Superfunneled Energy Landscape of Protein Evolution Unifies the Principles of Protein Evolution, Folding, and Design

Zhiqiang Yan and Jin Wang

Phys. Rev. Lett. **122**, 018103 — Published 9 January 2019

DOI: 10.1103/PhysRevLett.122.018103

Super funneled energy landscape of protein evolution unifies the principles of protein evolution, folding and design

Zhiqiang Yan[1] and Jin Wang[1,2,*]
[1] *State Key Laboratory of Electroanalytical Chemistry, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun, Jilin 130022, China*
[2] *Department of Chemistry & Physics, State University of New York at Stony Brook, Stony Brook, NY 11790, USA*

(Dated: December 17, 2018)

Evolution is essential for shaping the biological functions. Darwin proposed the selection as the driving force for evolution upon mutations. While mutations are clear, the quantification of the selection force is still challenging. In this study, we identified and quantified both thermodynamic stability and kinetic accessibility as the selection forces for protein evolution. The protein evolution can be viewed and quantified as a trajectory moving along a super funneled energy landscape with a line attractor at the bottom. The resulting evolved sequences and structures show strong protein characteristics including the hydrophobic core, high designability and fast folding. The evolution principle uncovered here is validated on real proteins and shed light on the protein design.

Evolution is essential for shaping the biological function. Darwin proposed the selection or fitness as the driving force for evolution upon mutations. While the mutations are understood reasonably well, the quantification of the selection force is still challenging. At the molecular level, it is still unclear what the dominant selection fitness are and how they are selected from both the structures and sequences. Protein evolution works both by selection and random mutation [1]. Persistent occurrence of protein mutations provides opportunities for proteins to be improved over the course of evolution and even so at the present stage of evolution. It has long been realized that the naturally occurring proteins in living organisms belong to a small ensemble of sequences distinctly different from the random sequences [2]. The proteins typically have a high degree of thermodynamic, kinetic and structural specificities different from random heteropolymers of amino acids [3–5]. Funneled folding energy landscape has been suggested for the explanation of the folding of natural sequences with minimal frustration, in contrast to the random sequences which are frustrated in their low energy conformations [5–8]. This answers the question of how the natural sequence searches its native conformation in the structure space. One could also ask how the proteins found in nature are selected not only from the vast structure space but also from even larger sequence space (Fig. 1) as a consequence of evolution. This raises another paradox on protein evolution similar as the Levinthal's paradox on protein folding [4], which is the seemly infinite time for searching through the vast sequence space in contrast to the finite time of protein evolution for function.

*Electronic address: jin.wang.1@stonybrook.edu

Previously, the evolutionary mechanisms have been explored by employing different fitness criteria to trace the evolutionary process [9–14]. However, the quantification of the dominant selection pressure for protein evolution is still challenging. Previous protein evolution studies explored the search in sequence space based on a fixed target structure. However, the structural transition can occur with even a single point mutation in the evolution process [15, 16]. Therefore, the protein evolution should be considered as the search in both sequence and structure space rather than evolving sequences only based on a fixed target structure in previous studies [9–14]. Attempts to address these issues would undoubtedly help to understand protein evolution principle and design novel protein sequence-structure pairs with potential functions [17–19].

In light of the knowledge encoded in most of the existing protein structures and sequences, the evolution has been converging into an ensemble of sequences whose structures satisfies the folding requirements [20, 21], except for inherently disordered proteins (IDP) [22]. The characteristic folding requirements of natural proteins can also reduce misfolding and aggregation propensities that hamper cellular functions and lead to diseases [23, 24]. Importantly, the folding requirements are not typical characteristics of random polymers. Thus, we suggest that protein folding is not an accidental event but the result of the actions or constrains from natural selection. In this view, the emergence of the special ensemble of naturally occurring protein sequences and structures is driven by the folding requirements of both thermodynamic stability and kinetic accessibility [25–27].

Energy landscape theory of protein folding has been proved to be fruitful in explaining the folding mechanism [6–8, 28]. We presented the detailed derivation for quantifying selection fitness from the energy landscape

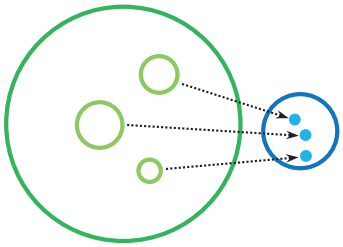

FIG. 1: Schematic diagram for the naturally occurring proteins in the sequence and structure space. Each subspace (areas inside the light green circle) corresponds to a set of natural occurring sequences that possess the same structure as their ground-state native structure (light blue point).

theory of protein folding in the Supplemental Material. In short, the theory argues that there are two critical temperatures to characterize the protein folding. The folding transition temperature T_f represents a first-order phase transition from denatured state to native state. The native state is thermodynamically stable below T_f until protein undergoes a glass transition at a temperature T_g. At T_g, protein is trapped in one frozen state where the native state is no longer kinetically accessible. The larger the ratio T_f/T_g is, the less chances of the protein is trapped on the way to its folding. For a protein to fold, it must kinetically access to the native state and be thermodynamically stable. Naturally occurring proteins should satisfy this folding requirement through evolution. Analytical studies have shown that the ratio of T_f and T_g can be expressed as $T_f/T_g = \Lambda + \sqrt{\Lambda^2 - 1}$, where $\Lambda = \sqrt{K_B/2S_0}\ \delta E/\Delta E$ (Fig. S1, Fig. S2 and see details in Supplemental Materia). δE represents the energy gap or slope of energy landscape characterizing the difference between the energy of the native ground state conformation (E_N) and the average energy of the conformation ensemble ($\overline{E}$), i.e. $\overline{E} - E_N$. ΔE is the variance of energies or the width of the energy distribution, i.e. $\sqrt{\langle E^2\rangle - \langle E\rangle^2}$, it can be used to quantify the roughness of the energy landscape. S_0 represents the entropy or the size of the protein. Λ is a quantitative measure of the landscape topography of protein folding as the slope against the roughness modularized by the size. The larger the Λ is, the more funneled protein folding energy landscape shape is against the vast number of states and roughness. The relationship between the T_f/T_g and Λ indicates that maximizing T_f/T_g is equivalent of maximizing the value of Λ [5, 29–31]. Furthermore, landscape topography measure Λ is strongly correlated to the kinetic speed of folding. Thus landscape topography determines the kinetic accessibility of folding. The folding requirement is therefore associated to the underlying landscape topography. At a particular temperature higher than T_g, the thermodynamic stability of the native ground state is quantified as $\Delta G = -K_B T \ln(P_N/P_D) = E_N + \ln[\sum_{E>E_N} n(E)\exp(-E/K_B T)]$, where P_N and P_D are the probabilities of the sequence in its unique ground state and denatured state [32, 33] (see details in Supplementary Materia). A funnelled and minimally frustrated landscape with stable native state can be achieved if protein evolves with the optimization of ΔG for thermodynamic stability and Λ for kinetic accessibility. Λ and ΔG provide mathematical foundations and formulations which can be quantified as the selection force or fitness of protein evolution for not only the protein folding at a given sequence, but more importantly also for the evolution in both sequence and structure space, which we focus on in this Letter.

Different from previous studies in which the target structure was fixed [11–14], we simulated the evolutionary process of protein population as an adaptive walk in both the sequence and structure spaces via random mutation under the selection pressures of optimizing the thermodynamic stability and kinetic accessibility [34] (Fig. S3 and Fig. S4). With quantified Shannon entropies of sequence and structure space, the evolution dynamics can be visualized as the movements on a projected energy landscape in structure and sequence space (Fig. 2). The bowl-like energy landscape of the sequence evolution (Fig. 2a) indicates that the evolution in sequence space first proceeds as the number of sequences gradually reduced and then reached a plateau. The sequences still keep evolving at the plateau stage through the exploration of the basin in the sequence space. The plateau stage could be viewed as an evolution close line attractor where the size of the sequence space is no longer changed but the energy still gradually decreases. The size of the local basin of sequence space is determined by the dominant ground-state structure at the bottom of the funnel-like energy landscape of structure evolution (Fig. 2b). The sequences in the line attractor possess the same evolved and dominant structure as their nondegenerate ground state.

With the bowl-like energy landscape of sequence evolution and funnel-like energy landscape of structure evolution, a super funneled energy landscape for protein evolution is quantified and visualized as Fig. 2c. The super funneled energy landscape traces through the sequence entropy and structure entropy as the axis of the cross-section ellipsoid. It is funnelled towards the line attractor where multiple sequences encode the same dominant structure. The topography of the evolution energy landscape has a similar funneled shape as the folding landscape of a single protein. Both sequences and structures experience alterations with clear energy and entropy reduction compensations along the downhill of the funnel during the evolution process. Random sequences as well as their ground-state structures locate at the top of the funnel while the evolved sequences and their ground-state structure cluster at the bottom of the funnel. Both sequence and structure entropies decrease as the energy descend in the energy landscape until the line attractor arrives. The super funnelled energy landscape describes how proteins evolve both in the sequence and structure space.

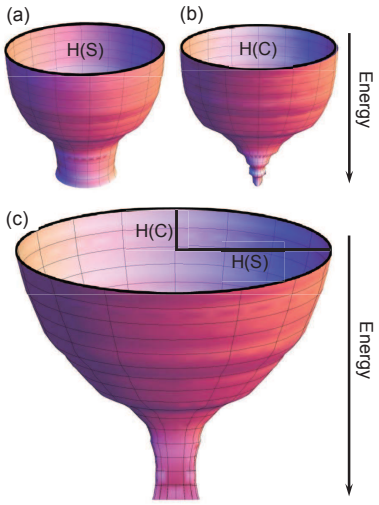

FIG. 2: Quantified energy landscape of protein evolution in sequence and structure space. (a) Bowl-like energy landscape for the protein evolution in the sequence space. (b) Funnel-like energy landscape for the protein evolution in the structure space. (c) Super funneled energy landscape of protein evolution in both sequence and structure space. One semiaxis of the ellipsoid stratum is the sequence entropy (H(S)) while the other one is the structure entropy (H(C)) (see details in Supplemental Materia).

It is known that evolved natural protein structures exhibit a high degree of regularity that is absent from random compact structures. The most obvious characteristics is the formation of a packed hydrophobic core as a structural component of globular protein structures (mostly protein domains). Optimization of the hydrophobic core has been the central goal for computational protein design [39–41]. Hydrophobic force has also long been recognized as a dominant factor of protein folding and hydrophobic collapse has been observed during protein folding, experimentally and computationally [42–44]. The hydrophobic core is a hallmark of natural and designed protein structures. Whether hydrophobic/hydrophilic residues segregation pattern is formed in the evolved structures is significant to justify the faithfulness of the evolution protocol in reproducing the protein evolution history and generating the evolved structures similar to naturally occurring proteins.

For the evolved structures, the probabilities of the hydrophobic residues residing on four types of lattices are apparently decreased from the Center to the Corner lattices rather than equally distributed (Fig. 3a, Table S1). The hydrophobic residues prefer to reside in the interior locations while the hydrophilic residues prefer to reside in the exterior locations (Fig. S5). It suggests that the evolved structure has a well-packed hydrophobic core in the interior of protein model (Fig. 3b). The result shows that the generation of the packing pattern of hydrophobic core is an inherent outcome of the protein evolution with the thermodynamic stability and kinetic accessability imposed on the protein evolution. It explains the importance of hydrophobic core for protein folding. Practically, automatical generation of the hydrophobic core from the evolution simulation would provide a valuable tool to design novel proteins with well-organized hydrophobic/hydrophylic segregation patten.

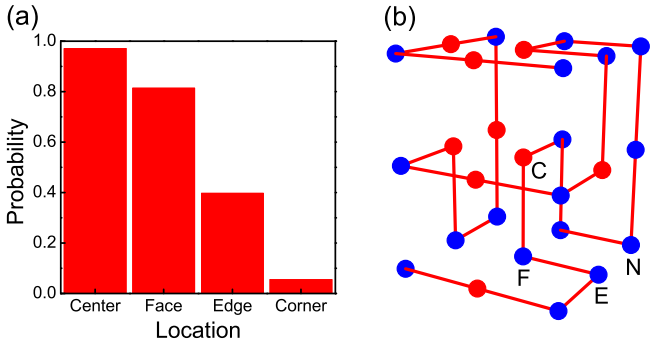

FIG. 3: Hydrophobic/hydrophilic residues segregation pattern. (a) The probabilities of hydrophobic residues residing on four types of locations in the structure, including Center (C), Face (F), Edge (E) and Corner (N). (b) Hydrophobic core is formed in a typical evolved structure with one of its evolved sequence mapping on the structure, hydrophobic residues are marked in red and hydrophilic residues in blue; the designability (N_S) of this structure is 52, and the evolved sequence is KTEGKVHDGTDPCKVKWQMEKCDCKCE.

Protein designability is another specific characteristics of natural proteins, which is defined as the number of the protein sequences taking the same protein structure as their ground-state conformation [32]. A highly designable structure is compatible with a large volume of sequences which have this structure as unique lowest-energy conformation [32, 45]. The highly designable structures are rare, but on average more stable and folding faster than other structures [32, 46–48]. Natural proteins are known to fold to a limited number of folds. It was estimated that the total number of different natural protein folds is only about 1000 [49, 50]. This leads to the assumption that naturally occurring proteins are a set of highly designable structures [32, 51]. Herein, it is important to uncover whether and why the highly designable structures are preferred during evolution.

The estimated distribution of the designability (N_S) shows that a strong bias existed in the distribution, i.e. very few structures have much higher N_S than those of other structures (Fig. 4a). To illustrate the selection preference of highly designable structures, three types of structures are classified from the structure space, they are random structures, designable structures and evolved structures (see definitions in the Supplemental Material). Because majority of structures are poorly designable or undesignable (Fig. 4a), the average N_S of the random structures (=0.61) is very small. Even with the consideration of the designability in selecting the structures, the average N_S of the designable structures is still only 6.10. However, the average N_S of the evolved structures

is increased significantly to 18.7. Compared to the random and designable structures, highly designable structures are much more preferred in the evolved structures. Protein evolution is driven by random mutation and selection pressures. Random mutation is equivalent to random selection of sequences in the sequence space, i.e. the designability is automatically involved in the evolution. Therefore, it can be concluded that the preference of the rare structures with high designability is the collaborative product of the selection pressure and random mutation. The improvement of thermodynamic stability and kinetic accessibility tend to evolve structures into high designability.

In addition, the increase of thermodynamic stability and kinetic accessibility also tend to enlarge the energy gap between the ground-state conformation and the excited-state conformation ($\delta = |E_N - E_C|$) (Fig. 4b,c). As seen in Fig. 4d, the degree of designability (N_S) correlates well with the magnitude of the gap (δ). Therefore, larger thermodynamic stability and kinetic accessibility often lead to larger gaps which can give rise to larger designability. This is consistent with some previous studies [27, 52, 53]. Larger energy gap of a sequence implies a greater ability to tolerate mutations, i.e. more mutations can be accommodated without losing the ability to fold into the same ground-state structure. In this sense, high designability implies large number of tolerate mutations. Due to the high designability and the large gaps favored by the evolution, natural proteins are compatible with large number of mutations. This explains why natural proteins are robust to mutation and also marginally stable [26, 54, 58, 59].

Except IDP, naturally occurring proteins fold in a biologically reasonable time scale. This requires natural sequences to fold fast into their native ground-state structures different from random sequences. It can be seen from the distribution of MFPT (mean first passage time) that on average the evolved sequences fold much faster than random sequences [55] (Fig. S6). Herein, it is conjectured that the fast folding of naturally occurred proteins arises as a consequence of evolutionary selections aimed at ensuring that funnel-like energy landscapes is achieved, as the funnel-like energy landscapes guarantees kinetic and thermodynamic requirements of protein folding [5–8, 31].

Naturally occurring proteins are generally composed of one or more functional domains which can fold and evolve independently [60–62]. Herein, two of the smallest protein domains (Villin headpiece (VHP) domain and WW domain) with different structure classes were chosen as off-lattice models [63, 64] for the evolution of real proteins (Fig. S7, Table S2-S5 and see details in Supplemental Materia). It is found that both native and evolved sequences have higher kinetic accessibility and thermodynamic stability than random sequences without evolution (Fig. S9). The evolved sequences for both VHP and WW domains show high similarities of hydrophobic preference to the native sequences, especially conserved residues of

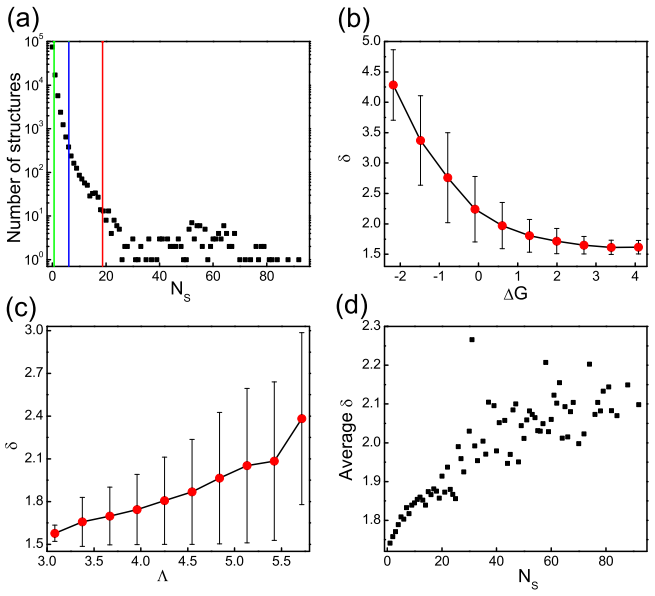

FIG. 4: Designability of evolved structures. (a) The number of structures as a function of designability (N_S). The average N_S for the random, designable and evolved structures are marked with green, blue and red lines respectively. (b) The relation between the thermodynamic stability (ΔG) and the energy gap (energy difference between the ground-state conformation and the exited-state conformation,$\delta = |E_N - E_C|$). (c) The relation between the kinetic accessibility (Λ) and the energy gap (δ). (d) Average energy gap (δ) as a function of designability (N_S).

the hydrophobic core [65] (Fig. 5 and Fig. S8). This suggests that the evolution protocol can faithfully produce the characteristic hydrophobic core of real proteins. For VHP domain, the dissimilarities of hydrophobic preference are mainly from the binding residues (Fig. S8(a) and (c)). This can be attributed to the fact that proteins are also evolved for functional binding, i.e. the requirement for functional binding during evolution is at the expense of the requirement for folding [30, 77, 78], in other words, trading the folding stability for binding function. For example, previous experimental and computational studies [79–84] have demonstrated that mutating the binding residues K24 and K29 on native sequences to hydrophobic residues can largely enhance the thermodynamic stability and increase the folding rate of VHP domain. The mutations are consistent with the hydrophobic preference of these two residues on evolved sequences (Fig. S8(a)). As validated from Fig. S9, the average thermodynamic stability and kinetic accessibility of evolved sequences are both higher than those of naturally occurred sequences. It demonstrates that the folding landscapes of the evolved sequences generally exhibit deeper folding funnels towards the folded state than natural sequences. The conclusions are insensitive to the residue mutation probability (Fig. S10 and Table S6-S8) and robust on interaction potentials [85] (Fig. S11).

Our study explains how the naturally occurring pro-

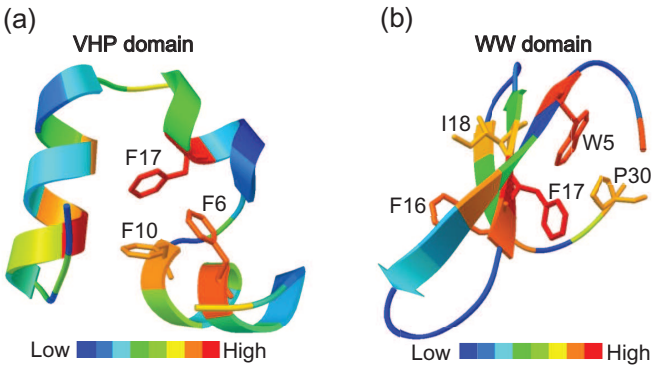

FIG. 5: Evolved sequences of real proteins. (a) VHP domain, (b) WW domain. The residues constitute the hydrophobic core are labeled and the hydrophobic preferences of the residues for evolved sequences are represented with color spacing.

teins emerge in the evolution and why they are special with a high degree of regularities and specificities which are absent in the random sequences. This evolution principle learned from this study provides valuable insights and practical way for the design of novel proteins [18, 19]. We conclude that the supper funnelled energy landscape unifies the principles of protein folding, evolution and design.

Acknowledgments

This work was financially supported by Ministry of Science and Technology of China (Grant No. 2016YFA0203200), National Natural Science Foundation of China (Grant No. 91430217 and 21403208), Natural Science Foundation of Jilin Province (Grant No. 20180101241JC) and National Science Foundation (Grant No. NSF-PHY-76066).

[1] M. J. Harms and J. W. Thornton, Nat. Rev. Genet. **14**, 559 (2013).
[2] A. D. Keefe and J. W. Szostak, Nature **410**, 715 (2001).
[3] C. B. Anfinsen, Science **181**, 223 (1973).
[4] C. Levinthal, in *Proceedings of a meeting held at Allerton House* (University of Illinois Press: Urbana, IL, USA, 1969), pp. 22–24.
[5] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).
[6] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992).
[7] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins **21**, 167 (1995).
[8] K. A. Dill and H. S. Chan, Nat. Struct. Biol. **4**, 10 (1997).
[9] Y. Xia and M. Levitt, Curr. Opin. Struct. Biol. **14**, 202 (2004).
[10] P. B. Chi and D. A. Liberles, Protein Sci. **25**, 1168 (2016).
[11] L. A. Mirny, V. I. Abkevich, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **95**, 4976 (1998).
[12] Y. Xia and M. Levitt, Proc. Natl. Acad. Sci. U.S.A. **99**, 10382 (2002).
[13] E. Bornberg-Bauer and H. S. Chan, Proc. Natl. Acad. Sci. U.S.A. **96**, 10689 (1999).
[14] T. Yomo, S. Saito, and M. Sasai, Nat. Struct. Mol. Biol. **6**, 743 (1999).
[15] G. Bouvignies, P. Vallurupalli, D. F. Hansen, B. E. Correia, O. Lange, A. Bah, R. M. Vernon, F. W. Dahlquist, D. Baker, and L. E. Kay, Nature **477**, 111 (2011).
[16] T. Sikosek, H. Krobath, and H. S. Chan, PLoS Comput. Biol. **12**, e1004960 (2016).
[17] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker, Science **302**, 1364 (2003).
[18] P.-S. Huang, G. Oberdorfer, C. Xu, X. Y. Pei, B. L. Nannenga, J. M. Rogers, F. DiMaio, T. Gonen, B. Luisi, and D. Baker, Science **346**, 481 (2014).
[19] P.-S. Huang, S. E. Boyken, and D. Baker, Nature **537**, 320 (2016).
[20] F. Morcos, N. P. Schafer, R. R. Cheng, J. N. Onuchic, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **111**, 12408 (2014).
[21] C. M. Dobson, Nature **426**, 884 (2003).
[22] A. K. Dunker, I. Silman, V. N. Uversky, and J. L. Sussman, Current Opin. Struct. Biol. **18**, 756 (2008).
[23] M. Bucciantini, E. Giannoni, F. Chiti, F. Baroni, L. Formigli, J. Zurdo, N. Taddei, G. Ramponi, C. M. Dobson, and M. Stefani, Nature **416**, 507 (2002).
[24] A. E. Lobkovsky, Y. I. Wolf, and E. V. Koonin, Proc. Natl. Acad. Sci. U.S.A. **107**, 2983 (2010).
[25] S. A. Lim, K. M. Hart, M. J. Harms, and S. Marqusee, Proc. Natl. Acad. Sci. U.S.A. **113**, 13045 (2016).
[26] F. O. Tzul, D. Vasilchuk, and G. I. Makhatadze, Proc. Natl. Acad. Sci. U.S.A. **114**, E1627 (2017).
[27] E. Y. Klein, D. Blumenkrantz, A. Serohijos, E. Shakhnovich, J.-M. Choi, J. V. Rodrigues, B. D. Smith, A. P. Lane, A. Feldman, and A. Pekosz, mSphere **3**, e00554 (2018).
[28] P. G. Wolynes, Philos. Trans. Royal Soc. Lond. A Math. Phys. Eng. Sci. **363**, 453 (2005).
[29] S. S. Plotkin, J. Wang, and P. G. Wolynes, J. Chem. Phys. **106**, 2932 (1997).
[30] J. Wang and G. M. Verkhivker, Phys. Rev. Lett. **90**, 188101 (2003).
[31] C.-L. Lee, G. Stell, and J. Wang, J. Chem. Phys. **118**, 959 (2003).
[32] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).
[33] H. S. Chan and K. A. Dill, Annu. Rev. Biophys. Biophys. Chem. **20**, 447 (1991).
[34] See Supplemental Materia for the sequence and structure model emplolyed in the protein evolution simulation, which includes Refs. [35-38].
[35] V. Tozzini, Curr. Opin. Struct. Biol. **15**, 144 (2005).
[36] C. Clementi, Curr. Opin. Struct. Biol. **18**, 10 (2008).
[37] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **101**, 1519 (1994).
[38] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623

(1996).
[39] E. P. Baldwin and B. W. Matthews, Curr. Opin. Biotechnol. **5**, 396 (1994).
[40] S. Sun, R. Brem, H. S. Chan, and K. A. Dill, Protein Eng. Des. Sel. **8**, 1205 (1995).
[41] B. Borgo and J. J. Havranek, Proc. Natl. Acad. Sci. U.S.A. **109**, 1494 (2012).
[42] K. A. Dill, Biochemistry **29**, 7133 (1990).
[43] V. R. Agashe, M. Shastry, and J. B. Udgaonkar, Nature **377**, 754 (1995).
[44] Y. Duan and P. A. Kollman, Science **282**, 740 (1998).
[45] H. Li, C. Tang, and N. S. Wingreen, Biochemistry **95**, 4987 (1998).
[46] R. Mélin, H. Li, N. S. Wingreen, and C. Tang, J. Chem. Phys. **110**, 1252 (1999).
[47] T. Wang, J. Miller, N. S. Wingreen, C. Tang, and K. A. Dill, J. Chem. Phys. **113**, 8329 (2000).
[48] P. Koehl and M. Levitt, Proc. Natl. Acad. Sci. U.S.A. **99**, 1280 (2002).
[49] C. Chothia, Nature **357**, 543 (1992).
[50] C. Orengo, D. T. Jones, and J. M. Thornton, Nature **372**, 631 (1994).
[51] E. G. Emberly, N. S. Wingreen, and C. Tang, Proc. Natl. Acad. Sci. U.S.A. **99**, 11163 (2002).
[52] R. A. Broglia, G. Tiana, H. E. Roman, E. Vigezzi, and E. Shakhnovich, Phys. Rev. Lett. **82**, 4727 (1999).
[53] H. Li, C. Tang, and N. S. Wingreen, Proteins **49**, 403 (2002).
[54] D. M. Taverna and R. A. Goldstein, J. Mol. Biol. **315**, 479 (2002).
[55] See Supplemental Materia for Monte Carlo simulation of protein folding and the definition of MFPT, which includes Refs. [56-57].
[56] E. Shakhnovich, M. Karplus, et al., Nature **369**, 248 (1994).
[57] A. Gutin, V. Abkevich, and E. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **92**, 1282 (1995).
[58] J. D. Bloom, S. T. Labthavikul, C. R. Otey, and F. H. Arnold, Proc. Natl. Acad. Sci. U.S.A. **103**, 5869 (2006).
[59] D. M. Taverna and R. A. Goldstein, Proteins **46**, 105 (2002).
[60] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, et al., Nucleic Acids Res. **44**, D279 (2015).
[61] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, J. Mol. Biol. **247**, 536 (1995).
[62] N. L. Dawson, T. E. Lewis, S. Das, J. G. Lees, D. Lee, P. Ashford, C. A. Orengo, and I. Sillitoe, Nucleic Acids Res. **45**, D289 (2016).
[63] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, Science **334**, 517 (2011).
[64] E. G. Baker, G. J. Bartlett, K. L. Porter Goff, and D. N. Woolfson, Acc. Chem. Res. **50**, 2085 (2017).
[65] See Supplemental Materia for the discription of hydrophobic residues and the binding residues on the seuqnces of VHP and WW domains, which includes Refs. [66-76].
[66] C. J. McKnight, P. T. Matsudaira, and P. S. Kim, Nat. Struct. Biol. **4**, 180 (1997).
[67] B. S. Frank, D. Vardar, D. A. Buckley, and C. J. McKnight, Protein Sci. **11**, 680 (2002).
[79] T. K. Chiu, J. Kubelka, R. Herbst-Irmer, W. A. Eaton, J. Hofrichter, and D. R. Davies, Proc. Natl. Acad. Sci. U.S.A. **102**, 7517 (2005).
[69] J. Meng, D. Vardar, Y. Wang, H.-C. Guo, J. F. Head, and C. J. McKnight, Biochemistry **44**, 11963 (2005).
[70] S. Xiao, Y. Bi, B. Shan, and D. P. Raleigh, Biochemistry **48**, 4607 (2009).
[71] J. W. Brown, D. Vardar-Ulu, and C. J. McKnight, J. Mol. Biol. **393**, 608 (2009).
[72] M. J. Macias, M. Hyvönen, E. Baraldi, J. Schultz, M. Sudol, M. Saraste, and H. Oschkinat, Nature **382**, 646 (1996).
[73] M. J. Macias, V. Gervais, C. Civera, and H. Oschkinat, Nat. Struct. Mol. Biol. **7**, 375 (2000).
[74] M. A. Verdecia, M. E. Bowman, K. P. Lu, T. Hunter, and J. P. Noel, Nat. Struct. Mol. Biol. **7**, 639 (2000).
[75] M. J. Macias, S. Wiesner, and M. Sudol, FEBS lett. **513**, 30 (2002).
[76] M. Jäger, M. Dendle, and J. W. Kelly, Protein Sci. **18**, 1806 (2009).
[77] J. D. Bloom, C. O. Wilke, F. H. Arnold, and C. Adami, Biophys. J. **86**, 2758 (2004).
[78] M. Manhart and A. V. Morozov, Proc. Natl. Acad. Sci. U.S.A. **112**, 1797 (2015).
[79] T. K. Chiu, J. Kubelka, R. Herbst-Irmer, W. A. Eaton, J. Hofrichter, and D. R. Davies, Proc. Natl. Acad. Sci. U.S.A. **102**, 7517 (2005).
[80] J. Kubelka, T. K. Chiu, D. R. Davies, W. A. Eaton, and J. Hofrichter, J. Mol. Biol. **359**, 546 (2006).
[81] Y. Bi, J.-H. Cho, E.-Y. Kim, B. Shan, H. Schindelin, and D. P. Raleigh, Biochemistry **46**, 7497 (2007).
[82] H. Lei, X. Deng, Z. Wang, and Y. Duan, J. Chem. Phys. **129**, 10B612 (2008).
[83] H. Lei, C. Chen, Y. Xiao, and Y. Duan, J. Chem. Phys. **134**, 05B613 (2011).
[84] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, Proc. Natl. Acad. Sci. U.S.A. **109**, 17845 (2012).
[85] See Supplemental Materia for the dicussions of protein evolution using different mutation probabilities and interaction potentials, which includes Refs. [86-88].
[86] T. Lazaridis and M. Karplus, Curr. Opin. Struct. Biol. **10**, 139 (2000).
[87] D. Hinds and M. Levitt, Proc. Natl. Acad. Sci. U.S.A. **89**, 2536 (1992).
[88] P. D. Thomas and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **93**, 11628 (1996).