Two Views on the Protein Folding Puzzle

Alexei V. Finkelstein, Oxana V. Galzitskaya, Sergiy O. Garbuzynskiy, Azat J. Badretdin, Dmitry N. Ivankov, and Natalya S. Bogatyreva

1 Introduction

1.1 Overview of the Protein Folding Problem

The ability of proteins to fold spontaneously puzzled protein science for a long time. It is well known that a protein chain (actually, the chain of a globular protein) can spontaneously fold into its unique native 3D structure [1, 2]. 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 [3, 4]: it is at least 2^{100} but rather may be 3^{100} or 10^{100} (or even 100^{100}) for a 100-residue chain, because at least 2 ("right" and "wrong"), but more likely 3 (α, β, "coil") or 10 [5] (or even (10_for_φ) $\times$ (10_for_Ψ) = 100 [3, 4]) conformations are possible for each residue (Fig. 1).

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

A. V. Finkelstein (✉) · O. V. Galzitskaya · S. O. Garbuzynskiy
Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region, Russian Federation
e-mail: afinkel@vega.protres.ru

A. J. Badretdin
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

D. N. Ivankov · N. S. Bogatyreva
Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region, Russian Federation

Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain

Universitat Pompeu Fabra (UPF), Barcelona, Spain

R. P. Mondaini (ed.), *Trends in Biomathematics: Modeling, Optimization and Computational Problems*, https://doi.org/10.1007/978-3-319-91092-5_27

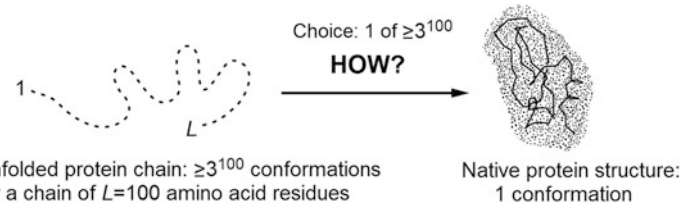

Fig. 1 The Levinthal's choice problem

least $\sim 2^{100}$ ps (or 3^{100} or 10^{100} or even 100^{100} ps), 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 [3, 4] (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, Levithal 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 [6, 7].

Nevertheless, most of the suggested hypotheses on protein folding are based on the "kinetic control assumption."

Ahead of Levinthal, Phillips [8] 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 [9, 10].

Wetlaufer [11] hypothesized formation of the folding nucleus by adjacent residues of the protein chain. However, in vitro experiments show that this is far not always so [12].

Ptitsyn [13] proposed a model of hierarchical folding, i.e., a stepwise involvement of different interactions and formation of different folding intermediate states. This hypothesis has some important advantages and drawbacks; below, it will be considered in detail, together with some interesting implications [14, 15] that follow from the Ptitsyn's model.

Alongside with approaches based on various hypotheses of "kinetic choice" of the protein native structure, some models and theories are based on the idea of the "stability choice" of this structure.

In particular, various "folding funnel" models [16–19] have become popular for illustrating and describing fast folding processes. These models, which have their own important advantages and drawbacks, will be considered below in detail as well.

On the top of that, the free-energy barrier separating the folded (native) and unfolded states of protein chains has been investigated, and the estimated rate of overcoming of this barrier [20, 21] turned out to be in a good concordance with experimental results (see below).

The difficulty of the "kinetics vs. stability" problem, underlined by Levinthal, 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?

1.2 Overview of the Basic Thermodynamic Facts Related to Protein Folding

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 $\sim$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 [22].
2. Protein unfolding is reversible [2]; moreover, the denatured and native states of a protein can be in a kinetic equilibrium [23]; and there is an "all-or-none" transition between them [5]. The latter means that only two states of the protein molecule, native and denatured, are present (close to the midpoint of the folding–unfolding equilibrium) in a visible quantity, while all others, "semi-native" or misfolded, are virtually absent.

   (Notes: (1) the "all-or-none" transition makes the protein function reliable: like a light bulb, the protein either works or not; (2) 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 [6, 24–27].)
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 [5] (and these two states have equal stability at mid-transition, naturally).

(Notes: For the below theoretical analysis, it is essential that (1) 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; (2) 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 [22], come within "energy." This terminology is commonly used to concentrate on the main problem of sampling the protein chain conformations.)

The abovementioned "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 the kinetics of this transition, and just this height is to be estimated to solve the Levinthal's paradox.

1.3 Is the Levinthal's Paradox a Paradox Indeed?

However, to begin with, it is not out of place considering whether the "Levinthal's paradox" is a paradox indeed. Bryngelson and Wolynes [28] mentioned that this "paradox" is based on the absolutely flat (and therefore unrealistic) "golf course" model of the protein potential energy surface (Fig. 2a), and somewhat later Leopold et al. [16], following the line of Gō and Abe [29], considered more realistic (tilted and biased to the protein's native structure) energy surfaces and introduced the "folding funnels" (Fig. 2b), which seemingly eliminate the "paradox" at all.

Its not as simple as that, though...

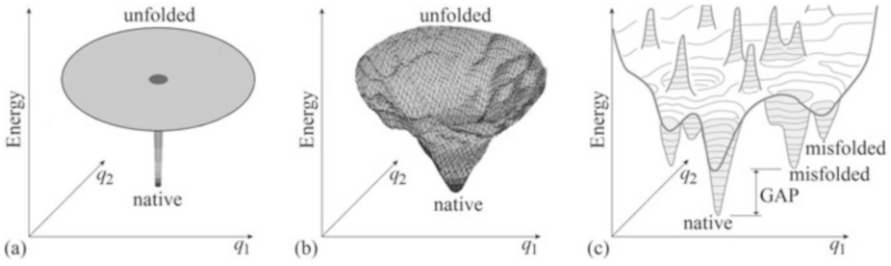

Fig. 2 Schematic illustration of basic models of the energy landscape of protein chains. (**a**) The "golf course" (Levinthal's) model of the protein potential energy landscape. (**b**) The "funnel" model of the protein potential energy landscape. The funnel is centered in the lowest-energy ("native") structure. (**c**) In more detail: the bumpy potential energy landscape of a protein chain. A wide (of many $k_B T_{\mathrm{melt}}$, 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

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 [30, 31], 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 [17, 32, 33]. 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) [28, 34–36]. Anyhow, the "energy funnel" directs movement toward 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 [37] of the straightforwardly presented funnel models [19, 38] shows that close to the midpoint of the folding–unfolding equilibrium they cannot *simultaneously* explain both the major features observed in protein folding: (1) its nonastronomical time, and (2) the "all-or-none" transition, i.e., coexistence of native and unfolded protein molecules during the folding process.

By the way, the stepwise mechanism of protein folding [13], taken per se, also cannot [39] *simultaneously* explain these two major features observed in protein folding. Rather, it states that the folding must be fast *if* each subsequent folding intermediate is much more stable than the preceding one (and thus, if the native fold is much more stable than the unfolded state of the chain).

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 very special nucleation funnels [20, 21]: those, considering the separation of the unfolded and native phases within the folding chain (now called the "capillarity theory" [40]).

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 [23, 41], and then observable intermediates often arise; but let us consider not the fastest but the simplest case... (We have to note that this special attention to the mid-transition conditions differs our approach [20, 21, 42] from those that prevailed from 1960s to the middle of 1990s.)

Since the "all-or-none" transition requires a large energy gap between the most stable structure and misfolded ones [6, 24–27] (Fig. 2c), 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 or even 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. 3). *If* the free energy went downhill along the entire pathway, a 100-residue chain would 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 [43].

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, i.e., the least stable ("barrier") state on the reaction pathway. According to the conventional transition state theory [44–46], the time of the multistep process of overcoming the barrier is estimated as

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

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

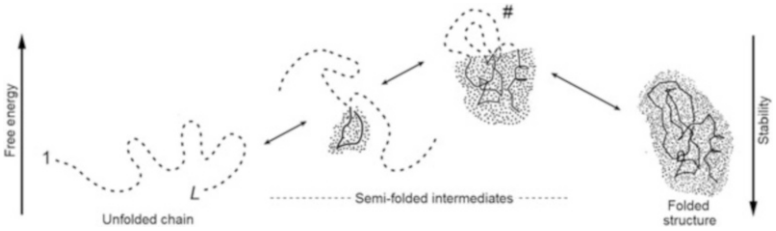

Fig. 3 A scheme [21] of a sequential folding pathway of some globular structure (if passed in the opposite direction, it is the sequential unfolding pathway of this structure). At each step of sequential folding, one residue leaves the coil and takes its final position in the structure. The free energy of intermediates is elevated due to the interface of folded and unfolded phases. The # sign indicates the most unstable ("transition") state. The folded part (shaded) of semi-folded intermediates which constitute the optimal ("low-free-energy") pathway must be compact (having a small boundary between the folded and unfolded phases). The bold lines 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

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 [29].

Let us consider a *sequential wetlaufer1* folding pathway (Fig. 3). 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 [47], the direct and reverse reactions follow the same pathway and have equal rates when both the end states have equal stability. (This law follows from the second law of thermodynamics. It is 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 $A \underset{2}{\overset{1}{\rightleftharpoons}} B$, 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. 3) by minimal interfaces [20, 21, 48, 49].

(In this connection, it is not out of place mentioning that, odd enough, protein unfolding, in contrast to folding, has been never 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 one, that is, how the protein gains a huge energy required for unfolding... This shows that it is easier to imagine how to unfold any protein structure than 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. 3) 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. 3 and 4a, the number of restored interactions grows (and their total energy decreases, see Fig. 4c) 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 nonlinear *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 dotted line in Fig. 4c, if the folded parts do not form a compact piece, as in Fig. 4b.

The entropy decrease is also *approximately* proportional to the number n of residues that have taken their final positions (Fig. 4d).

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. 3 and 5). 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 [50, 51] to have a very slow, logarithmic dependence on its length. This again results in a nonlinear *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) [20]; see also the later rigorous mathematical papers [52, 53].

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

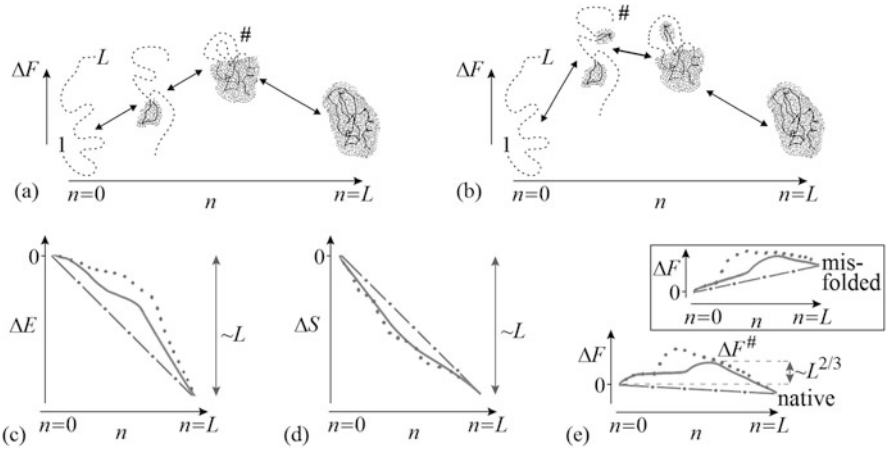

Fig. 4 Schematic illustration of sequential folding/unfolding with compact (**a**) and non-compact (**b**) semi-folded intermediates and the change of energy (**c**), entropy (**d**), and free energy (**e**) along these sequential folding/unfolding pathways close to the point of thermodynamic equilibrium 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 bar-dotted lines show the linear (proportional to the number of already folded residues n) parts of $\Delta E(n)$ and $\Delta S(n)$. The nonlinear 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 (bar-dotted 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 $\Delta F^{\#}$ ("the barrier") over the bar-dotted free-energy baseline 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 [20, 21]

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. 4e and 6) 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 [20, 21, 42, 49]), 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.

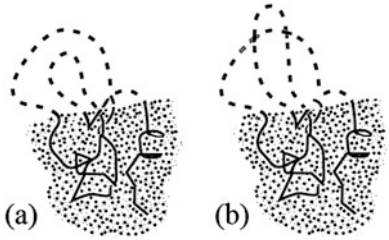

Fig. 5 A compact semi-folded intermediate with protruding unfolded loops. Its growth corresponds to a shift of the boundary between the folded (globular) and unfolded parts. Successful folding requires correct knotting (**a**) of loops: the structure with incorrect knotting (**b**) cannot change directly 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 [42], the search for correct knotting can only slow down the folding twofold or at most fourfold; thus, the search for correct chain knotting does not limit the folding rate of normal size protein chains. Adapted from [42]

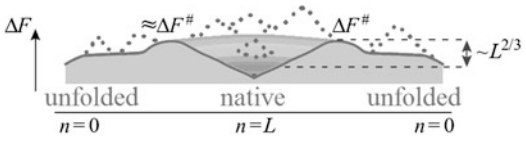

Fig. 6 This purely illustrative drawing shows how entropy converts the energy funnel (illustrated in Fig. 2b) into a "volcano-shaped" (as it is called now [54]) *free-energy* folding landscape with free-energy barriers (Fig. 4e) on each pathway leading from an unfolded conformation to the native fold. 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. 4a), the rocks (denoted by dotted lines) present a landscape including non-compact semi-folded intermediate structures (shown in Fig. 4b). More accurate but less beautiful scheme of a free-energy landscape is shown in Fig. 2 in [48]

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

$$L^{2/3} \cdot \frac{1}{4}\varepsilon \,, \tag{2a}$$

where $\varepsilon \approx 1.3$ kcal/mol $\approx 2k_B T_{\mathrm{mel}}$ is the average heat of protein melting per residue [5] (this is the first empirical parameter used by the theory), and $\approx \frac{1}{4}$ is the fraction of interactions lost by an interface residue. Thus,

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

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. 5).

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

$$(\Delta S^{\#})_{\text{lower}} = \frac{1}{6} L^{2/3} \left[-\frac{5}{2} k_B \ln(3\, L^{1/3}) \right], \tag{3a}$$

where $\frac{1}{6} L^{2/3}$ 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. 5), since the 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). $3L^{1/3} \equiv \left(\frac{L}{2}\right) / \left(\frac{1}{6} L^{2/3}\right)$ is the average number of residues in such a loop (equal to the number of unfolded residues divided by the number of loops), and $-\frac{5}{2} k_B \ln(3L^{1/3})$ 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 [20, 21]). Having $L \sim 100$ (actually, this approximation is good for the whole range of $L = 10$–1000), we obtain

$$(\Delta S^{\#})_{\text{lower}} \approx -k_B L^{2/3} \tag{3b}$$

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

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

where $\tau \approx 10\,\text{ns}$ [43] (this is the second and the last empirical parameter used in the theory).

The folding time depends on the size and the shape (see above) of the folding protein's native structure.

The physical reason for this "non-Levinthal" estimate is that (1) 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 (2) the free energy results only from surface effects which are relatively weak.

The observed protein folding times span (Fig. 7) 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 $10\,\text{ns} \times \exp(0.5L^{2/3})$ to $10\,\text{ns} \times \exp(1.5L^{2/3})$, 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 Sect. 4), 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. 7 shows that a chain of $L \lesssim$ 80–90 residues will find its most stable fold within minutes (or faster) even under "nonbiological" mid-transition conditions, where folding is known [23, 41] 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 [49]: 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 [49].

One thing is left to be said:

The "quasi-Levinthal" search over intermediates with different chain knotting (Fig. 5) can, in principle, be a "Levinthal-like" 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 (see [42] and references therein) which cannot fold within a reasonable time (according to Eq. (4)) in any case.

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 *unfolding* rather than *folding*. We have considered *unfolding* because it is easier to outline a good *un*folding pathway (and time, see above) 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. 5 and 6) 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

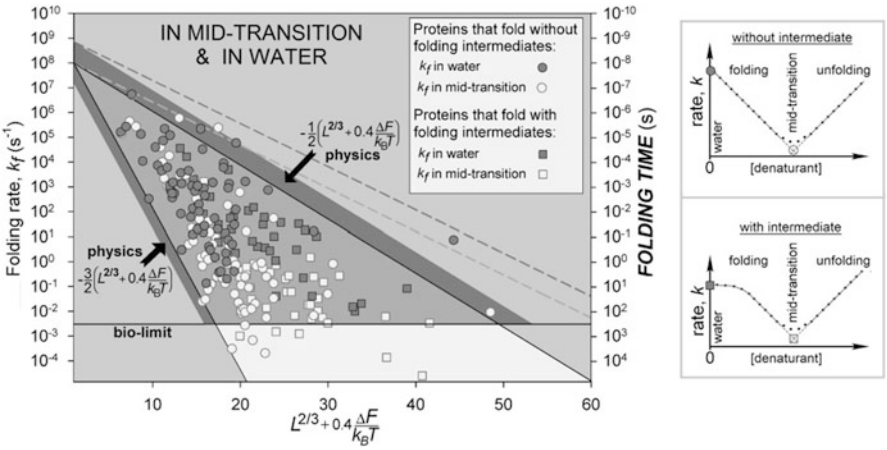

Fig. 7 Main panel: experimentally measured in vitro folding rate constants in water (under approximately "biological" conditions) and at mid-transition for 107 single-domain proteins (or separate domains) without SS bonds and covalently bound ligands (though the rates for proteins with and without SS bonds are principally the same [55]). Triangle: the region allowed by physics; its gray part (with the dark 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., nonbiological conditions. The light-gray dashed line limits the area allowed only for oblate (1:2) and oblong (2:1) globules at mid-transition; the dark-gray 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 [49]. Supplementary panels: Typical forms of "chevron plots" for the folding/unfolding kinetics of proteins that fold without and with folding intermediates (after [41])

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 biological 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 [4] is huge indeed: as many as from 3^{100} to 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. 2b) 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.

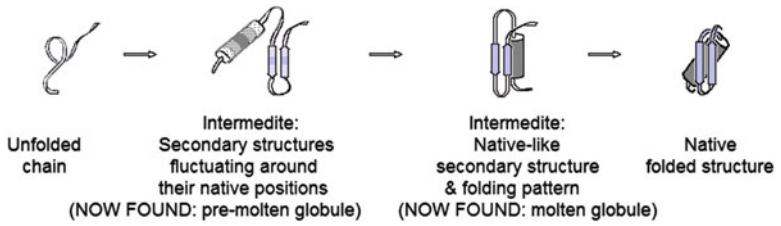

Fig. 8 A scheme of Ptitsyn's [13] hypothetical mechanism of stepwise protein folding. Cylinders: α-helices; arrows: β-strands. Both predicted in 1973 intermediates have been observed in 1980s–1990s [62]

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 [56, 57], but our aim is not to calculate the protein folding rate, but to estimate its lower limit only (which is very different from the somewhat contradictive [58] theory of the native-like topomer search by simulation).

An overview of protein structures shows that interactions occurring in the chains are mainly connected with secondary structures [13, 59–61]. 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 [13] in his model (Fig. 8) of stepwise protein folding.

It turns out that the number of conformations at the level of secondary structures is by many orders of magnitude smaller than that of conformations of amino acid residues of the chain [14]: 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 than $\sim L^N$ with the number of residues 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. 9).

The number of architectures (i.e., types of dense stacks of secondary structures) is small (cf. [59, 60, 63]), usually $\sim$10 or less for a given set of secondary structures (Fig. 9a), 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 (see Fig. 9b–e).

The maximal number of packings, i.e., all combinations of positions of N elements in the given protein architecture, cannot exceed $N!$ (Fig. 9b).

The maximal number of topologies, i.e., all combinations of directions of these elements cannot exceed 2^N (Fig. 9c).

Transverse shifts and tilts of an element within each dense packing are prohibited (Fig. 9d).

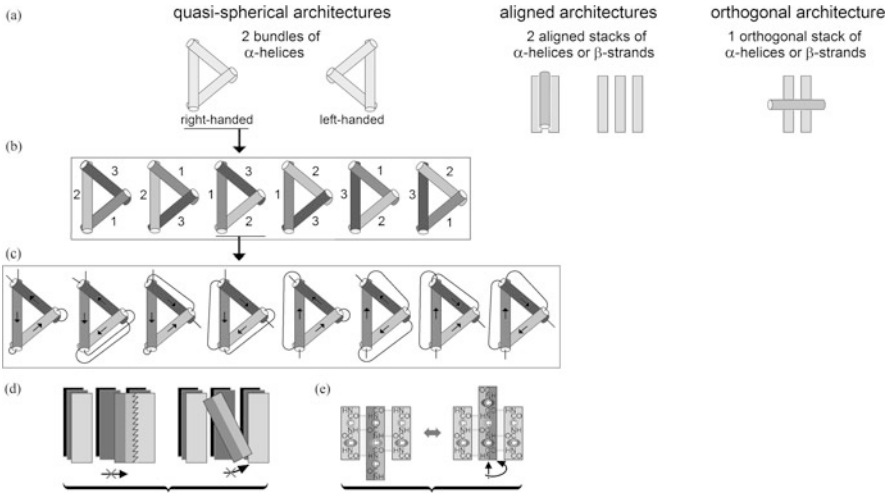

Fig. 9 A scheme of estimate of the conformation space volume at the level of secondary structure assembly. Adapted from Supplement to [14]. (**a**) Architectures of $N = 3$ structural elements. (**b**) Packings: $N! = N \times (N-1) \times \ldots \times 2 \times 1$. (**c**) Topologies: 2^N. (**d**) Transverse shifts and tilts: prohibited. (**e**) Coupled shift and rotation

Longitudinal shifts and turns about the axes of secondary structure elements within a dense packing are coupled (this is shown in Fig. 9e 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 [64]); as a result, each α or β element can have about L/N (that is, about the mean chain length per element) 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 $L \gg N \gg 1$) [14].

This number can be somewhat 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. [65]. 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 seems to be as follows. Most of secondary structures are determined by local amino acid

sequences [13, 66]. Anyhow, the choice of "to be or not to be" for a secondary structure element adds only 1 state to the number L/N of its possible shift/turn states (already taken into account), and conversion only duplicates it, which is not significant (see [67]).

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 [61]. 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, the mean number of residues in "secondary structure + loop" element is

$$L/N \approx 4.5L^{1/3} - 3L^{1/3}, \tag{6}$$

(which, at $L \sim 100$, is close to the value of $L/N = 15 \pm 5$ found from protein statistics [54]), and the mean number of "secondary structure + loop" elements is

$$N \approx \frac{L^{2/3}}{4.5} — \frac{L^{2/3}}{3}. \tag{7}$$

Thus, the value L^N (the sampling volume) is within the range

$$\sim L^{\frac{L^{2/3}}{4.5}} \equiv \exp\left(\left[\frac{\ln(L)}{4.5}\right] \times L^{2/3}\right) — \sim L^{\frac{L^{2/3}}{3}} \equiv \exp\left(\left[\frac{\ln(L)}{3}\right] \times L^{2/3}\right) \tag{8}$$

Analogous scaling was obtained [52, 53] from mathematical consideration of complexity of the choice problem. Also, one can see that, since $\ln(L)/4.5 \approx 1$ and $\ln(L)/3 \approx 1.5$ for $L \sim 100$, the estimate given by Eq. (8) is, eventually, more or less close to the upper limit outlined by Eq. (4).

Taking, from experiments on folding of the smallest proteins [68, 69], 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. 10) the upper limit of experimental folding times observed for small ($L \lesssim$ 80–90) residue 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. 7; note that folding of these small proteins is, as we have concluded, 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. 10 shows for the most slowly folding proteins of such size). It means only that the native fold-leading "energy funnel," working at the level of secondary structures, has to accelerate (for some, rapidly folding proteins) the folding process by several orders of magnitude (as Fig. 10 shows for the majority of proteins), rather

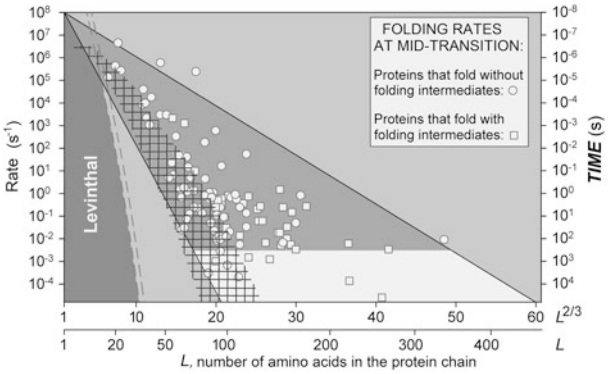

Fig. 10 Sampling rate and folding rate. Folding rates (circles and squares) are shown for proteins experimentally studied at mid-transition (i.e., at equal stability of their folded and unfolded states); the dark- gray/light-gray triangle shows the predicted (from consideration of *unfolding*!) range of these rates (cf. Fig. 7). 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, and 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 [70]

than for all proteins and 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 [56, 57]). Figure 10 shows that the "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)–(8) 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 of surface residues or—nearly the same—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 [24, 27], 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 [52] and Steinhofel et al. [53], as well as our estimates based on unfolding pathways [20, 21, 48, 49], 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 (or rather, sampling) 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. 7) 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. [71, 72]).

A few more things remain to be said:

1. Our estimate of the number of the secondary structure ensembles (i.e., the energy minima to be sampled) is independent (see Eqs. (5)–(8)) on stability of these ensembles. The influence of the native state stability (ΔF) on the folding *time* is considered below.
2. Our basic estimate of the folding *time*, Eq. (4), referred, up to now, to $\Delta F = 0$, i.e., to the point of equilibrium between the unfolded and native states—but here the observed folding time is at a maximum and can exceed by orders of magnitude the folding time under native conditions [41].

   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 [41]), the theoretical analysis [21, 61, 73, 74] 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 [49] as

$$\text{TIME} \sim \tau \times \exp\left[(1 \pm 0.5) \times (L^{2/3} + 0.4 \times \Delta F/RT)\right]; \quad (9)$$

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 [41, 48, 61, 73, 74]). Equation (9) 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 ($\ln(\text{TIME}) \sim L^{1/2}$) was obtained [75]. 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 [27, 76] that, at the temperature providing the fastest folding, the folding time grows with the chain length as $\ln(\text{TIME}) \sim A \times \ln(L)$, 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 [26, 40].

3. Here, it is worth mentioning that some, quite rare proteins are "metamorphic" [77]: 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 [78]. 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 [61, 74] 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...
4. Equations (4), (9) 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%) [79]. 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 [33, 80]). 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), (9) reflects their presence at the surface of the semi-folded globule (Fig. 5). Therefore, CO% is more or less proportional to this factor (1 ± 0.5) [81]. 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 [49, 81, 82] (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 $\mathrm{AbsCO} = \mathrm{CO\%} \times L$, which scales as $L^{2/3}$ [81] and combines the effect of protein fold shape [79, 82] with the main effect of protein size.

5. 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) [83]) were not quite successful up to now [84].
6. 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 that sequences of these chains ensure a significant stability to only one of their folds); this is because (1) these chains can overcome free-energy barrier at the pathway to their most stable folds, independently of their complexity (Fig. 7), and (2) they are able to sample all their folds at the level of secondary structure formation and assembly (Fig. 10) 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).
7. All told above is also applicable to in vivo folding, because NMR studies of ^{15}N, ^{13}C-labeled nascent chains of small protein state that "polypeptides [at ribosomes] remain unstructured during elongation but fold into a compact, native-like structure when the entire sequence is available" [85, 86]; thus, there is no principal difference between in vivo and in vitro protein folding.

Acknowledgements We are grateful to O.B. Ptitsyn, A.M. Gutin, and E.I. Shakhnovich for seminal discussions at the initial stages of our work.

The first part of this work has been partially supported by the Howard Hughes Medical Institute Awards and the Russian Academy of Sciences Program "Molecular and Cell Biology" (Grant Nos. 01200957492, 01201358029); the second part has been partially supported by the Russian Science Foundation Grant No. 14-24-00157.

References

1. C.B. Anfinsen, E. Haber, M. Sela, F.H. White Jr., Proc. Natl. Acad. Sci. U. S. A. **47**, 1309 (1961)
2. C.B. Anfinsen, Science **181**, 223 (1973)
3. C. Levinthal, J. Chim. Phys. Chim. Biol. **65**, 44 (1968)

4. C. Levinthal, in *Mössbauer Spectroscopy in Biological Systems: Proceedings of a Meeting Held at Allerton House, Monticello*, ed. by P. Debrunner, J.C.M. Tsibris, E. Munck (University of Illinois Press, Urbana-Champaign, 1969), p. 22
5. P.L. Privalov, Adv. Protein Chem. **33**, 167 (1979)
6. A. Šali, E. Shakhnovich, M. Karplus, J. Mol. Biol. **235**, 1614 (1994)
7. V.I. Abkevich, A.M. Gutin, E.I. Shakhnovich, Biochemistry **33**, 10026 (1994)
8. D.C. Phillips, Sci. Am. **215**, 78 (1966)
9. D.P. Goldenberg, T.E. Creighton, J. Mol. Biol. **165**, 407 (1983)
10. V.P. Grantcharova, D.S. Riddle, J.V. Santiago, D. Baker, Nat. Struct. Biol. **5**, 714 (1998)
11. D.B. Wetlaufer, Proc. Natl. Acad. Sci. U. S. A. **70**, 697 (1973)
12. K.F. Fulton, E.R.G. Main, V. Dagett, S.E. Jackson, J. Mol. Biol. **291**, 445 (1999)
13. O.B. Ptitsyn, Dokl. Akad. Nauk SSSR (Moscow, in Russian) **210**, 1213 (1973)
14. A.V. Finkelstein, S.O. Garbuzynskiy, ChemPhysChem **16**, 3373 (2015)
15. A.V. Finkelstein, A.J. Badretdin, O.V. Galzitskaya, D.N. Ivankov, N.S. Bogatyreva, S.O. Garbuzynskiy, Phys. Life Rev. **20** (2017). https://doi.org/10.1016/j.plrev.2017.01.025 [Epub ahead of print]
16. P.E. Leopold, M. Montal, J.N. Onuchic, Proc. Natl. Acad. Sci. U. S. A. **89**, 8721 (1992)
17. P.G. Wolynes, J.N. Onuchic, D. Thirumalai, Science **267**, 1619 (1995)
18. K.A. Dill, H.S. Chan, Nat. Struct. Biol. **4**, 10 (1997)
19. D.J. Bicout, A. Szabo, Protein Sci. **9**, 452 (2000)
20. A.V. Finkelstein, A.Ya. Badretdinov, Mol. Biol. (Moscow, Eng. Trans.) **31**, 391 (1997)
21. A.V. Finkelstein, A.Ya. Badretdinov, Fold. Des. **2**, 115 (1997)
22. C. Tanford, Adv. Protein Chem. **23**, 121 (1968)
23. T.E. Creighton, Prog. Biophys. Mol. Biol. **33**, 231 (1978)
24. E.I. Shakhnovich, A.M. Gutin, Nature **346**, 773 (1990)
25. A.M. Gutin, E.I. Shakhnovich, J. Chem. Phys. **98**, 8174 (1993)
26. O.V. Galzitskaya, A.V. Finkelstein, Protein Eng. **8**, 883 (1995)
27. E.I. Shakhnovich, Chem. Rev. **106**, 1559 (2006)
28. J.D. Bryngelson, P.G. Wolynes, J. Phys. Chem. **93**, 6902 (1989)
29. N. Gō, H. Abe, Biopolymers **20**, 991 (1981)
30. J.T. Ngo, J. Marks, Protein Eng. **5**, 313 (1992)
31. R. Unger, J. Moult, Bull. Math. Biol. **55**, 1183 (1993)
32. M. Karplus, Fold. Des. **2**(Suppl. 1), S69 (1997). http://www.sciencedirect.com/science/journal/13590278
33. B. Nölting, *Protein Folding Kinetics: Biophysical Methods*, Chaps. 10, 11, 12 (Springer, Berlin, 2010)
34. J.D. Bryngelson, P.G. Wolynes, Proc. Natl. Acad. Sci. U. S. A. **84**, 7524 (1987)
35. J.D. Bryngelson, J.N. Onuchic, N.D. Socci, P.G. Wolynes, Proteins **21**, 167 (1995)
36. A.V. Finkelstein, A.Ya. Badretdinov, A.M. Gutin, Proteins **23**, 142 (1995)
37. N.S. Bogatyreva, A.V. Finkelstein, Protein Eng. **14**, 521 (2001)
38. R. Zwanzig, A. Szabo, B. Bagchi, Proc. Natl. Acad. Sci. U. S. A. **89**, 20 (1992)
39. A.V. Finkelstein, J. Biomol. Struct. Dyn. **20**, 311 (2002)
40. P.G. Wolynes, Proc. Natl. Acad. Sci. U. S. A. **94**, 6170 (1997)
41. A. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*, Chaps. 2, 15, 18, 19 (W.H. Freeman & Co, New York, 1999)
42. A.V. Finkelstein, A.Ya. Badretdinov, Fold. Des. **3**, 67 (1998)
43. R. Zana, Biopolymers **14**, 2425 (1975)
44. H. Eyring, J. Chem. Phys. **3**, 107 (1935)
45. L. Pauling, *General Chemistry*, Chap. 16 (W.H. Freeman & Co., New York, 1970)
46. N.M. Emanuel, D.G. Knorre, *The Course in Chemical Kinetics*, 4th Russian edn., Chaps. III (§2), V (§§2, 5) (Vysshaja Shkola, Moscow, 1984)
47. L.D. Landau, E.M. Lifshitz, *Statistical Physics*. A Course of Theoretical Physics, vol. 5, 3rd edn. (Elsevier, Amsterdam, 1980) §§7, 8, 150
48. O.V. Galzitskaya, A.V. Finkelstein, Proc. Natl. Acad. Sci. U. S. A. **96**, 11299 (1999)

49. S.O. Garbuzynskiy, D.N. Ivankov, N.S. Bogatyreva, A.V. Finkelstein, Proc. Natl. Acad. Sci. U. S. A. **110**, 147 (2013)
50. H. Jacobson, W. Stockmayer, J. Chem. Phys. **18**, 1600 (1950)
51. P.J. Flory, *Statistical Mechanics of Chain Molecules*, Chap. 3 (Interscience Publishers, New York, 1969)
52. B. Fu, W. Wang, Lect. Notes Comput. Sci. **3142**, 630 (2004)
53. K. Steinhofel, A. Skaliotis, A.A. Albrecht, Lect. Notes Comput. Sci. **4175**, 252 (2006)
54. G.C. Rollins, K.A. Dill, General mechanism of two-state protein folding kinetics. J. Am. Chem. Soc. **136**, 11420 (2014)
55. O.V. Galzitskaya, D.N. Ivankov, A.V. Finkelstein, FEBS Lett. **489**, 113 (2001)
56. D.A. Debe, M.J. Carlson, W.A. Goddard 3rd., Proc. Natl. Acad. Sci. U. S. A. **96**, 2596 (1999)
57. D.E. Makarov, K.W. Plaxco, Protein Sci. **12**, 17 (2003)
58. S. Wallin, H.S. Chan, Protein Sci. **14**, 1643 (2005)
59. M. Levitt, C. Chothia, Nature **261**, 552 (1976)
60. C. Chothia, A.V. Finkelstein, Ann. Rew. Biochem. **59**, 1007 (1990)
61. A.V. Finkelstein, O.B. Ptitsyn, *Protein Physics. A Course of Lectures*, Chaps. 7, 10, 13, 17–21 (Academic, An Imprint of Elsevier Science, Amsterdam, 2002)
62. O.B. Ptitsyn, Adv. Protein Chem. **47**, 83 (1995)
63. A.G. Murzin, A.V. Finkelstein, J. Mol. Biol. **204**, 749 (1988)
64. F.H.C. Crick, Acta Crystallogr. **6**, 689 (1953)
65. O.B. Ptitsyn, A.V. Finkelstein, Quart. Rev. Biophys. **13**, 339 (1980)
66. D.T. Jones, J. Mol. Biol. **292**, 195 (1999). Current version of the program: http://bioinf.cs.ucl.ac.uk/psipred/
67. A.V. Finkelstein, S.A. Garbuzynskiy, Biofizika (in Russian) **61**, 5 (2016)
68. V. Muñoz, P.A. Thompson, J. Hofrichter, W.A. Eaton, Nature **390**, 196 (1997)
69. S. Mukherjee, P. Chowdhury, M.R. Bunagan, F. Gai, J. Phys. Chem. B **112**, 9146 (2008)
70. A.V. Finkelstein. http://atlasofscience.org/are-users-satisfied-with-single-sign-on-technologies-in-er/
71. A.R. Ubbelohde, *Melting and Crystal Structure*, Chaps. 2, 5, 6, 10–12, 14, 16 (Clarendon Press, Oxford, 1965)
72. V.V. Slezov, *Kinetics of First-Order Phase Transitions*, Chaps. 3–5, 8 (Wiley-VCH, Weiheim, 2009)
73. A.V. Finkelstein, O.V. Galzitskaya, Phys. Life Rev. **1**, 23 (2004)
74. A.V. Finkelstein, O.B. Ptitsyn, *Protein Physics. A Course of Lectures*, 2nd edn., Chaps. 7, 10, 13, 18, 19–21 (Academic, An Imprint of Elsevier Science, Amsterdam, 2016)
75. D. Thirumalai, J. Phys. I. (Orsay, Fr.) **5**, 1457 (1995)
76. A.M. Gutin, V.I. Abkevich, E.I. Shakhnovich, Phys. Rev. Lett. **77**, 5433 (1996)
77. A.G. Murzin, Science **320**, 1725 (2008)
78. Y. Tsutsui, R.D. Cruz, P.L. Wintrode, Proc. Natl. Acad. Sci. U. S. A. **109**, 4467 (2012)
79. K.W. Plaxco, K.T. Simons, D. Baker, J. Mol. Biol. **277**, 985 (1998)
80. B. Nölting, W. Schälike, P. Hampel, F. Grundig, S. Gantert, N. Sips, W. Bandlow, P.X. Qi, J. Theor. Biol. **223**, 299 (2003)
81. D.N. Ivankov, S.O. Garbuzynskiy, E. Alm, K.W. Plaxco, D. Baker, A.V. Finkelstein, Protein Sci. **12**, 2057 (2003)
82. D.N. Ivankov, N.S. Bogatyreva, M.Yu. Lobanov, O.V. Galzitskaya, PLoS One **4**, e6476 (2009)
83. A.V. Finkelstein, N.S. Bogatyreva, S.O. Garbuzynskiy, FEBS Lett. **587**, 1884 (2013)
84. M. Corrales, P. Cuscó, D.R. Usmanova, H.C. Chen, N.S. Bogatyreva, G.J. Filion, D.N. Ivankov, PLoS One **10**, e0143166 (2015)
85. C. Eichmann, S. Preissler, R. Riek, E. Deuerling, Proc.Natl. Acad. Sci. U. S. A. **107**, 9111 (2010)
86. Y. Han, A. David, B. Liu, J.G. Magadan, J.R. Bennink, J.W. Yewdell, S.-B. Qian, Proc. Natl. Acad. Sci. U. S. A. **109**, 12467 (2012)