary elements (helices and strands). The maximal "Levinthal-like" sampling rate (10^{12} s$^{-1}/3^L$, allowing for 3 possible states: α, β, coil) is shown by the double dashed line; the lines for "Levinthal-like" sampling rates with 10 or 100 possible states of a residue would have been much below (in the dark-gray zone). Adapted from [22].

of surface residues or to the number of the secondary structures N, which are both proportional to $L^{2/3}$. The physical reason is that in a dense globule all independent degrees of freedom are connected only with its surface, because the globule's density prohibits independent rearrangements of residues in its interior [86,88], just like the secondary structure prohibits independent movements of residue backbones inside it. From this point of view, the used secondary structure elements are not necessary for estimating the scaling law (estimates by Fu and Wang [37] and Steinhofel et al. [91], as well as our estimates [23,24,40,46], did not use secondary structures), though these structures do form the protein core, and they are useful for refinement of the principal law.

4. Discussion and conclusion

We have viewed the pathways through the "volcano-shaped" (illustrated in Fig. 5) folding landscape both from outside, i.e., from the "volcano" foot, and from inside, that is, from its crater. In this way we investigated the free energy barrier separating the folded and unfolded states of a protein chain from its both sides. We have passed it there and back again and obtained two views on the protein folding puzzle; these two views solve the Levinthal's paradox.

The barrier side facing the folded state is easier for investigation because it is easier to outline a reasonable *un*folding pathway from any given fold than a good folding pathway to a fold that is still unknown for the chain. The view from inside of the folding funnel gave us an estimate of the range of unfolding times, and then we used the detailed balance principle to find the folding time.

The view from outside of the folding funnel gave us only the upper limit of the folding time.

It is worth mentioning that the unfolding-based estimate gives both the upper and lower estimates of the folding time, while the folding-based estimate gives its upper limit only.

The same scheme can be applicable to formation of the native protein structure not only from the coil (which we used in this study for simplicity) but also from the molten globule or from another intermediate. However, for these scenarios, all the estimates would be much more cumbersome due to more complicated nature of the denatured state of the protein, while these processes do not demonstrate (in experiment, see Fig. 6) any drastic advantage in the folding rate. Therefore, we now will not go beyond the simplest case of the coil-to-native globule transition.

It is not out of place mentioning that something similar to the Levinthal's problem must exist in crystallization (which resembles protein folding, because atoms of a few sorts have to acquire a particular conformation among plentiful others in "yet unknown" for them crystal); though, to our best knowledge, it did not attract there as much attention as in the protein science (cf. [90,95]).

A few more things remain to be said:

1. Our estimate of the number of energy minima to be sampled, eqs. (5)–(6), is the upper estimate, which does not take into account that some of these minima can have very high energy and therefore will be practically inaccessible for the folding protein chain.

2. This estimate of the number of energy minima is independent of stability of the native state under physiological conditions. The influence of this stability is considered below.

3. Our basic estimate of the folding time, eq. (4), refers to $\Delta F = 0$, i.e., to the point of equilibrium between the unfolded and native states – here the observed folding time is at a maximum and can exceed by orders of magnitude the folding time under native conditions [19].

How will the folding time change when the native state becomes somewhat more stable than the coil (that is, $\Delta F < 0$)? In accordance with experiment (see [19]), the theoretical analysis [24,26,29,30] shows that as long as $-\Delta F$ is small, about a few $k_B T$, so that no stable intermediates arise, the folding time decreases with increasing stability, and, theoretically, it can be estimated [46] as

$$TIME \sim \tau \times \exp\left[(1 \pm 0.5) \times \left(L^{2/3} + 0.4 \times \Delta F/RT\right)\right]; \tag{8}$$

the multiplier 0.4 corresponds to the approximate theoretical estimate of the average fraction of a chain involved in the folding nucleus, so that $0.4 \times \Delta F$ is the approximate change of the nucleus free energy. (The overview of other details of folding nuclei is out of the scope of this paper; one can find them in [19,26,29,30,40].) Equation (8) gives a unified approximate estimate of folding rates occurring under various conditions (see Fig. 6).

For the case of a very high native fold stability ($-\Delta F \gg k_B T$), another but similar to eq. (4) scaling law ($\ln(TIME) \sim L^{1/2}$) was obtained [92]. Then protein folding is the fastest, because it essentially goes "downhill" in energy all the way; but the "downhill slope" has (due to protein heterogeneity) random bumps, whose energy is proportional to $L^{1/2}$. However, numerical experiments with lattice protein chains have shown [52,88] that, at the temperature providing the fastest folding, the folding time grows with the chain length as $\ln(TIME) \sim A \times \ln(L)$, where the coefficient A equals to 6 for chains with "random" sequences and 4 for sequences selected to fold most rapidly (i.e., for chains having a large energy gap between the most stable fold and other ones). This emphasizes once again the dependence of the folding rate on experimental conditions and on the difference in stability between the lowest-energy fold and its competitors [39,101].

4. Here, it is worth mentioning that some, quite rare proteins are "metamorphic" [68]: they are observed in two or more distinct folds. Of interest for us are those very few in number (e.g., serpin) that first obtain some "native", that is, working structure, work in the cell or a test-tube for an hour or so, and then acquire another, non-working but more stable structure [94]. Significantly, this transition is not connected with a change in the protein's environment (aggregation, as in amyloids, or formation of some complexes). Thus, the chain of such a protein has two stable folds: one of them folds faster, the other is more stable. It seems, though, that such "metamorphic" (or "polymorphous") proteins are and must be very rare: theoretical estimates [29] show that the amino acid sequence coding for one stable chain fold (i.e., whose energy is separated by a wide gap from energies of others) is a kind of wonder by itself, but the sequence coding for two stable folds is a squared wonder...

5. Equations (4), (8) estimate the *range* of possible folding rates rather than folding rates of an individual protein, which, even for proteins of the same size, may differ (Fig. 6) from one another by orders of magnitude. The influence of a particular protein chain fold shape upon the folding rate can be estimated using a phenomenological "contact order" parameter (CO%) [74]. CO% is equal to the average distance along the chain between residues that are in contact in the native protein fold divided by the chain length (see also [71,72]). A high CO% value reflects the presence of many long closed loops in the protein fold, while a high value of (1 ± 0.5) factor in eqs. (4), (8) reflects their presence in a semi-folded globule (Fig. 4). Therefore, CO% is more or less proportional to this factor (1 ± 0.5) [53]. CO% by itself is a good predictor of folding rates of proteins equal in size, but it fails to compare folding rates of small proteins with those of large ones, because CO% decreases approximately in proportion to $L^{-1/3}$ with increasing protein size L [46,53,54] (which reflects a low entangling of chains forming large domains), – while the folding rate decreases, on the average, with increasing protein size (Fig. 6).

Therefore, a really good predictor of protein folding rates is AbsCO = CO% × L, which scales as $L^{2/3}$ [53] and combines the effect of protein fold shape [54] with the main effect of protein (and thus also nucleus) size [41,44]. The attempts to use machine learning and information provided by protein sequences to raise the quality of predictions over the level achieved with AbsCO (or ln(AbsCO) [33]) were not quite successful up to now [11, for more details, see references therein].

Coming back to the Levinthal's paradox, we can conclude that it is solved for protein chains of less than 100 amino acid residues (provided sequences of these chains ensure a significant stability to only one of their folds); this is because (i) these chains can overcome free energy barriers at the pathway to their most stable folds, independently of their complexity (Fig. 6), and (ii) they are able to sample all their folds at the level of secondary structure formation and assembly (Fig. 8) and find the most stable one. As to the chains of larger proteins, they can sample only relatively simple (not too entangled) folds, and it remains a question whether some another fold can be more stable than the native one (which is indeed observed for some "exceptional" proteins like serpin, having a 400-residue chain).

Acknowledgements

We are grateful to O.B. Ptitsyn, A.M. Gutin and E.I. Shakhnovich for seminal discussions and E.V. Serebrova for assistance in manuscript preparation.

The first part of this work has been supported by the Russian Academy of Sciences "Molecular and Cell Biology" Program Grant Nos. 01200957492, 01201358029, and the second part – by the Russian Science Foundation Grant No. 14-24-00157.

References

[1] Abkevich VI, Gutin AM, Shakhnovich EI. Specific nucleus as a transition state for protein folding: evidence from the lattice model. Biochemistry 1994;33:10026–31.
[2] Anfinsen CB, Haber E, Sela M, White FH Jr. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci USA 1961;47:1309–14.
[3] Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–30.
[4] Bicout DJ, Szabo A. Entropic barriers, transition states, funnels, and exponential protein folding kinetics: a simple model. Protein Sci 2000;9:452–65.
[5] Bogatyreva NS, Finkelstein AV. Cunning simplicity of protein folding landscapes. Protein Eng 2001;14:521–3.
[6] Bryngelson JD, Wolynes PG. Spin glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci USA 1987;84:7524–8.
[7] Bryngelson JD, Wolynes PG. Intermediates and barrier crossing in a random energy model (with applications to protein folding). J Phys Chem 1989;93:6902–15.
[8] Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins 1995;21:167–95.
[9] Chothia C, Finkelstein AV. The classification and origins of protein folding patterns. Annu Rev Biochem 1990;59:1007–39.
[10] Chou PY, Fasman GD. Prediction of protein conformation. Biochemistry 1974;13:222–45.
[11] Corrales M, Cuscó P, Usmanova DR, Chen H-C, Bogatyreva NS, Filion GJ, et al. Machine learning: how much does it tell about protein folding rates? PLoS ONE 2015;10(11):e0143166.
[12] Creighton TE. Experimental studies of protein folding and unfolding. Prog Biophys Mol Biol 1978;33:231–97.
[13] Crick FHC. The packing of α-helices: simple coiled coils. Acta Crystallogr 1953;6:689–97.
[14] Debe DA, Carlson MJ, Goddard WA 3rd. The topomer-sampling model of protein folding. Proc Natl Acad Sci USA 1999;96:2596–601.
[15] Dill KA, Chan HS. From Levinthal to pathways to funnels. Nat Struct Biol 1997;4:10–9.
[16] Dill KA, MacCallum JL. The protein-folding problem, 50 years on. Science 2012;338:1042–6.
[17] Emanuel NM, Knorre DG. The course in chemical kinetics. 4th edn. Moscow: Vysshaja Shkola; 1984 (in Russian). Chapters III (§ 2), V (§§ 2, 5).
[18] Eyring H. The activated complex in chemical reactions. J Chem Phys 1935;3:107–15.
[19] Fersht A. Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding. NY: W. H. Freeman & Co.; 1999. Chapters 2, 15, 18, 19.
[20] Fersht AR. Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. Proc Natl Acad Sci 2000;97:1525–9.
[21] Finkelstein AV. Cunning simplicity of a hierarchical folding. J Biomol Struct Dyn 2002;20:311–3.
[22] Finkelstein AV. Two views on the protein folding puzzle, http://atlasofscience.org/two-views-on-the-protein-folding-puzzle/, 2015.
[23] Finkelstein AV, Badretdinov AYa. Physical reason for fast folding of the stable spatial structure of proteins: a solution of the Levinthal paradox. Mol Biol (Mosc) 1997;31:391–8.
[24] Finkelstein AV, Badretdinov AYa. Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. Fold Des 1997;2:115–21.
[25] Finkelstein AV, Badretdinov AYa. 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 1998;3:67–8.
[26] Finkelstein AV, Galzitskaya OV. Physics of protein folding. Phys Life Rev 2004;1:23–56.

[27] Finkelstein AV, Garbuzynskiy SO. 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 2015;16:3373–8.

[28] Finkelstein AV, Garbuzynskiy SO. Solution of Levinthal's paradox is possible at the level of the formation and assembly of protein secondary structures. Biophysics 2016;61:1–5.

[29] Finkelstein AV, Ptitsyn OB. Protein physics. A course of lectures. Amsterdam – Boston – London – New York – Oxford – Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo: Academic Press, an Imprint of Elsevier Science; 2002. Chapters 7, 10, 13–21.

[30] Finkelstein AV, Ptitsyn OB. Protein physics. 2nd edn. Amsterdam – Boston – Heidelberg – London – New York – Oxford – Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo: Academic Press, an Imprint of Elsevier Science; 2016.

[31] Finkelstein AV, Badretdinov AYa, Gutin AM. Why do protein architectures have a Boltzmann-like statistics? Proteins 1995;23:142–50.

[32] Finkelstein AV, Badretdinov AYu, Ptitsyn OB. Short alpha-helix stability. Nature 1990;345:300.

[33] Finkelstein AV, Bogatyreva NS, Garbuzynskiy SO. Restrictions to protein folding determined by the protein size. FEBS Lett 2013;587:1884–90.

[34] Finkelstein AV, Ivankov DN, Garbuzynskiy SO, Galzitskaya OV. Understanding the folding rates and folding nuclei of globular proteins. Curr Protein Pept Sci 2007;8:521–36.

[35] Finkelstein AV, Ivankov DN, Garbuzynskiy SO, Galzitskaya OV. Understanding the folding rates and folding nuclei of globular proteins. In: Dunn BM, editor. Frontiers in protein and peptide sciences, vol. 1. 2014. p. 91–138. eBook series. Chapter 5.

[36] Flory PJ. Statistical mechanics of chain molecules. NY: Interscience Publishers; 1969. Chapter 3.

[37] Fu B, Wang W. A $2^{0(\mathrm{n}^{1-1/\mathrm{d}}\cdot\log(\mathrm{n}))}$ time algorithm for d-dimensional protein folding in the HP-model. Lecture Notes in Computer Science, vol. 3142. 2004. p. 630–44.

[38] Fulton KF, Main ERG, Dagett V, Jackson SE. Mapping the interactions present in the transition state for unfolding/folding of FKBP12. J Mol Biol 1999;291:445–61.

[39] Galzitskaya OV, Finkelstein AV. Folding of chains with random and edited sequences: similarities and differences. Protein Eng 1995;8:883–92.

[40] Galzitskaya OV, Finkelstein AV. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. Proc Natl Acad Sci USA 1999;96:11299–304.

[41] Galzitskaya OV, Glyakina AV. Nucleation-based prediction of the protein folding rate and its correlation with the folding nucleus size. Proteins 2012;80:2711–27.

[42] Galzitskaya OV, Higo J, Finkelstein AV. Alpha-helix and beta-hairpin folding from experiment, analytical theory and moleculare dynamics simulations. Curr Protein Pept Sci 2002;3:191–200.

[43] Galzitskaya OV, Ivankov DN, Finkelstein AV. Folding nuclei in proteins. FEBS Lett 2001;489:113–8.

[44] Galzitskaya OV, Garbuzynskiy SO, Ivankov DN, Finkelstein AV. Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. Proteins 2003;51:162–6.

[45] Garbuzynskiy SO, Kondratova MS. Structural features of protein folding nuclei. FEBS Lett 2008;582:768–72.

[46] Garbuzynskiy SO, Ivankov DN, Bogatyreva NS, Finkelstein AV. Golden triangle for folding rates of globular proteins. Proc Natl Acad Sci USA 2013;110:147–50.

[47] Go N, Abe H. Noninteracting local-structure model of folding and unfolding transition in globular proteins. I. Formulation. Biopolymers 1981;20:991–1011.

[48] Goldenberg DP, Creighton TE. Circular and circularly permuted forms of bovine pancreatic trypsin inhibitor. J Mol Biol 1983;165:407–13.

[49] Grantcharova VP, Riddle DS, Santiago JV, Baker D. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. Nat Struct Biol 1998;5:714–20.

[50] Grantcharova V, Alm E, Baker D, Horwich AL. Mechanism of protein folding. Curr Opin Struct Biol 2001;11:70–82.

[51] Gutin AM, Shakhnovich EI. Ground state of random copolymers and the discrete random energy model. J Chem Phys 1993;98:8174–7.

[52] Gutin AM, Abkevich VI, Shakhnovich EI. Chain length scaling of protein folding time. Phys Rev Lett 1996;77:5433–6.

[53] Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco KW, Baker D, Finkelstein AV. Contact order revisited: influence of protein size on the folding rate. Protein Sci 2003;12:2057–62.

[54] Ivankov DN, Bogatyreva NS, Lobanov MYu, Galzitskaya OV. Coupling between properties of the protein shape and the rate of protein folding. PLoS ONE 2009;4:e6476.

[55] Jackson SE. How do small single-domain proteins fold? Fold Des 1998;3:R81–91.

[56] Jacobson H, Stockmayer W. Intramolecular reaction in polycondensations. I. The theory of linear systems. J Chem Phys 1950;18:1600–6.

[57] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202;
Current version of the program: http://bioinf.cs.ucl.ac.uk/psipred/.

[58] Karplus M. The Levinthal paradox: yesterday and today. Fold Des 1997;2(Suppl. 1):S69–75.

[59] Landau LD, Lifshitz EM. Statistical physics. 3rd edn. A course of theoretical physics, vol. 5. Amsterdam – Boston – Heidelberg – London – New York – Oxford – Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo: Elsevier; 1980. §§ 7, 8, 150.

[60] Lappalainen I, Hurley MG, Clarke J. Plasticity within the obligatory folding nucleus of an immunoglobulin-like domain. J Mol Biol 2008;375:547–59.

[61] Leopold PE, Montal M, Onuchic JN. Protein folding funnels: a kinetic approach to the sequence–structure relationship. Proc Natl Acad Sci USA 1992;89:8721–5.

[62] Levinthal C. Are there pathways for protein folding? J Chim Phys 1968;65:44–5.

[63] Levinthal C. How to fold graciously. In: Debrunner P, Tsibris JCM, Munck E, editors. Mössbauer spectroscopy in biological systems: proceedings of a meeting held at Allerton House, Monticello, Illinois. Urbana-Champaign, IL: University of Illinois Press; 1969. p. 22–4.

[64] Levitt M, Chothia C. Structural patterns in globular proteins. Nature 1976;261:552–8.

[65] Makarov DE, Plaxco KW. The topomer search model: a simple, quantitative theory of two-state protein folding kinetics. Protein Sci 2003;12:17–26.

[66] Mukherjee S, Chowdhury P, Bunagan MR, Gai F. Folding kinetics of a naturally occurring helical peptide: implication of the folding speed limit of helical proteins. J Phys Chem B 2008;112:9146–50.

[67] Muñoz V, Thompson PA, Hofrichter J, Eaton WA. Folding dynamics and mechanism of beta-hairpin formation. Nature 1997;390:196–9.

[68] Murzin AG. Metamorphic proteins. Science 2008;320:1725–6.

[69] Murzin AG, Finkelstein AV. General architecture of α-helical globule. J Mol Biol 1988;204:749–70.

[70] Ngo JT, Marks J. Computational complexity of a problem in molecular structure prediction. Protein Eng 1992;5:313–21.

[71] Nölting B. Protein folding kinetics: biophysical methods. NY: Springer; 2010. Chapters 10, 11, 12.

[72] Nölting B, Schälike W, Hampel P, Grundig F, Gantert S, Sips N, et al. Structural determinants of the rate of protein folding. J Theor Biol 2003;223:299–307.

[73] Phillips DC. The three-dimensional structure of an enzyme molecule. Sci Am 1966;215:78–90.

[74] Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–94.

[75] Pauling L. General chemistry. NY: W.H. Freeman & Co.; 1970. Chapter 16.

[76] Privalov PL. Stability of proteins: small globular proteins. Adv Protein Chem 1979;33:167–241.

[77] Ptitsyn OB. Stages in the mechanism of self-organization of protein molecules. Dokl Akad Nauk SSSR 1973;210:1213–5 (in Russian).

[78] Ptitsyn OB, Finkel'shtein AV. Relation of the secondary structure of globular proteins to their primary structure. Biofizika 1970;15(5):757–68 (in Russian).

[79] Ptitsyn OB, Finkelstein AV. Similarities of protein topologies: evolutionary divergence, functional convergence or principles of folding? Q Rev Biophys 1980;13:339–86.

[80] Ptitsyn OB, Finkelstein AV. Theory of protein secondary structure and algorithm of its prediction. Biopolymers 1983;22:15–25.

[81] Robson B, Vaithilingam A. Protein folding revisited. Prog Mol Biol Transl Sci 2008;84:161–202.

[82] Rollins GC, Dill KA. General mechanism of two-state protein folding kinetics. J Am Chem Soc 2014;136:11420–7.

[83] Šali A, Shakhnovich E, Karplus M. Kinetics of protein folding. A lattice model study of the requirements for folding to the native state. J Mol Biol 1994;235:1614–36.

[84] Schulz GE, Barry CD, Friedman J, Chou PY, Fasman GD, Finkelstein AV, et al. Comparison of predicted and experimentally determined secondary structure of adenyl kinase. Nature 1974;250:140–2.

[85] Shakhnovich EI, Finkelstein AV. Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is the first order phase transition. Biopolymers 1989;28:1667–80.

[86] Shakhnovich EI, Gutin AM. Formation of unique structure in polypeptide-chains theoretical investigation with the aid of a replica approach. Biophys Chem 1989;34:187–99.

[87] Shakhnovich EI, Gutin AM. Implications of thermodynamics of protein folding for evolution of primary sequences. Nature 1990;346:773–5.

[88] Shakhnovich EI. Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. Chem Rev 2006;106:1559–88.

[89] Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, et al. Atom-level characterization of structural dynamics of proteins. Science 2010;330:341–6.

[90] Slezov VV. Kinetics of first-order phase transitions. Weiheim: Wiley–VCH; 2009. Chapters 3–5, 8.

[91] Steinhofel K, Skaliotis A, Albrecht AA. Landscape analysis for protein folding simulation in the H-P model. Lecture Notes in Computer Science, vol. 4175. 2006. p. 252–61.

[92] Thirumalai D. From minimal models to real proteins: time scales for protein folding kinetics. J Phys I 1995;5:1457–69.

[93] Tanford C. Protein denaturation. Adv Protein Chem 1968;23:121–282.

[94] Tsutsui Y, Cruz RD, Wintrode PL. Folding mechanism of the metastable serpin α1-antitrypsin. Proc Natl Acad Sci USA 2012;109:4467–72.

[95] Ubbelohde AR. Melting and crystal structure. Oxford: Clarendon Press; 1965. Chapters 2, 5, 6, 10–12, 14, 16.

[96] Unger R, Moult J. Finding the lowest free energy conformation of a protein is an NP-hard problem: proof and implications. Bull Math Biol 1993;55:1183–98.

[97] Wallin S, Chan HS. A critical assessment of the topomer search model of protein folding using a continuum explicit-chain model with extensive conformational sampling. Protein Sci 2005;14:1643–60.

[98] Wang J, Oliveira RJ, Chu X, Whitford PC, Chahine J, Han W, et al. Topography of funneled landscapes determines the thermodynamics and kinetics of protein folding. Proc Natl Acad Sci USA 2012;109:15763–8.

[99] Wensley BG, Gärtner M, Choo WX, Batey S, Clarke J. Different members of a simple three-helix bundle protein family have very different folding rate constants and fold by different mechanisms. J Mol Biol 2009;390:1074–85.

[100] Wetlaufer DB. Nucleation, rapid folding, and globular intrachain regions in proteins. Proc Natl Acad Sci USA 1973;70:697–701.

[101] Wolynes PG. Folding funnels and energy landscapes of larger proteins within the capillarity approximation. Proc Natl Acad Sci USA 1997;94:6170–5.

[102] Wolynes PG. Evolution, energy landscapes and the paradoxes of protein folding. Biochimie 2015;119:218–30.

[103] Wolynes PG, Onuchic JN, Thirumalai D. Navigating the folding routes. Science 1995;267:1619–20.

[104] Zana R. On the rate determining step for helix propagation in the helix–coil transition of polypeptides in solution. Biopolymers 1975;14:2425–8.

[105] Zwanzig R, Szabo A, Bagchi B. Levinthal's paradox. Proc Natl Acad Sci USA 1992;89:20–2.