Biophysical Reviews (2023) 15:661–669
https://doi.org/10.1007/s12551-023-01088-z

REVIEW

Protein folding rate evolution upon mutations

Jorge A. Vila[1]

Received: 10 April 2023 / Accepted: 24 June 2023 / Published online: 15 July 2023

Abstract
Despite the spectacular success of cutting-edge protein fold prediction methods, many critical questions remain unanswered, including why proteins can reach their native state in a biologically reasonable time. A satisfactory answer to this simple question could shed light on the slowest folding rate of proteins as well as how mutations—amino-acid substitutions and/or post-translational modifications—might affect it. Preliminary results indicate that (i) Anfinsen's dogma validity ensures that proteins reach their native state on a reasonable timescale regardless of their sequence or length, and (ii) it is feasible to determine the evolution of protein folding rates without accounting for epistasis effects or the mutational trajectories between the starting and target sequences. These results have direct implications for evolutionary biology because they lay the groundwork for a better understanding of why, and to what extent, mutations—a crucial element of evolution and a factor influencing it—affect protein evolvability. Furthermore, they may spur significant progress in our efforts to solve crucial structural biology problems, such as how a sequence encodes its folding.

Keywords Mutations · Evolution · Folding rate · Post-translational modifications · Levinthal paradox · Anfinsen dogma · Protein marginal stability

Introduction

Evolution and protein folding are intertwined processes. Indeed, protein sequences, encoded by DNA, determine their tridimensional structure (Anfinsen 1973), which in turn determines their function, while evolution could alter either one by mutations. Then, does the folding rate—which is a measure of how quickly or slowly a protein folds from its unfolded forms to its native state—restrains the mutation frequency? If this were the case, what would be its impact on evolution? Whatever the answer to these questions, protein folding cannot happen in cosmic times (~ 10^{27} years), as foreseen by an exhaustive sampling of all possible conformations for a 100-residue protein (Zwanzig et al. 1992), because the observed folding rates in water for single-domain two-state proteins are smaller than ~ 10 s (Garbuzynskiy et al. 2013). As the reader may be aware, several possible solutions to this apparent contradiction, also known as Levinthal's paradox (Levinthal 1968), exist in the literature (Zwanzig et al. 1992; Dill and Chan 1997; Karplus 1997; Rooman et al. 2002; Ben-Naim 2012; Finkelstein and Garbuzynskiy 2013; Martinez 2014; Ivankov and Finkelstein 2020; Finkelstein et al. 2022). However, the existence of numerous solutions to this paradox does not assure a clear answer to the following key question: why can proteins reach their native state in a biologically reasonable time? As a strategy to answer this question, we will prove that a reasonable estimation of the height of the activation barrier (see Fig. 1), separating the native state from the highest free-energy native-like conformation—beyond which the protein unfolds or becomes non-functional—will enable us to determine the slowest folding/unfolding time for two-state monomeric proteins. Before resuming the analysis, let us recall the last question. Should we focus on why—rather than on how—proteins reach their native state in a biologically reasonable time? This dilemma does not have a simple solution because both are relevant queries. Indeed, the interrogative how is associated with determining the mechanism, e.g., the routes or pathway/s of the folding/unfolding (Sali, et al. 1994; Wolynes et al. 1995; Lazaridis and Karplus 1997; Jackson 1998; Lindorff-Larsen et al. 2011; Englander and Mayne 2014; Wolynes 2015; Li and

✉ Jorge A. Vila
jv84@cornell.edu

[1] IMASL-CONICET, Universidad Nacional de San Luis, Ejército de Los Andes 950, 5700 San Luis, Argentina

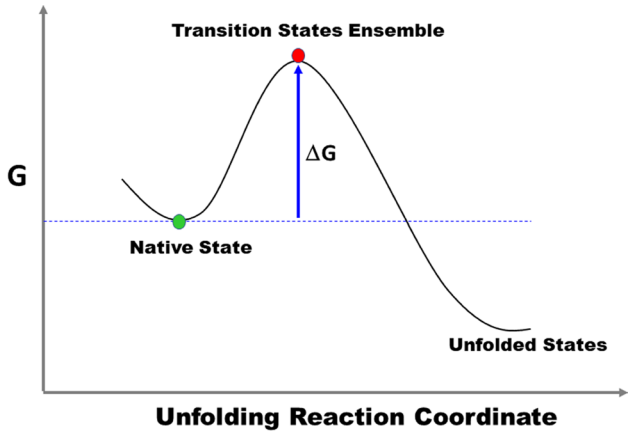

Fig. 1 The Gibbs free-energy profile (G) for a two-state protein unfolding is sketched out in broad strokes. The native state and the highest point of the free-energy profile are highlighted as green- and red-filled dots, respectively. The Gibbs free-energy gap between these two states is indicated by ΔG

Gong 2022), while the why is associated with identifying the main factors—independently of the mechanism—governing the folding/unfolding process. An attempt to answer how proteins reach their native state in a biologically reasonable time has been recently analyzed (Ivankov and Finkelstein 2020). Therefore, we choose to focus on why two-state proteins reach their native state in a biologically reasonable time because, in the first place, it questions our basic knowledge of the main factors determining protein folding rate changes and, hence, poses a preliminary problem to one of the most critical unanswered questions in structural biology: how a sequence encodes its folding. Second, it will help to understand the origins of protein folding rate evolution after amino-acid substitutions and/or post-translational modifications.

Overall, we start by determining the slowest folding rate for a two-state monomeric protein, i.e., by providing an answer to why proteins fold in a biologically reasonable time. Arguments, such as that life would not have emerged if it took the age of the universe for a protein to fold, or that proteins should fold fast enough in a cell—not to be degraded—could be, at first glance, plausible answers. None of these ideas, however, could adequately describe the nature of the key factors determining how protein folding/unfolding rates evolve in response to amino acid substitutions and/or post-translational modifications. For this reason, this phenomenon is examined here in terms of (i) protein-marginal stability (Dinner and Karplus 2001; Vila 2019; Martin and Vila 2020; Vila 2021) and (ii) arguments from the transition state theory (Ivankov and Finkelstein 2020). Unless otherwise stated, the terms "folding" and "unfolding" shall be used interchangeably from this point on.

Results and discussion

I.- Two-state protein folding time scales

Among the possible solutions to the time scales for protein folding, we distinguish three studies that have determined a plausible relation between protein length (N), with N being the number of residues, and folding time logarithm ($\ln \tau$), namely, $\ln \tau \sim N^{1/2}$ (Thirumalai 1995), $\sim \ln (N)$ (Gutin et al. 1996), and $\sim N^{2/3}$ (Finkelstein and Badretdinov 1997; Wolynes 1997). Although an analysis of such a relationship is vital, given the strongly observed anti-correlation—between N and $\ln \tau$—for three-state folding proteins ($R \sim -0.80$) (Galzitskaya et al. 2003), it is also equally important to highlight that such a relationship for two-state folding proteins is nearly inexistent ($R \sim -0.07$) (Galzitskaya et al. 2003). Therefore, we will focus on determining a plausible explanation for the latter. For this purpose, we will resolve the slowest folding/unfolding time (τ_{max}) for a monomeric two-state protein in terms of the result obtained for the marginal-stability upper bound of proteins obtained via a statistical-mechanics analysis of the partition function in the thermodynamic limit, also known as "the infinite chain limit" (Vila 2019, 2021). Therefore, for two-state proteins of any sequence and length (N), the expected value for the slowest folding/unfolding time (τ_{max}) will hold if the following conjecture and facts are plausible:

1. The folding approach for monomeric two-state proteins is a reversible thermodynamic-driven process (Privalov 1979; Matouschek et al. 1989)
2. The two-state protein unfolding model shown in Fig. 1 alludes to a process in which the thermodynamics and kinetic stability happen only between the native-state and unfolded states, which are separated by an energetic barrier higher than thermal fluctuation energy (Akmal and Muñoz 2004; Kuwajima 2020). In other words, folded and unfolded states are separated by an ensemble of a high-energy set of structures, i.e., the transition state ensemble (TSE), representing the energetic barrier for the process (Privalov 1979; Matouschek et al. 1989; Itzhaki et al. 1995; Englander 2000; Fersht and Daggett 2002; Akmal and Muñoz 2004; Shakhnovich 2006). In this simple unfolding model, there are no stable intermediate states necessary to complete the process
3. We will focus our attention on the analysis of the unfolding rather than on the folding process because the former enables us to make a quick estimation of the height of the Gibbs free-energy difference (ΔG) between the native state (representing a well-defined reference point) and the highest point of the TSE (see Fig. 1). The lat-

ter is feasible since the "detailed balance principle" demands that the TSE be the same for unfolding and folding processes (Ivankov and Finkelstein 2020), e.g., as shown by the analysis of the rates and equilibria of folding from ~ 100 mutants strategically distributed throughout the protein chymotrypsin inhibitor 2 (Itzhaki et al. 1995). This conjecture is in line with the observed folding/unfolding data from 108 proteins (70 showing two-state kinetics) that demonstrate that the logarithm of the folding and unfolding rates is well correlated ($R \sim 0.8$) and that such correlation is better for two than that for multiple-state proteins (Glyakina and Galzitskaya, 2020)

4. The largest size of the Gibbs free-energy barrier (ΔG) between the native state and the highest point of the free-energy profile (see Fig. 1) is assumed to be given by the protein marginal-stability upper bound limit, i.e., $\Delta G \sim 7.4$ kcal/mol, which (i) is a universal feature of proteins, i.e., was obtained regardless of their sequence or length (Vila 2019; Vila 2021); (ii) is a consequence of Anfinsen's dogma validity (Vila 2019; 2021); and (iii) represents a threshold beyond which a conformation will unfold and become non-functional (Martin and Vila 2020; Vila 2021; 2022)
5. The word "mutation" usually refers to an amino-acid substitution in the protein sequence as a result of a nucleotide pair replacement (Kimura 1968). This is a very well-known phenomenon in the protein folding/unfolding field because it could alter protein stability (Privalov and Tsalkova, 1979; Tokuriki et al. 2008; Tokuriki and Tawfik 2009; Socha and Tokuriki 2013; Martin and Vila 2020), structure (Koehl and Levitt 2002), function (Tokuriki et al. 2008; Otwinowski 2018), and evolvability (Kimura 1968; Bloom et al. 2006; Kurahashi et al. 2018; Vila 2022) through a variety of mechanisms. As such, there has been considerable interest in understanding the structural and energetic consequences of such amino acid substitutions. Interestingly, an alteration that also has a significant impact on the protein structure, stability, and function occurs through post-translational modifications (PTMs), a phenomenon that refers to an amino acid side-chain modification in some proteins after their biosynthesis. In this regard, it is worth noting the existence of more than 400 types of PTMs, among which phosphorylation, acetylation, methylation, and glycosylation, are the most common (Khoury et al. 2011). Notably, N-linked glycoproteins, which are the result of a reversible enzyme-directed reaction, are a particularly interesting case of PTM since more than 50% of all eukaryotic are glycoproteins (Shental-Bechor and Levy 2008; Ellis et al. 2012), and hence, there is considerable interest in predicting the structural and functional consequences of such site-specific modifications (Chen et al. 2010; Garay et al. 2016; Ramazi and Zahiri 2021; Weaver et al. 2022). PTMs are particularly relevant to biology because they increase proteomic diversity by several orders of magnitude (Spoel 2018). All of this enables us to conjecture that each PTM could be thought of as a different amino acid from the 20 naturally occurring ones. Then, unless otherwise noted, the word "mutation" will merely refer to a protein sequence modification, and, thus, its effects on the protein structure, stability, and foldability rate will be analyzed without making any distinction among these phenomena
6. It is assumed that point mutations mainly affect the native-state stability (Zeldovich et al. 2007). This assumption is equivalent to assuming an average ϕ-value—a technique commonly used to examine the kinetic effects on the protein folding upon a point mutation (Matouschek et al. 1989; Itzhaki et al. 1995; Campos 2022)—closer to ~ 0 than to ~ 1. In line with this, the average ϕ-value—of more than 800 mutations for 24 two-state proteins—is $<\phi> \sim 0.24$ (Naganathan and Muñoz 2010)
7. The unfolding Gibbs free energy (ΔG_U) between the wild-type (*wt*) and the mutant (*m*) protein can be effortlessly computed as $\Delta\Delta G_U = (\Delta G_U^m - \Delta G_U^{wt})$ (Bigman and Levy 2018). This definition—together with assumption 6—enables us to propose (Vila 2022) a reasonable strategy to assess the change in the protein marginal stability upon point mutations ($\Delta\Delta G$), namely, as $\Delta\Delta G \sim \Delta\Delta G_U$
8. The best candidates for simulations of all-atom molecular dynamics are proteins that fold at or close to the speed limit, simply because such simulations are computationally intensive. This has inspired experimentalists to look for proteins that fold rapidly as well as to develop other proteins that fold even more quickly. For this reason, the folding speed limit (τ_0) of two-state proteins (the barrier-less limit) has been discussed at great length in the literature (Zana 1975; McCammon 1996; Hagen et al. 1996; Mayor et al. 2000; Krieger et al. 2003; Yang and Gruebele 2003; Akmal and Muñoz 2004; Muñoz et al. 2008; Ivankov and Finkelstein 2020; Glyakina and Galzitskaya 2020; Muñoz and Cerminara 2016; Chung and Eaton 2018; Eaton 2021), and there is a consensus that it should be within the following range of values

$$\sim 10^{-8}\ [\text{sec}] < \tau_0 < \sim 10^{-5}\ [\text{sec}] \quad (1)$$

Let us quickly show how these constraints on the folding rate impact the ability of proteins to evolve. If a given 100-residue two-state protein cannot fold faster than $\tau_0 \sim 10^{-8}$ (or $\sim 10^{-5}$) seconds, and if life began on earth around a billion ($\sim 10^9$) years ago, its protein space

size (Maynard Smith 1970) would contain at most ~ 10^{24} (or ~ 10^{21}) sequences. If this were the case, the average mutation rate per amino acid (ξ) should be ≤ ~ 1.74 (or ≤ ~ 1.62) since ξ must satisfy $\xi^{100} = \sim 10^{24}$ (or ~ 10^{21}). The fact that $\xi < 2$ is of paramount importance from an evolutive point of view because it means that only a fraction of a given protein sequence is available for an amino acid substitution at any one time, in agreement with both previous estimations of the protein space size (Vila 2020) and existent pieces of evidence (Margoliash and Smith 1965; Sarkisyan et al 2016). From an evolutionary perspective, an in-depth discussion of an accurate estimation of the protein space size in light of the factors that govern it is of utmost importance (Mandecki 1998; Dryden et al. 2008; Romero and Arnold 2009; Ivankov 2017), as well as it is of practical interest for studies of directed evolution (Arnold, 2009).

9. The time (τ) to overcome the free-energy barrier ΔG (shown in Fig. 1) may be computed by using an argument from the transition state theory (Ivankov and Finkelstein 2020) as

$$\tau = \tau_0 \, exp(\beta \Delta G) \text{ [sec]} \quad (2)$$

in which the lower and upper bound of the pre-exponential factor (τ_0) is given in Eq. (1), $\beta = 1/RT$, R is the gas constant and T is the absolute temperature (298 K for all the calculations). If the free energy barrier vanishes ($\Delta G \sim 0$), a downhill, barrierless, or one-state unfolding (Garcia-Mira et al., 2002; Naganathan et al. 2005; Muñoz et al. 2008) occurs in times given by τ_0.

10. After assuming the validity of all of the above conjectures and facts, it is possible to determine the following range of τ_{max} values from Eq. (2) (with $\Delta G \sim 7.4$ kcal/mol and τ_0 given by Eq. 1)

$$\sim 10^{-3} \text{ [sec]} \leq \tau_{max} \leq \sim 1 \text{ [sec]} \quad (3)$$

The results of simulations on the protein folding (Sali et al. 1994; Karplus 1997; Lindorff-Larsen et al. 2011) and the observed folding rates for 65 two-state proteins that fold in an aqueous solution under biological conditions (Garbuzynskiy et al. 2013; Ivankov and Finkelstein 2020) attest that this time window for the slowest folding rate, τ_{max}, is acceptable from a biological point of view. This result is a consequence of the fact that there is an upper bound on the marginal stability of proteins (~ 7.4 kcal/mol), which seems to be a universal property of biomolecules and macromolecular complexes (Martin and Vila 2020; Vila 2021) and arises from the validity of Anfinsen's dogma (Vila 2019, 2021; Martin and Vila 2020).

Overall, the range of variation for τ_{max} shown in Eq. (3) for a two-state protein (i) does not depend on the chain length, which is consistent with the observation that chain length has a nearly null correlation ($R \sim -0.07$) with the folding time logarithm (Plaxco et al. 2000; Galzitskaya et al. 2003), (ii) provides the answer to the central question of Levinthal's paradox's of how long it takes for a protein to reach its native state, and (iii) is a standard that will allow us to evaluate the impact of amino acid substitutions and/or post-translational modifications on the rates of protein folding, which we will examine in the next section.

II.- Evolution of protein folding rate in light of mutations

If the free-energy barrier height (ΔG) rules the unfolding (and folding) time τ for a two-state protein; then, a single-point mutation could affect it by either increasing (stabilizing) or decreasing (destabilizing) the marginal stability. Let us start by examining the physics that rules the phenomenon of protein folding time changes upon mutations. The ratio between the wild-type protein folding time (τ_{wt}) and that of this protein upon a point mutation (τ_m) can be computed—after assuming that τ_0 is insensitive to mutations (Socci et al. 1996; Muñoz and Eaton 1999)—using Eq. (2) as (Chaudhary et al. 2015; Ivankov and Finkelstein 2020)

$$\Delta\tau_m = (\tau_m/\tau_{wt}) \sim exp(\beta \Delta\Delta G_m) \Rightarrow RT \, ln\Delta\tau_m \sim \Delta\Delta G_m \quad (4)$$

where $\Delta\Delta G_m = (\Delta G_m - \Delta G_{wt}) \sim \Delta\Delta G_U$ is the change, upon a single-point mutation, between the mutant and the wild-type Gibbs free-energy gap (ΔG), respectively. The key takeaway from this analysis is that the protein marginal-stability change upon a mutation ($\Delta\Delta G_m$) provides the necessary and sufficient information to accurately estimate, via a Boltzmann factor, the evolution of the folding rates ($\Delta\tau_m$). The physics underpinning this conclusion follows. Mutations affect, mainly, the stability of the native state (Zeldovich et al. 2007) and, to a lesser extent, the ensemble of high-energy native-like structures that coexist with it, i.e., the transition state ensemble, shown in Fig. 1). This hypothesis is supported by convincing theoretical simulations of the amide hydrogen exchange mechanism on proteins (Vendruscolo et al. 2003), as well as the results of a high-resolution structure determination method indicating that high-energy native-like structures may be required for protein function (Stiller et al. 2022).

Since $\Delta\Delta G_m$ is a state function, Eq. 4 will be valid for any number of j (≥ 2) consecutive mutations and, hence, it can be generalized straightforwardly by replacing $m \rightarrow j$ because $(\Delta G_1 - \Delta G_{wt}) + \sum_{k=2}^{j} (\Delta G_k - \Delta G_{k-1}) = (\Delta G_j - \Delta G_{wt}) = \Delta\Delta G_j$. This generalization is particularly relevant to determine the evolution of folding rates upon mutations because many forms

of post-translational modifications may occur in tandem (Khoury et al. 2011). Additionally, this property of $\Delta\Delta G_m$ should have a profound impact on evolutionary biology research. To illustrate this, imagine evolution as a walk across Protein Space, i.e., one where "…functional proteins must form a continuous network which can be traversed by unit mutational steps without passing through nonfunctional intermediates…" (Maynard Smith 1970). Then, if we assign a "fitness value" to each functional protein in that sequence space, which is a measurement of how effectively each protein may perform an expected function (Romero and Arnold 2009), it becomes clear that starting from an arbitrary functional protein, nature can follow any mutational trajectory to achieve a specified "fitness target" (see Fig. 2), if there is no penalty for doing so (Weinreich et al. 2006). This simple illustration shows that it is not necessary to predict mutational trajectories (Sailer and Harms 2017a, b) or account for epistasis effects (Breen and et al. 2012; Starr and Thornton 2016; Miton and Tokuriki 2016; Sailer and Harms 2017a, b; Sailer and Harms 2017b; Domingo et al. 2019; Vila 2022), a phenomenon which occurs when the total effect of two or more mutations is different from the sum of those effects, to determine the evolution of the folding rate. However, if a particular mutational trajectory has a higher probability than all the others, epistasis effect considerations (Romero and Arnold 2009; Sailer and Harms 2017b) may be crucial to understanding the reason for such a preference. Overall, Eq. (4) enables us to calculate the evolution of the protein folding rate after *j* consecutive mutations, regardless of the paths that evolution takes in the protein space, as follows:

$$\tau_j \sim \tau_{wt}\, e^{\beta\Delta\Delta G_j} \tag{5}$$

In light of all of this, directed evolution studies (Arnold, 2009; Romero and Arnold 2009; Socha and Tokuriki 2013) that look for protein sequences that carry out a desired function in a specific amount of time would undoubtedly benefit from knowing all of the parameters influencing protein evolvability, especially those governing changes in folding rates as a result of mutations.

In general, the above analysis confirms that evolution influences, through mutations, unfolding/folding time scales by altering the height and composition of the energetic barrier but not their rate-limiting step set by the physics of folding, namely, by the largest-possible change in the free energy barrier ($|\Delta\Delta G| < \sim 7.4$ kcal/mol). From a thermodynamic standpoint, this barrier defines a threshold beyond which a two-state protein unfolds or becomes non-functional and, from a kinetic viewpoint, the time ceiling for the unfolding process.

In the following subsections, the magnitude of the τ_j changes upon mutations will be illustrated by using data from (a) post-translational modifications (PTMs) and (b) amino acid substitutions.

(a) Post-translational modifications

Among all possible PTMs effects, we choose N-linked glycosylation—a covalent attachment of carbohydrate to certain residues of a protein—because it is, on the one hand, one of the most common PTM in eukaryotes (Shental-Bechor and Levy 2008; Hanson et al. 2009; Ellis et al. 2012) and, on the other, one for which detailed information of its effects on two-state protein folding rates is well documented (Hanson et al. 2009). Thus, a study on the observed folding energetics of the mono-N-glycosylated adhesion domain of the human immune cell receptor cluster of differentiation 2 (hCD2ad)—a protein with a β-sandwich topology—reveals that the N-glycan first saccharide unit is responsible for a stabilization-free-energy ($\Delta\Delta G_1$) of ~ 2.3 kcal/mol and a 50-fold rate slower than that of the nonglycosylated protein (Hanson et al. 2009). This observed folding rate change upon a PTM

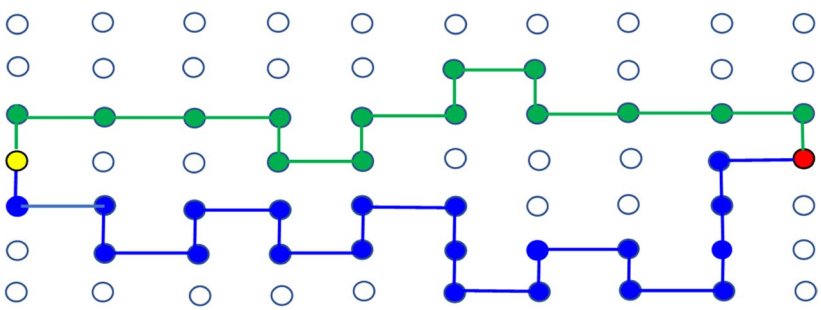

Fig. 2 Cartoon of the Protein Space (Maynard Smith, 1970) as a model of evolution, where each circle represents a functional protein that differs from any of its nearest neighbors by one amino acid. The yellow- and red-filled circles—the wild-type (wt) and the target-sequence (ts), respectively—represent the starting and finishing functional proteins of any possible mutational trajectory. The green- and blue-filled circles illustrate two arbitrary mutational trajectories, each representing a walk in the protein space of 15 and 21 mutational steps (amino-acid substitutions), respectively. Then, for any mutational trajectory, the following relation for the protein marginal-stability evolution holds $\Delta\Delta G = (\Delta G_{ts} - \Delta G_{wt})$. Consequently, $\tau_{ts} \sim \tau_{wt}\, e^{\beta\Delta\Delta G}$, with $\beta = 1/RT$ (see text for details)

is fully consistent with Eq. (5) which—for such change on the protein marginal stability—predicts an unfolding rate of $\tau_1 \sim 48$-fold slower than that of the nonglycosylated wild-type (τ_{wt}) at room temperature (298 K). We focused on the analysis of the effect of the first N-linked glycan because it affects the thermodynamics and kinetics of the protein folding by 65% out of 100% of the total N-glycan contributions to HCD2ad (Hanson et al. 2009).

It is worth noting that glycosylation does not always lead to a more stable protein structure. Indeed, there is evidence that the contrary occurs for O-glycosylation of the serum vitamin D binding protein for which each event destabilizes the protein by ~ -1 kcal/mol (Spiriti et al. 2008). Then, the unfolding speed will be ~ fivefold faster than that of the nonglycosylated one (τ_{wt}), as indicated by Eq. (5).

(b) Amino acid substitutions

An analysis to determine the magnitude of the τ_j changes upon amino acid substitution is, actually, unnecessary because the existence of large databases providing detailed information on the changes in protein stability upon single-point mutations makes their computation trivial. Indeed, ThermoMutDB (Xavier et al. 2021) is a manually curated database containing ~ 8,800 entries that collect experimental information on the effect of single-point mutations on protein stability ($\Delta\Delta G_1$), together with available experimental structural information. Then, the corresponding values for ln ($\Delta\tau_1$) or τ_1 can be straightforwardly computed by using Eqs. (4) and (5), respectively. At this point, it is worth noting that the ThermoMutDB database contains nearly all (~98%) single-point mutated proteins whose report $|\Delta\Delta G_1|$ values are $\leq \sim 7.4$ kcal/mol, confirming the hypothesis that protein marginal stability cannot exceed this threshold (Vila 2019, 2021).

Conclusions

The analysis has made it possible for us to find a straightforward answer to a key question that sits at the heart of Levinthal's paradox: how long does it take for a protein to achieve its native state? As proved, it takes seconds—not years, as suggested by a naïve solution to the dilemma—for a two-state protein of any sequence or length to acquire its native state. Also, it helped us to comprehend why proteins reach their native state within a biologically acceptable timeframe, specifically because the largest-possible change in the two-state protein free-energy barrier (~ 7.4 kcal/mol) is a consequence of the validity of the thermodynamic hypothesis—or Anfinsen's dogma—a limit set by the physics of folding. Furthermore, we have shown that the evolution of protein folding rates is primarily driven by changes in the marginal stability of proteins caused by amino acid substitutions and/or post-translational modifications. This dependence ensures that, given a starting and a target sequence, whatever the mutational paths in sequence space or epistasis effects are, they will not have an impact on the determination of the evolution of the protein folding rate. This is an important result since the evolutionary trajectories are unpredictable, and the estimation of the epistasis effects is a daunting task. Moreover, if folding/unfolding speed becomes a bottleneck in the search for new proteins and functions, the prediction of the folding rate becomes important, and all factors influencing it should be thoroughly investigated. The analysis offered from this point of view may well be a good place to start.

Overall, this review focuses on protein sequence changes caused by mutations—amino-acid substitutions and/or post-translational modifications—and their impact on protein folding rates, a phenomenon closely related to one of the most important unanswered questions in structural biology: how a sequence encodes its folding. In this regard, we have learned that some properties of two-state proteins, such as their slowest folding time, are sequence-independent. As already explained, this is a consequence of a universal feature of proteins, namely, the existence of a marginal-stability upper bound limit beyond which the protein unfolds or becomes non-functional. Then, all biologically relevant processes must take place under this stability threshold and, hence, are sequence-dependent, since the latter determines the tridimensional structure of proteins, which in turn regulates its function. Therefore, finding a solution to the abovementioned question becomes critical and highly relevant in this context, as state-of-the-art numerical methods have so far been unable to solve it. The current study, we firmly believe, will encourage researchers to continue looking for solutions to this and other unsolved structural biology problems.

Acknowledgements The author acknowledges support from the IMASL-CONICET-UNSL.

Funding This work was funded by the Agencia Nacional de Promoción de la Investigación, el Desarrollo Tecnológico y la Innovación (PICT-02212), Argentina.

Data availability Not applicable.

Code of availability Not applicable.

Declarations

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Conflict of interest The author declares no competing interest.

References

Akmal A, Muñoz V (2004) The nature of the free energy barriers to two-state folding. Proteins 57:142–152. https://doi.org/10.1002/prot.20172

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

Arnold FH (2009) How proteins adapt: lessons from directed evolution. Cold Spring Harbor Symposia Quantitative Biology 74:41–46. https://doi.org/10.1101/sqb.2009.74.046

Ben-Naim A (2012) Levinthal's paradox revisited, and dismissed. Open J Biophys 2:23–32. https://doi.org/10.4236/ojbiphy.2012.22004

Bigman LS, Levy Y (2018) Stability effects of protein mutations: the role of long-range contacts. J Phys Chem B 122:11450–11459. https://doi.org/10.1021/acs.jpcb.8b07379

Bloom JD, Labthavikul ST, Otey CR (2006) Arnold FH (2006) Protein stability promotes evolvability. Proc Natl Acad Sci USA 103:5869–5874. https://doi.org/10.1073/pnas.0510098103

Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA (2012) Epistasis as the primary factor in molecular evolution. Nature 490:535–538. https://doi.org/10.1038/nature11510

Campos LA (2022) Mutational analysis of protein folding transition states: phi values. Methods Mol Biol 2376:3–30. https://doi.org/10.1007/978-1-0716-1716-8_1

Chaudhary P, Naganathan AN, Gromiha M (2015) Folding RaCe: a robust method for predicting changes in protein folding rates upon point mutations. Bioinformatics 31:2091–2097. https://doi.org/10.1093/bioinformatics/btv091

Chen MM, Bartlett AI, Nerenberg PS, Friel CT, Hackenberger CP, Stultz CM, Radford SE, Imperiali B (2010) Perturbing the folding energy landscape of the bacterial immunity protein Im7 by site-specific N-linked glycosylation. Proc Natl Acad Sci USA 107:22528–22533. https://doi.org/10.1073/pnas.1015356107

Chung HS, Eaton WA (2018) Protein folding transition path times from single molecule FRET. Curr Opin Struct Biol 48:30–39. https://doi.org/10.1016/j.sbi.2017.10.007

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

Dinner AR, Karplus M (2001) The roles of stability and contact order in determining protein folding rates. Nat Struct Biol 8:21–22. https://doi.org/10.1038/83003

Domingo J, Baeza-Centurion P, Lehner B (2019) The causes and consequences of genetic interactions (epistasis). Annu Rev Genomics Hum Genet 20:433–460. https://doi.org/10.1146/annurev-genom-083118-014857

Dryden DTF, Thomson AR, White JH (2008) How much of protein sequence space has been explored by life on Earth? J R Soc Interface 5:953–956. https://doi.org/10.1098/rsif.2008.0085

Eaton WA (2021) Modern kinetics and mechanism of protein folding: a retrospective. J Phys Chem B 125:3452–3467. https://doi.org/10.10021/acs.jpcb.1c0206

Ellis CR, Maiti B, Noid WG (2012) Specific and nonspecific effects of glycosylation. J Am Chem Soc 134:8184–8193. https://doi.org/10.1021/ja301005f

Englander SW (2000) Protein folding intermediates and pathways studied by hydrogen exchange. Annu Rev Biophys Biomol Struct 29:213–238. https://doi.org/10.1146/annurev.biophys.29.1.213

Englander SW, Mayne L (2014) The nature of protein folding pathways. Proc Natl Acad Sci USA 111:15873–15880. https://doi.org/10.1073/pnas.1411798111

Fersht AR, Daggett V (2002) Protein folding and unfolding at atomic resolution. Cell 108:573–582. https://doi.org/10.1016/S0092-8674(02)00620-7

Finkelstein AV, Badretdinov AY (1997) Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. Fold Des 2:115–121.https://doi.org/10.1016/S1359-0278(97)00016-3

Finkelstein AV, Garbuzynskiy SO (2013) Levinthal's question answered … again? J Biomol Struct Dyn 31:1013–1015. https://doi.org/10.1080/07391102.2012.748544

Finkelstein AV, Bogatyreva NS, Ivankov DN et al (2022) Protein folding problem: enigma, paradox, solution. Biophys Rev 14:1255–1272. https://doi.org/10.1007/s12551-022-01000-1

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

Garay PG, Martin OA, Scheraga HA, Vila JA (2016) Detection of methylation, acetylation, and glycosylation of protein residues by monitoring ^{13}C chemical-shift changes: a quantum-chemical study. PeerJ 4:e2253. https://doi.org/10.7717/peerj.2253

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

Garcia-Mira MM, Sadqi M, Fischer N, Sanchez-Ruiz JM, Muñoz V (2002) Experimental identification of downhill protein folding. Science 298:2191–2195. https://doi.org/10.1126/science.1077809

Glyakina AV, Galzitskaya OV (2020) How quickly do proteins fold and unfold, and what structural parameters correlate with these values? Biomolecules 10:197. https://doi.org/10.3390/biom10020197

Gutin AM, Abkevich VI, Shakhnovich EI (1996) Chain length scaling of protein folding time. Phys Rev Lett 77:5433–5436. https://doi.org/10.1103/PhysRevLett.77.5433

Hagen SJ, Hofrichter J, Szabo A, Eaton WA (1996) Diffusion-limited contact formation in unfolded cytochrome c: estimating the maximum rate of protein folding. Proc Natl Acad Sci U S A 93:11615–11617. https://doi.org/10.1073/pnas.93.21.11615

Hanson SR, Culyba EK, Hsu TL, Wong CH, Kelly JW, Powers ET (2009) The core trisaccharide of an N-linked glycoprotein intrinsically accelerates folding and enhances stability. Proc Natl Acad Sci USA 106:3131–3136. https://doi.org/10.1073/pnas.0810318105

Itzhaki LS, Otzen DE, Fersht AR (1995) The structure of the transition state for folding of chymotrypsin inhibitor 2 analyzed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. J Mol Biol 254:260–288. https://doi.org/10.1006/jmbi.1995.0616

Ivankov DN (2017) Exact correspondence between walk in nucleotide and protein sequence spaces. PLoS ONE 12:e0182525. https://doi.org/10.1371/journal.pone.0182525

Ivankov DN, Finkelstein AV (2020) Solution of Levinthal's paradox and a physical theory of protein folding times. Biomolecules 10:250. https://doi.org/10.3390/biom10020250

Jackson SE (1998) How do small single-domain proteins fold? Fold Des 3:R81-91. https://doi.org/10.1016/S1359-0278(98)00033-9

Karplus M (1997) The Levinthal paradox: yesterday and today. Fold Des 2:S69–S75. https://doi.org/10.1016/S1359-0278(97)00067-9

Khoury GA, Baliban RC, Floudas CA (2011) Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. Sci Rep 13(1):90. https://doi.org/10.1038/srep00090

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626. https://doi.org/10.1038/217624a0

Koehl P, Levitt M (2002) Protein topology and stability define the space of allowed sequences. Proc Natl Acad Sci USA 99:1280–1285. https://doi.org/10.1073/pnas.032405199

Krieger F, Fierz B, Bieri O, Drewello M, Kiefhaber T (2003) Dynamics of unfolded polypeptide chains as a model for the earliest steps in protein folding. J Mol Biol 332:265–274. https://doi.org/10.1016/S0022-2836(03)00892-1

Kurahashi R, Sano S, Takano K (2018) Protein evolution is potentially governed by protein stability: directed evolution of an esterase from the hyperthermophilic archaeon Sulfolobus tokodaii. J Mol Evol 86:283–292. https://doi.org/10.1007/s00239-018-9843-y

Kuwajima K (2020) The molten globule, and two-state vs. non-two-state folding of globular proteins. Biomolecules 10:407. https://doi.org/10.3390/biom10030407

Lazaridis T, Karplus M (1997) "New view" of protein folding reconciled with the old through multiple unfolding simulations. Science 278:1928–1931. https://doi.org/10.1126/science.278.5345.1928

Levinthal C (1968) Are there pathways for protein folding? Journal De Chimie Physique 65:44–45. https://doi.org/10.1051/jcp/04419 68650

Li Y, Gong H (2022) Identifying a feasible transition pathway between two conformational states for a protein. J Chem Theory Comput 18:4529–4543. https://doi.org/10.1021/acs.jctc.2c00390

Lindorff-Larsen K, Piana S, Dror RO, Shaw DE (2011) How fast-folding proteins fold. Science 334:517–520. https://doi.org/10.1126/science.1208351

Mandecki W (1998) The game of chess and searches in protein sequence space. Trends Biotechnol 16:200–202. https://doi.org/10.1016/S0167-7799(98)01188-3

Margoliash E, Smith EL. (1965) Structural and Functional Aspects of Cytochrome c in Relation to Evolution, pp. 221-242, in Evolving genes and proteins: a symposium held at the Institute of Microbiology of Rutgers, with support from the National Science Foundation. Edited by Vernon Bryson and Henry J. Vogel. New York; London: Academic Press. https://wellcomecollection.org/works/bxp4duyj/items?canvas=7. https://doi.org/10.1016/B978-1-4832-2734-4.50023-1

Martin AO, Vila JA (2020) The marginal stability of proteins: how the jiggling and wiggling of atoms is connected to neutral evolution. J Mol Evol 88:424–426. https://doi.org/10.1007/s00239-020-09940-6

Martinez L (2014) Introducing the Levinthal's protein folding paradox and its solution. J Chem Educ 91:1918–1923. https://doi.org/10.1021/ed300302h

Matouschek A, Kellis JT Jr, Serrano L, Fersht AR (1989) Mapping the transition state and pathway of protein folding by protein engineering. Nature 340:122–126. https://doi.org/10.1038/340122a0

Maynard Smith J (1970) Natural selection and the concept of a protein space. Nature 225:563–564. https://doi.org/10.1038/225563a0

Mayor U, Johnson CM, Daggett V, Fersht AR (2000) Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. Proc Natl Acad Sci USA 97:13518–13522. https://doi.org/10.1073/pnas.250473497

McCammon JA (1996) A speed limit for protein folding. Proc Natl Acad Sci U S A 93:11426–11427. https://doi.org/10.1073/pnas.93.21.11426

Miton CM, Tokuriki N (2016) How mutational epistasis impairs predictability in protein evolution and design. Protein Sci 25:1260–1272. https://doi.org/10.1002/pro.2876

Muñoz V, Cerminara M (2016) When fast is better: protein folding fundamentals and mechanisms from ultrafast approaches. Biochem J 473:2545–2559. https://doi.org/10.1042/BCJ20160107

Muñoz V, Eaton WA (1999) A simple model for calculating the kinetics of protein folding from three-dimensional structures. Proc Natl Acad Sci USA 96:11311–11316. https://doi.org/10.1073/pnas.96.20.11311

Muñoz V, Sadqi M, Naganathan AN, de Sancho D (2008) Exploiting the downhill folding regime via experiment. HFSP J 2:342–353. https://doi.org/10.2976/1.2988030

Naganathan AN, Muñoz V (2010) Insights into protein folding mechanisms from large scale analysis of mutational effects. Proc Natl Acad Sci USA 107:8611–8616. https://doi.org/10.1073/pnas.1000988107

Naganathan AN, Perez-Jimenez R, Sanchez-Ruiz JM, Muñoz V (2005) Robustness of downhill folding: guidelines for the analysis of equilibrium folding experiments on small proteins. Biochemistry 44:7435–7449. https://doi.org/10.1021/bi050118y

Otwinowski J (2018) Biophysical inference of epistasis and the effects of mutations on protein stability and function. Mol Biol Evol 35:2345–2354. https://doi.org/10.1093/molbev/msy141

Plaxco KW, Simons KT, Ruczinski I, Baker D (2000) Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics. Biochemistry 39:11177–11183.https://doi.org/10.1021/bi000200n

Privalov PL (1979) Stability of proteins: small globular proteins. Adv Protein Chem 33:167–241. https://doi.org/10.1016/S0065-3233(08)60460-X

Privalov PL, Tsalkova TN (1979) Micro- and macro-stabilities of globular proteins. Nature 280:694–696. https://doi.org/10.1038/280693a0

Ramazi S, Zahiri J (2021) Posttranslational modifications in proteins: resources, tools and prediction methods. Database (Oxford). 2021:baab012. https://doi.org/10.1093/database/baab012

Romero PA, Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol 10:866–876. https://doi.org/10.1038/nrm2805

Rooman M, Dehouck Y, Kwasigroch JM, Biot C, Gilis D (2002) What is paradoxical about Levinthal paradox? J Biomol Struct Dyn 20:327–329. https://doi.org/10.1080/07391102.2002.10506850

Sailer ZR, Harms MJ (2017) Molecular ensembles make evolution unpredictable. Proc Natl Acad Sci USA 114:11938–11943. https://doi.org/10.1073/pnas.1711927114

Sailer ZR, Harms MJ (2017) High-order epistasis shapes evolutionary trajectories. PLoS Comput Biol 13(5):e1005541. https://doi.org/10.1371/journal.pcbi.1005541

Sali A, Shakhnovich E, Karplus M (1994) How does a protein fold? Nature 369:248–251. https://doi.org/10.1038/369248a0

Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV et al (2016) Local fitness landscape of the green fluorescent protein. Nature 533:397–401. https://doi.org/10.1038/nature17995

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

Shental-Bechor D, Levy Y (2008) Effect of glycosylation on protein folding: a close look at thermodynamic stabilization. Proc Natl Acad Sci USA 105:8256–8261. https://doi.org/10.1073/pnas.0801340105

Socci ND, Onuchic JN, Wolynes PG (1996) Diffusive dynamics of the reaction coordinate for protein folding funnels. J Chem Phys 104:5860–5868. https://doi.org/10.1063/1.471317

Socha RD, Tokuriki N (2013) Modulating protein stability - directed evolution strategies for improved protein function. FEBS J 280:5582–5595. https://doi.org/10.1111/febs.12354

Spiriti J, Bogani F, van der Vaart A, Ghirlanda G (2008) Modulation of protein stability by O-glycosylation in a designed Gc-MAF analog. Biophys Chem 134:157–167. https://doi.org/10.1016/j.bpc.2008.02.005

Spoel SH (2018) Orchestrating the proteome with post-translational modifications. J Exp Bot 69:4499–4503. https://doi.org/10.1093/jxb/ery295

Starr TN, Thornton JW (2016) Epistasis in protein evolution. Protein Sci 25:1204–1218. https://doi.org/10.1002/pro.2897
Stiller JB et al (2022) Structure determination of high-energy states in a dynamic protein ensemble. Nature 603(7901):528–535. https://doi.org/10.1038/s41586-022-04468-9
Thirumalai D (1995) From minimal models to real proteins: time-scales for protein folding. J Phys I (france) 5:1457–1467. https://doi.org/10.1051/jp1:1995209
Tokuriki N, Tawfik DS (2009) Protein dynamics and evolvability. Science 324:203–207. https://doi.org/10.1126/science.1169375
Tokuriki N, Stricher F, Serrano L, Tawfik DS (2008) How protein stability and new functions trade off. PLoS Comput Biol 4(2):e1000002. https://doi.org/10.1371/journal.pcbi.1000002
Vendruscolo M, Paci E, Dobson CM, Karplus M (2003) Rare fluctuations of native proteins sampled by equilibrium hydrogen exchange. J Am Chem Soc 125(51):15686–15687. https://doi.org/10.1021/ja036523z
Vila JA (2019) Forecasting the upper bound free energy difference between protein native-like structures. Physica A 533:122053. https://doi.org/10.1016/j.physa.2019.122053
Vila JA (2020) About the protein space vastness. Protein J 39:472–475. https://doi.org/10.1007/s10930-020-09939-4
Vila JA (2021) Thoughts on the protein's native state. J Phys Chem Lett 12:5963–5966. https://doi.org/10.1021/acs.jpclett.1c01831
Vila JA (2022) Proteins' evolution upon point mutations. ACS Omega 7:14371–14376. https://doi.org/10.1021/acsomega.2c01407
Weaver GC, Arya R, Schneider CL, Hudson AW, Stern LJ (2022) Structural models for roseolovirus U20 and U21: non-classical MHC-I like proteins from HHV-6A, HHV-6B, and HHV-7. Front Immunol 13:864898. https://doi.org/10.3389/fimmu.2022.864898
Weinreich DM, Delaney NF, DePristo MA, Hartl DL (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 312:111–114. https://doi.org/10.1126/science.1123539
Wolynes PG (1997) Folding funnels and energy landscapes of larger proteins within the capillarity approximation. Proc Natl Acad Sci USA 94:6170–6175. https://doi.org/10.1073/pnas.94.12.6170
Wolynes PG (2015) Evolution, energy landscapes and the paradoxes of protein folding. Biochimie 119:218–230. https://doi.org/10.1016/j.biochi.2014.12.007
Wolynes PG, Onuchic JN, Thirumalai D (1995) Navigating the folding routes. Science 267:1619–1620. https://doi.org/10.1126/science.7886447
Xavier JS, Nguyen TB, Karmarkar M, Portelli S, Rezende PM, Velloso JPL, Ascher DB, Pires DEV (2021) ThermoMutDB: a thermodynamic database for missense mutations. Nucleic Acids Res 49:D475–D479. https://doi.org/10.1093/nar/gkaa925
Yang WY, Gruebele M (2003) Folding at the speed limit. Nature 423:193–197. https://doi.org/10.1038/nature01609
Zana R (1975) On the rate-determining step for helix propagation in the helix–coil transition of polypeptides in solution. Biopolymers 14:2425–2428. https://doi.org/10.1002/bip.1975.360141116
Zeldovich KB, Chen P, Shakhnovich EI (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. Proc Natl Acad Sci USA 104:16152–16157. https://doi.org/10.1073/pnas.0705366104
Zwanzig R, Szabo A, Bagchi B (1992) Levinthal's paradox. Proc Natl Acad Sci USA 89:20–22. https://doi.org/10.1073/pnas.89.1.20