The role of evolutionary selection in the dynamics of protein structure evolution

Amy I. Gilson[1], Ahmee Marshall-Christensen[1], Jeong-Mo Choi[1], Eugene I. Shakhnovich[1,*]

[1]Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford St., Cambridge MA, USA

* Correspondence should be addressed to E.I.S. shakhnovich@chemistry.harvard.edu

Abstract

Emergence of new protein structures has proved difficult to trace in nature and engineer in the laboratory. However, one aspect of structure evolution has proved immensely helpful for determining the three-dimensional structure of proteins from their sequences: in the vast majority of cases, proteins that share more than 30% sequence identity have similar structures. Below this mark is the "twilight zone" where proteins may have identical or very different structures. These observations form the foundational intuition behind structure homology modeling. Despite their importance, however, they have never received a comprehensive biophysical justification. Here we show that the onset of the twilight zone is more gradual for proteins with low contact density, a proxy for low thermodynamic stability, than proteins with high contact density. Then we present an analytical model that treats divergent fold evolution as an activated process, in analogy to chemical kinetics, where sequence evolution must overcome thermodynamically unstable evolutionary intermediates to discover new folds. This model explains the existence of a twilight zone and explains why its onset is more abrupt for some classes of proteins than for others. We test the assumptions of the model and characterize the dynamics of fold evolution using evolutionary simulations of model proteins and cell populations. Overall these results show how fundamental biophysical constraints directed evolutionary dynamics leading to the Universe of modern protein structures and sequences.

Introduction

A wide variety of protein structures exist in nature, however the evolutionary origins of this panoply of proteins remain unknown. While protein sequence evolution is easily traced in nature and produced in the laboratory, the emergence of new protein structures is rarely observed and difficult to engineer (1-3). One approach to studying structure evolution is to examine how proteins' structural similarity varies over a range of sequence identities. Such investigations proceed by aligning many pairs of proteins so that their sequence identity and structural similarity can be assessed (4-6). The result is a cusped relationship between sequence and structure divergence: Sequences reliably diverge up to 70% without significant evolution of protein structure. Below 30% sequence identity, the structural similarity between proteins abruptly decreases, giving rise to a "twilight zone" where little can be said about the relationship between sequence identity and structural similarity without more advanced methods. This finding has provided the foundation of one of the most important methods in protein biophysics: structure homology modeling (7, 8). Despite the fact that the plateau of high structural similarity above 30% sequence identity has been crucial for homology modeling and the abrupt onset of the twilight zone has motivated many of the advanced structure prediction methods, the cusped relationship between sequence and structural similarity has not yet received a detailed biophysical justification (9).

Randomly mutating 70% of the residues in a protein will almost surely unfold it as even a small number of point mutations can destroy a protein's structure (10). Therefore, it is clear that without evolutionary selection, the range of 100-30% sequence identity could not correspond to nearly identical structures. Previous models of protein structure evolution have been able to explain many fundamental features of the protein universe using purely physical models, without any selection. Dokholyan et al. constructed a protein domain universe graph in which protein domain nodes are connected by an edge if they are structurally similar. The resulting graph is scale-free, which they showed would be the result of duplication and structural divergence of proteins (11, 12). Similarly, the birth, death, innovation models developed by Koonin et al. explains the power law-like distribution of gene family sizes that exists in many genomes (13). However, because these works use neutral models, they are unable to explain the cusped sequence-structure relationship.

A small but growing collection of cases where protein structure evolution has been observed or inferred provides mechanistic insight into the role of selection in protein structure evolution. They show that it is possible for proteins to be within a small number of point mutations of a fold evolution event (14-16) but also that structure emergence may often pass through thermodynamically destabilized intermediates. Among Cro bacteriophage transcription factors, a pair of homologous proteins with 40% sequence identity (indicating that they share a recent common ancestor) but different structures was found. Subsequent studies showed that some Cro proteins might be just a few mutations away from changing fold (14, 17). Similarly, a trajectory of point mutations was engineered to convert a natural protein adopting a 3α fold into a 4β+α fold. A major result of this work was finding a rare sequence of mutations that avoided unfolded

intermediates (15, 18, 19). Computational investigations of protein fold evolution using model proteins of Protein Data Base structures also show that structure evolution traverses unstable intermediates and found that less stable proteins are more evolvable at the structural level (20, 21).

Here we explore the evolutionary dynamics of discovery of new protein structures. Given the fact that most globular proteins must adopt a well-defined 3D structure in order to function and the strong evidence that many fold evolution pathways require protein destabilization, we hypothesize that protein structure discovery requires crossing of valleys of low fitness on fitness landscapes corresponding to genetic encoding of evolutionarily transient, unstable proteins. Therefore, the strength of selection for folding stability under which a protein evolves may modulate its capacity to evolve a new structure. In this work we study evolution of protein folds using the data on sequence structure relationships in natural proteins and a variety of evolutionary models of increasing complexity and realism and find that the cusped relationship between structure and sequence divergence is a direct consequence of the interplay between evolutionary dynamics and biophysical constraint for protein folding stability. The sharpness of the "cusp," but not its position (70% of structural similarity for natural proteins) is determined by the strength of evolutionary selection for folding stability.

Results

Structure-sequence relationship in the protein universe

A cusp-like relationship between structure and sequence divergence has long been observed in the record (4-6, 9) when the sequence identity and structural similarity is determined for many pairs of protein domains. Here, we explore how selection for protein stability affects the shape of the relationship between sequence and structure divergence. We use the SCOP database of protein domains and examine the evolution of globular proteins as sequences diverge by aligning the sequences and structures of SCOP domains (See *Methods*, Fig. 1A) (22). Sequences were aligned using the Needleman-Wunsch algorithm as implemented in MATLAB and structures were aligned using the Template-Modeling (TM) algorithm and given a TM score between 0 and 1 to quantify their degree of structural similarity (23).

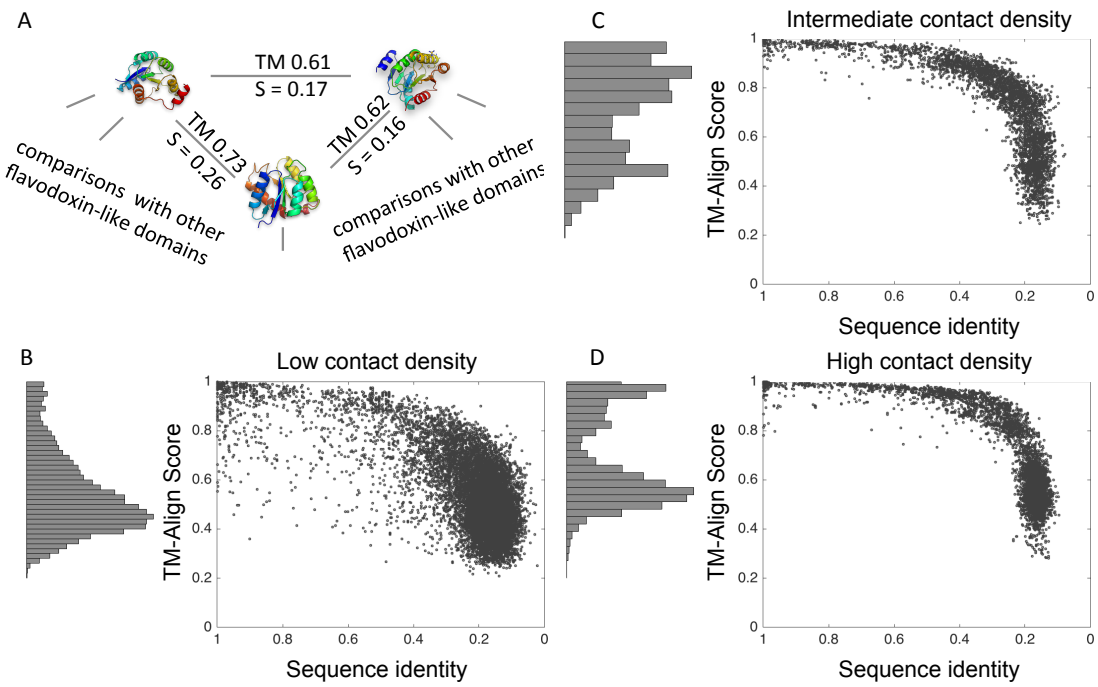

Figure 1: Stability and fold evolution of SCOP domains SCOP domains classified as α, β, α+β, α/β were used. Sequence identity and structural homology (TM-align score) for each pair of domains classified into the same fold were calculated. (A) An illustrative example from the α+β flavodoxin-like folds. The SCOP domains and their associated PDB ids from left to right are d1a04a2 (1A04), d1a2oa1 (1A2O), and d1eudb1 (1EUD). (B-D) The relationship between sequence divergence and structure evolution in SCOP domains. Domains are partitioned by contact density (B) bottom 10% contact density (<4.13 contacts/residue, N=12,671 data points) (C) middle 10% contact density (4.57<CD<4.65 contacts/residue. N=3,863 data points) (D) top 10% contact density (>4.93 contacts/residue, N=5,672 data points). Histograms at the left of plots B-D show that at high contact density, the distribution of structure similarity TM-align scores is bimodal.

We define thermodynamic stability of a protein as the fraction of molecules in the ensemble that resides in the native state. It is an experimentally observable thermodynamic property of proteins that is related to fitness and as such might be under evolutionary selection (24-28). For single domain proteins that fold as two state systems this quantity is directly related to the folding free energy ΔG through the Boltzmann relation of statistical mechanics:

$$P_{nat} = \frac{e^{-\frac{\Delta G}{k_B T}}}{1 + e^{-\frac{\Delta G}{k_B T}}} \tag{1}$$

where T is temperature and k_B is Boltzmann constant. Here we consider only single domains as defined in SCOP and use the number of amino acid contacts normalized by the domain length, the contact density (CD), as a proxy for the folding free energy ΔG of a protein in order to compensate for the relative deficiency of experimental

thermodynamic data (20). The connection between CD and folding free energy has been established through several lines of evidence. Inter-residue contacts stabilize a protein structure through van der Waals interactions, so the more contacts in a protein structure the more stable the structure. Indeed, proteins of thermophilic organisms have greater contact density than their mesophilic homologs (29). Finally, protein length correlates with both contact density and folding free energy. Correlation between length and contact density is a consequence of the globular structure of proteins where surface to volume ratio depends on total length. Folding free energy is an extensive property that increases with the total number of amino acids in a protein (30, 31). Generally lower free energy does not necessarily mean greater stability (P_{nat}). That is, a protein consisting of two identical thermodynamically independent domains each having folding free energy ΔG, will not be more stable than each domain in separation. However, for single-domain proteins that fold in a two-state manner folding free energy is directly related to stability as can be seen from Eq.1.

In Figs 1B-D (Dataset S1), the data are subdivided by contact density, such that only proteins in the bottom 10% (CD < 4.13 contacts per amino acid residue), middle 10% (4.65 > CD > 4.57 contacts per amino acid residue), and top 10% (CD > 4.93 contacts per amino acid residue), of contact density are compared (Fig 1B, C, and D respectively). For the most stable proteins, there is a bimodal distribution of TM-align scores with a peak at high TM-align scores corresponding to diverging proteins maintaining near identical structures and a peak at low TM-align scores corresponding to proteins that are unrelated or whose folds have evolved (Fig. 1B). However in the class of low contact density proteins there is only a peak at low TM-align score. For high and intermediate contact density domain, there is a pronounced cusp in the sequence-structure relationship. By contrast, the transition from homologous structures to diverged structures is more gradual for the low contact density domains, without a visible cusp. Overall, this analysis indicates that, as proteins diverge, the onset of structure evolution occurs earlier for less stable proteins than for more stable ones.

A possible confounding factor could be that proteins of different CD might be enriched in different structural classes so that our finding presented in Fig.1 might reflect differences in evolutionary scenarios for different protein structural classes. We analyzed that and found that there are no significant differences in the shape of the sequence-structure relationship when the data is subdivided by protein family or class (Fig. S1).

A simple analytical model for the twilight zone

An evolutionary trajectory of sequential amino acid substitutions can be imagined connecting proteins of any two structures. There are many experimental studies supporting the vision of an evolutionary landscape where sequences folding into stable structures are connected by sequences that do not adopt stable structures (18, 32-35). In general, a protein's stability determines the cellular amount of folded, and therefore functional, protein as well as the amount of unfolded protein, which not only is non-functional but also can form toxic aggregates (36). The fitness of an organism is thereby directly related to the ability of proteins to carry out their functions and therefore their stability.

Based on these understandings, we construct a simple model of protein structure evolution on a fitness landscape where sequences encoding high-fitness (thermodynamically stable) proteins form peaks separated by fitness valleys. (Fig. 2A). The model is based on three postulates. First, new folds are discovered in divergent evolution from existing folds. Second, in analogy with chemical kinetics we treat the events leading to structure evolution as an activated process, where wait times between fitness valley crossing events are exponentially distributed. The evolutionary reaction coordinate is a mutational path connecting two protein structure states. A free energy barrier separating two states in chemical kinetics is analogous to the evolutionary barrier formed by a fitness valley separating two stable protein structures that confer high fitness on their carrier organisms. Just as the ratio of the free energy barrier to temperature controls the rate of barrier crossing in chemical kinetics, the ratio of the effective fitness barrier to the strength of evolutionary selection controls the rate of structure discovery in the evolutionary context of the model (37). Proteins under strong evolutionary pressure in this model will be "fit" and stable since we assume stability is a main component of fitness throughout this paper. Third, we postulate that crossing the fitness valley leads to the discovery of novel folds that are structurally dissimilar to the original fold.

We begin by calculating the rate of structure evolution. From the Arrhenius equation, we have

$$k \propto e^{-\Delta G_f / T_{sel}} \tag{2}$$

where ΔG_f is the fitness barrier height and T_{sel} is the "evolutionary" selection temperature, which tunes the strength of selection acting on the evolving proteins. The probability, $q(t)$, that an ancestral structure (the structure at time $t=0$) is unchanged at time t follows immediately:

$$q(t) = e^{-kt} \tag{3}$$

Next, it is necessary to substitute the variable of time for the variable of sequence identity for two reasons. First, sequences can reach mutation saturation. Second, this transformation will permit direct comparison between the analytical and the bioinformatics results. The relationship for the expected Hamming distance between the evolving protein at time t with respect to the ancestral protein at time $t=0$ is given by

$$S(t) = \frac{l(a-1)}{a}\left(1-\left(1-\frac{a}{l(a-1)}\right)^t\right) \tag{4}$$

where l is the length of the protein and a is the number of amino acid types, typically 20 (see Supplementary Text for derivation). This expression only takes into account the point mutations that become fixed in the evolving population because the model only takes into account mutations that might contribute to the emergence of a new fold.

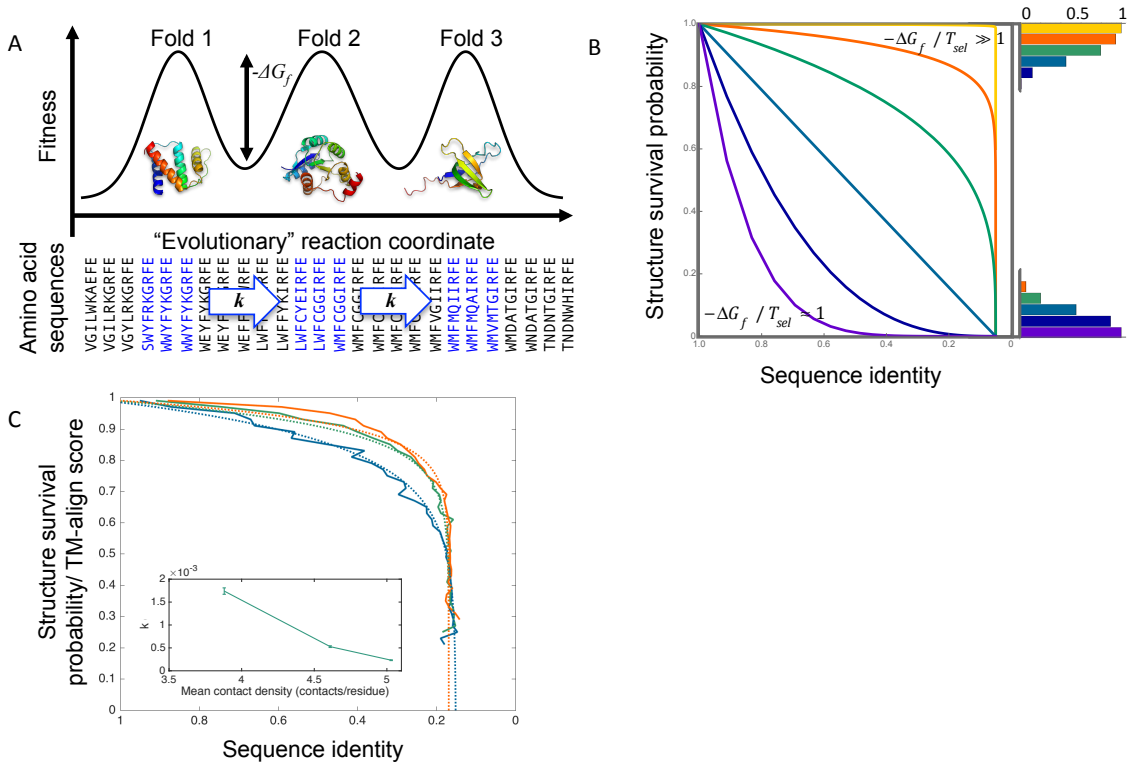

Figure 2: Energetic barriers to fold evolution (A) A schematic representation of the evolutionary process in which fold discovery events occur by overcoming fitness valleys along the "evolutionary" reaction coordinate. Sequences that are "more fit," i.e. that contribute to high fitness in the organisms they are part of are shown in blue. Families of homologous proteins are formed by sequence evolution between structure evolution events. The rate of fold discovery depends on the free energy barrier to evolving a new fold (a property of protein biophysics) and an evolutionary "temperature" corresponding to the strength of selection for folding stability under which proteins evolve such that $k \propto e^{\Delta G_f / T_{sel}}$. (B) The relationship between sequence identity and structural survival probability of an ancestral fold for a proteins with length 27 at several values of $\Delta G_f / T_{sel}$: from top to bottom, 9.7 (yellow), 5.8 (orange), 4.5 (green), 3.2 (light blue), 2.2 (dark blue) 1.4 (purple). The histograms show the probability that the ancestral structure has persisted (top) versus the probability that new structure has emerged (bottom) after at most 74% of the residues have been mutated. (C) Fit of the analytical model to real protein data. The model fit to the data is presented in dashed lines while binned bioinformatics data is presented in solid lines for each contact density subgroup: low contact density (blue), intermediate contact density (green), high contact density (orange). In the inset, the correlation between protein domain contact density and protein evolutionary rate, k, predicted analytically. Error bars indicate one standard deviation.

Excluding the time variable from equations (3) and (4), we determine the probability q that a structure has persisted after once its sequence is different at S from its ancestral $(t=0)$ sequence.

$$q(S)=\exp\left(-k\frac{\log\left(1-S\frac{a}{l(a-1)}\right)}{\log\left(1-\frac{a}{l(a-1)}\right)}\right) \tag{5}$$

Figure 2B shows the dependence of structure survival probability as a function of sequence identity ($q(S)$ v. $1-S$). When $-\Delta G_f / T_{sel} \approx 1$ selection pressure is so weak that the evolutionary barriers are not a significant impediment and the protein is free to diffuse through structure space as easily as it does through sequence space. As selection pressure increases and proteins become more stable, the barriers between protein structures begin to retard structure evolution giving rise to activated dynamics. When the ratio $-\Delta G_f / T_{sel} \gg 1$, the sequence identity degrades to random before the first structure evolution events take place, resulting in an abrupt decrease in $q(S)$ at very low S. Conceptually, this is because broad exploration of sequence space takes place over the course of many mutations before there arises a sequence that is stable yet which has a new ground state. The histograms in Figure 2B also indicate that differences in selection pressure may explain heterogeneity that has been observed in fold family sizes (38). For each of the selection ratios tested, the histogram shows the probability that the structure has (bottom) or has not (top) evolved after some evolutionary time. The curves and histograms for strong selection are consistent with the high contact density class of proteins while those for weaker selection are consistent with the low contact density class of proteins. Proteins evolving under weak selection pressure move rapidly through fold space, providing little time for gene duplication and sequence divergence to populate a particular fold family when compared to their strong selection counterparts.

Finally, we confirm that the analytical model predicts a decreasing rate of structure evolution, k, when fit to the three protein domains subgroups with increasing contact density. The bionformatic data featured in Fig. 1 was binned and averaged for each contact density subgroup and the rate of structure evolution k and twilight zone offset, c, were then fit to this bioinformatics data (See *Methods*). As shown in Figure 2C, this fit confirms both that the analytical model correctly reproduces the shape of the sequence-structure relation for proteins of different contact densities, and that contact density correlates negatively with protein structure evolution rate, k, (inset, Table S1). Interestingly, while it might be expected that the analytical model reproduces the correlation between contact density and k when comparing protein domain groups with very different contact densities, it also seems to be able to discriminate between protein subgroups with smaller differences in average contact density. Protein domains in the four structural classes, α, β, α/β, α+β, have average contact densities that range from 4.33 contacts/residue (α) to 4.66 contacts/residue (α/β). While differences in the shape of the

sequence-structure relation among the four classes remain hardly distinguishable by eye even after this binning and fitting, the fitted k values do decrease monotonically with the classes' increasing contact density indicating that structure evolution rates are sensitive even to modest differences in contact density (Fig S2, Table S1).

Structure evolution of model proteins

We now turn to explicit modeling of protein structure evolution in order to test the assumptions of our analytical model and to get mechanistic insight into the biophysics of fold emergence. Details of the model are provided in the *Methods* section. In brief, each model protein consists of 27 amino acid residues that fold into a compact 3×3×3 cube (39-41). All 103,346 possible compact structures of such model proteins have been enumerated, and, following Heo et al., we use a representative subset of randomly selected 10,000 conformations as our space of possible protein structures for computational efficiency throughout this work (42). Neighboring amino acid residues that are not connected by a covalent bond interact according to a Miyazawa-Jernigan (MJ) potential (43). In line with previous discussion, the stability of each model protein is represented by P_{nat}, the Boltzmann probability that a protein adopts its lowest energy (native) state. For any 27-mer sequence, P_{nat} can be determined exactly within this model (see *Methods*).

We ran many evolutionary simulations where new structures were evolved under a stability (P_{nat}) constraint as follows. Each generation, an amino acid substitution was introduced into the evolving protein. Stabilizing mutations that increase P_{nat} were always accepted, while destabilizing mutations were accepted according to the Metropolis criterion with a selective temperature T_{sel} that establishes stringency of evolutionary selection (44, 45) (see *Methods*, Fig. S3 and Dataset S2). Simulations ran for 1,000 generations (mutation attempts, illustrated in Fig. 3A) and the structure, sequence, and stability (P_{nat}) were recorded every 10 generations.

First, we quantify the structural similarity between the "wild-type" and mutant structure for each time step where a structure discovery event occurs in order to test the structural bimodality assumption made by the analytical model. The structural similarity of two model proteins is straightforwardly captured by the number of amino acid residue contacts that two structures have in common (Q-score, see *Methods*) (46).

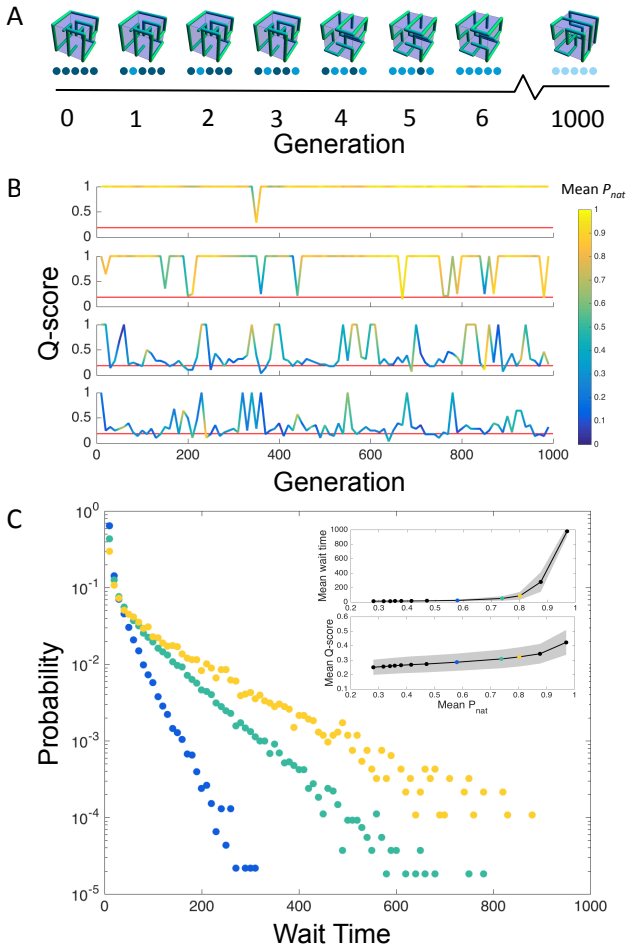

Figure 3: Dynamics of fold evolution (A) Schematic representation of Monte Carlo simulated evolution. Protein sequences (represented by a series of circles below each lattice structure, color indicates amino acid type) acquire mutations, periodically causing the protein to transition to a new structure. (B) Four representative fold evolution trajectories at different mean thermodynamic stabilities ($\bar{P}_{nat}$). Dips indicate structure evolution events. Trajectory color illustrates the thermodynamics stability of the evolving proteins, as described in the color bar. (C) The distribution of wait times between fold discovery events for three different $\bar{P}_{nat}$, $\bar{P}_{nat} = 0.58$ (blue), $\bar{P}_{nat} = 0.74$ (green), and $\bar{P}_{nat} = 0.80$ (yellow). In the insets and in (D), the colored data points also correspond to these $\bar{P}_{nat}$. The top inset shows the average wait time between fold discovery events for the different values of $\bar{P}_{nat}$ tested. The bottom inset shows the average structural similarity (Q-score) between the previous structure and the arising structure for fold discovery events.

In Figure 3B, representative individual trajectories from four selection regimes are plotted such that protein structure discovery events are reflected as dips in Q-score, and protein stability is indicated by color. The average Q-score, $\bar{Q}$, between a random pair of model proteins is 0.19 ± 0.08, as indicated by the red line in Fig. 3B. For each tested selection pressure, the average Q-scores associated with fold evolution events, $\bar{Q}_i$, is depicted in the bottom inset of Fig. 3C. $\bar{Q}_i$ ranges from 0.25 ± 0.11 when $\bar{P}_{nat} = 0.23$ to 0.42 ± 0.17 when $\bar{P}_{nat} = 0.97$. The connection between $\bar{Q}_i$ and $\bar{P}_{nat}$ likely reflects that discovering a mutant fold similar to the wild-type (high Q) avoids destabilizing evolutionary intermediates (Fig. S4). However, even strong selection pressure does not increase $\bar{Q}_i$ above ~0.43 because there exist so few structures with $Q > 0.43$ for any given evolving protein structure (Fig. S5). The underlying rarity of similar structures in the space of model protein structures erects evolutionary barriers to fold evolution and limits the capacity of structure evolution to occur incrementally. This is reflected in the distribution $P(Q_{t,t+10})$, which is bimodal with a peak at $Q = 1$ and $Q \approx 0.29$ (See below, Fig. 4A-D).

A key assumption of the analytical model was that fold evolution is an activated process (47). We tested this assumption by examining whether wait times between fold discovery events were distributed exponentially, which is the hallmark of an activated process, and found that indeed, model protein evolution does follow activated dynamics (Fig. 3C).
Alleviating purifying selection for stability accelerates the rate of structure evolution by allowing proteins to maintain lower stability on average (lower $\bar{P}_{nat}$). The average wait time between structure discovery events diminishes rapidly as simulations are run at higher T_{sel} such that the stability of the evolving proteins diminishes. When $\bar{P}_{nat} \leq 0.74$ selection pressure was so weak that nearly every recorded generation explored a different fold (Fig. 3B, Fig. 3C top insert), akin to the $-\Delta G_f / T_{sel} \approx 1$ regime in the analytical model. The abrupt transition from diffusive to activated dynamics is also apparent in the distribution of fold discovery events (Fig. S6A) which shows that the distribution is Poisson when $\bar{P}_{nat} \geq 0.74$, where structure evolution events are rare, and that the mean number of fold discovery events increases when $\bar{P}_{nat}$ falls below 0.74 (Fig. S6A, Dataset S2).

In summary, structure evolution of model proteins follows activated kinetics in the strong selection regime and newly discovered structures appear much different from parent ones, providing strong supports of the postulates of the analytical theory. Now we determine whether structure-sequence relationship reproduces the cusp shapes observed in real proteins (Fig.1). For unstable proteins the pace of structure evolution outstrips sequence evolution leading to a concave sequence-structure relationship (Fig. 4A), which is consistent with the $-\Delta G_f / T_{sel} \approx 1$ regime of the analytical model and not observed in

real proteins. The most interesting cases are at intermediate selection regimes where there is a sigmoidal cusp-like transition from high average structural similarity to low structural similarity at higher sequence divergence (Fig 4B-C). In this selection regime, proteins dwell in a particular structure while accumulating sequence mutations, yet are periodically, at longer time scales, able to transition to another structure, which is substantially different from the preceding one. For extremely stable proteins, by contrast, sequence can still evolve readily, but structure evolution is severely hindered (Fig. 4D). The increased dwell-time in a particular protein structure also allows sequences adopting that fold to proliferate over time, which is reflected in the correlation between $\overline{P}_{nat}$ and the mean structure family size (Figure 4E).

The analytical model and simulations indicate that a divergent scenario reproduces the key features of the sequence structure relationships shown in Fig.1. An important question is whether an alternative, convergent, mechanism of structure discovery can reproduce these relationships equally well. In Figures 4A and 4C we accompany the divergent evolution results described above with the results of convergent evolution simulations, shown in gray. To simulate convergent evolution, proteins start from sequences that stably fold into a randomly chosen structures but the fitness function favors sequence convergence to another, "target" protein (see *Methods* for details). Under weak selection for folding stability, the convergence of protein structure as sequences evolve follows a similar path as it does during divergent evolution (Fig. 4A). By contrast, under strong selection for stability, protein structures begin to converge at a much higher sequence identity than where they begin to diverge (approximately 85% and 25% respectively, Fig. 4C). Proteins converging under the constraint of stability also attain only $69 \pm 9\%$ sequence identity with respect to the target sequence after 3,000 mutation attempts and only 0.5% of the evolving proteins attained the target structure, compared to 100% in the weak selection regime.

Taken together, our results from divergent and convergent evolution scenarios indicate that selection pressure generates a peculiar hysteresis in evolutionary trajectories: When proteins diverge under selection for folding stability, they long retain the same structure because evolving a new structure involves passing through a fitness valley. Were two highly diverged (different sequence and structure) proteins to converge on the same sequence again, they would long retain *different* structures for the same reason. Such a scheme was realized experimentally in (15) in which two proteins with different sequences and structures were subjected to engineered single point mutations which increased their sequence identity up to 88% yet did not unfold the proteins or change their original structures.

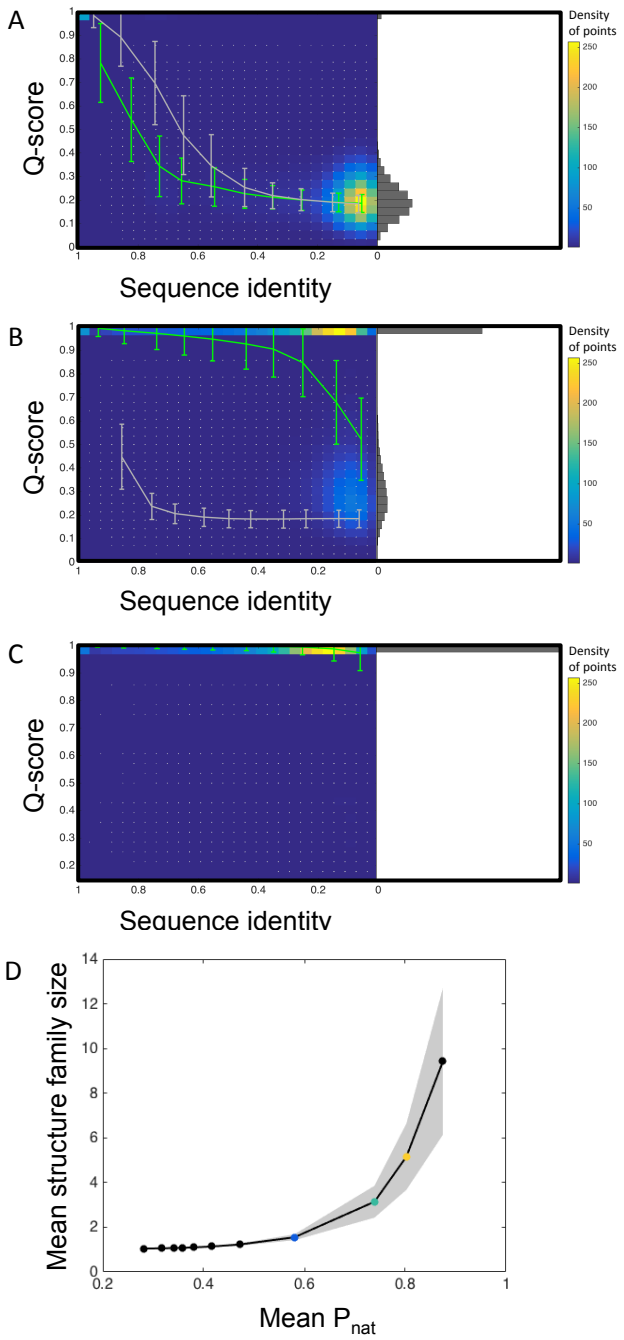

Figure 4: The relationship between sequence divergence and structure evolution for model proteins (A-C) The sequence identity and Q-score for each pair of proteins compared. White points indicate the data points from proteins arising in the same trajectory. The color scale indicates the density of points for each pair of Q-score, sequence identity values, and the average Q-score in each sequence identity bin of size 0.1 is calculated and plotted in green for divergent evolution simulations and in grey for convergent evolution simulations (see Methods) The histogram represents the frequency of Q-score pairs. (A) Weak selection for folding stability: $\overline{P}_{nat} = 0.32$. In this regime structure diverges more rapidly than sequence. The result of convergent evolution simulations without selection for stability is shown in grey. There is not

significant evolutionary hysteresis when selection for folding stability is low as can be seen in the similarity of the green and grey curves. (B) $\overline{P}_{nat} = 0.88$ When proteins evolve under strong selection for folding stability, as they do in both the divergent (green) and the convergent (grey) evolution simulations, there is clear evolutionary hysteresis: depending on whether a pair of proteins with 50% sequence identity are diverging or converging their structure will most likely be extremely similar or very different respectively). (C) $\overline{P}_{nat} = 0.97$: At such strong selection for thermodynamic stability, structure evolution is limited because stability valleys separating structures are not overcome on the time scale of simulations. (D) The mean structure family size as a function of $\overline{P}_{nat}$. The size of a structure family is the number of non-redundant sequences (sequence identity < 0.25) evolved in simulations that adopt a particular structure. Shaded regions show one standard deviation and colored points mark $\overline{P}_{nat} = 0.58$ (blue), $\overline{P}_{nat} = 0.74$ (green), and $\overline{P}_{nat} = 0.80$ (yellow).

Population size modulating selection pressure

Evolutionary simulations of individual model proteins, while biophysically realistic, lack biological realism: the concepts of fitness and selection are applicable to whole cells and populations rather than individual proteins. To address this shortcoming we performed simulations of an evolving population of competing single-cell model organisms, as depicted in Figure 5A and described in detail in *Methods* and (48). Model organisms have genes that encode 27-mer lattice model proteins. Each generation, an organism can die, divide, undergo a gene duplication event, or undergo a genetic point mutation. This evolutionary scheme mimics the natural emergence of protein families and was previously used to explain the power-law distribution of protein family and superfamily sizes observed in nature (48). In contrast to the previous model, selection pressure is here applied to the whole organisms rather than only the evolving proteins, which makes the situation more biologically realistic and non-trivial. The strength of selection (proxied by T_{sel} in the previous model of individual protein evolution) is determined by the population size (37, 49) in the present, more biologically realistic model.

Evolutionary runs were simulated over a range of maximum population sizes, $N_{\max}$, from 500 to 5,000 organisms, in replicates of 50. Whenever birth of new organisms drives the population size, N , above $N_{\max}$, $N - N_{\max}$ randomly selected organisms are removed to ensure constant population size, simulating a turbidostat (48).

After many generations, we recorded the extant proteins from all organisms in a particular evolutionary run, calculated their stabilities, and calculated the sequence identity and structural similarity (Q-score) of all pairs of proteins. We found that the evolutionary runs yielding unstable proteins have qualitatively different relationships between sequence and structure divergence than other replicas. In these cases, the extremely rapid turn-over of structures is manifested in a concave-up dependence of average Q-score on sequence identity (green and blue trajectories in Fig. 5B), as in the low P_{nat} regime of the Monte Carlo simulations (Fig. 4A). Once the population size

passes the critical threshold, of about 1,250 organisms, the characteristically cusped sequence-structure relationships become the most probable result of evolution (Fig. 5B, middle and bottom). On the other hand, we do not see a population size where structure evolution is almost entirely shut down, even at the largest population sizes probed in simulations. This could reflect that in larger populations, there is a larger influx of beneficial as well as deleterious mutations, so more neutral or beneficial fold evolution events can occur.

We found that the average P_{nat} of proteins at the end of 3,000 generations ($\bar{P}_{nat}$) strongly anti-correlates with the diversity of structures observed in the population, as reflected in the structure entropy:

$$H = -\sum_{i=1}^{10,000} p_i \log(p_i) \tag{6}$$

where p_i is the probability of finding structure i in the final generation of the simulations (inset of Fig. 5C). The correlation between $\bar{P}_{nat}$ and H reflects the fact that less stable proteins have faster rates of structure evolution. This correlation does not depend on population size (different colors in Fig 5B inset). Rather, we found that population size modulated the number of evolutionary replicas that failed to evolve stable proteins (Fig. 5C). This is also reflected in the relative peak heights of the Q-scores shown in Figure 5B.

Finally we examine the size of the average protein structure family across the range of $\bar{P}_{nat}$ explored in the simulations (Figure 5D). The size of protein structure family is defined as the number of genes encoding non-redundant protein sequences (sequence identity < 25%) but whose gene products adopt the same native state structure. In datasets of natural proteins, protein families of various sizes have been observed and overall, there is a positive correlation between family size and protein contact density (50). In the context of these simulations, which instantiate the crucial mechanisms of structure family creation and growth (sequence evolution, gene duplication, and structure evolution) we observe a strong positive correlation between $\bar{P}_{nat}$ of proteins at the end of an evolutionary replica and the average structure family size (Fig. 5D). The significance of the trend is further magnified by the observation that the total number of genes in a population at the end of a simulation actually correlates negatively with $\bar{P}_{nat}$ (Fig. S7). Overall, this observation supports the view that stable proteins are trapped in particular structures, providing more time for the number of sequences adopting this structure to grow.

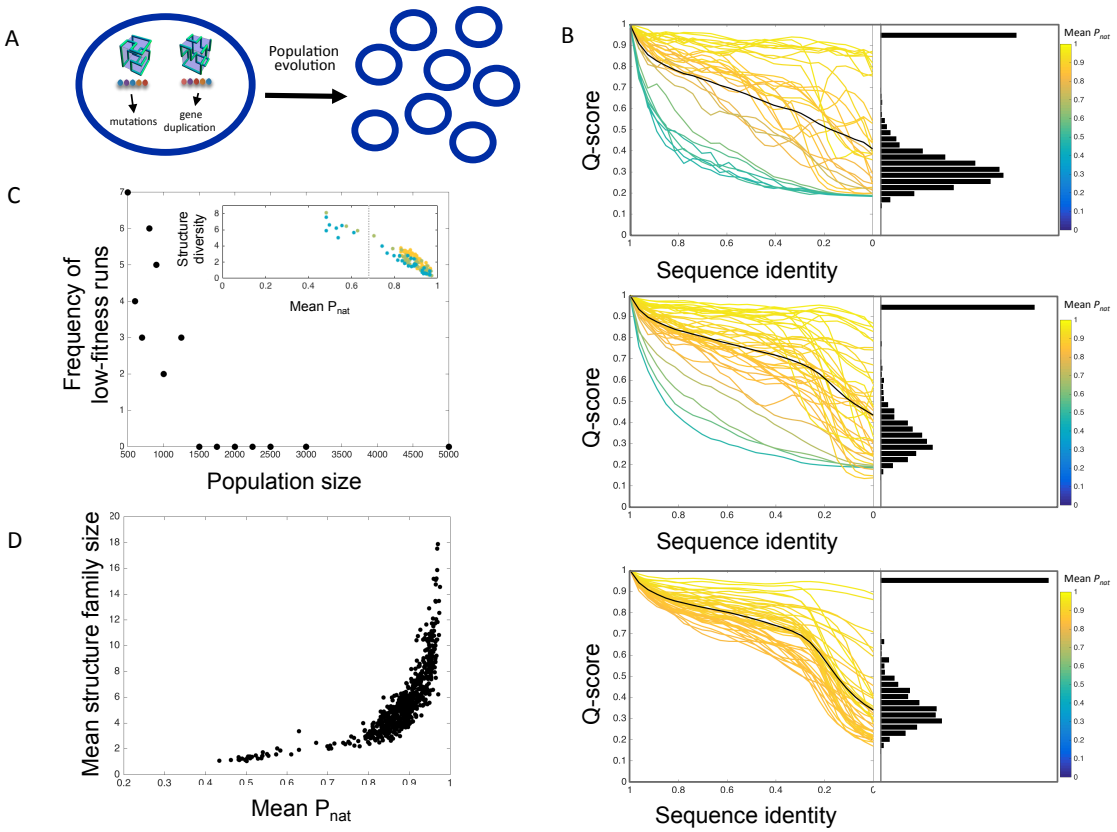

Figure 5: Population size modulates selection pressure (A) Schematic representation of the multi-scale evolution model. Each organism consists of genes that code for proteins and can be duplicated or acquire point mutations. Organisms evolve under selection pressure for protein folding stability. At the end of the simulation, sequence identity and Q-score are calculated for each pair of extant proteins in the population. (B) The number of replicas that evolved the full 3,000 generations yet failed to evolve stable proteins ($\overline{P}_{nat} < 0.68$) for each population size. Notably, above population sizes of 1,250 the selection pressure is apparently strong enough to drive evolution of stable proteins in each replica. Inset: the mean P_{nat} of proteins at the end of an evolutionary trajectory correlates with the diversity of structures. Dark green: population size =500, light green: population size = 1,250 and yellow: population size = 5,000 (C) The relationship between sequence divergence and structure evolution in population simulations. For each trajectory, the average Q-scores are plotted in a color indicating $\overline{P}_{nat}$ of proteins at the end of the simulation. The black line is the average over the individual trajectories. Top panel: Population size = 500. Middle panel: Population size = 1,250, Bottom panel: Population size = 5,000. Histograms showing the bimodal distribution of Q-scores are shown for each population size. Consistent with the results described above, in the realistic context of population dynamics, the characteristic cusp-like dependence of structural similarity (Q-score) on sequence identity only emerges at strong selection pressure. Furthermore, strong evolutionary pressure increases relative height of the high Q-score peak in the bimodal Q-score distribution relative to the low Q-score peak. (D) Strong evolutionary pressure increases the number of generations between fold discovery events and because it generates more stable proteins, is also likely to lead to larger

protein family sizes through gene duplication in this model. Indeed, the largest protein families arise in simulations where evolved model proteins have a high mean P_{nat} (r=0.9, p=0).

Discussion

We presented a simple analytical model of protein structure evolution that explains why there is a cusped relationship between structure and sequence evolution. Using the model, we were able to explain the role that contact density, a proxy for protein stability, plays in structure evolution in the bioinformatics protein record. Importantly, our work indicates that the cusped structure-sequence relationship with approximate 30% sequence similarity threshold is not the result of a biophysical constraint that 70% of a protein's residues must be changed in order for a new structure to emerge, but rather that continual sequence evolution under the constraint of protein folding stability degrades sequence identity over the time scale needed to find the mutations that could encode a switch to an alternative *stable* structure. Evolutionary simulations with model proteins strongly corroborate the model as well as the idea of a fitness barrier blocking fold evolution.

Intuitively, our results show that strong selection for stability hinders structure evolution, while weak selection permits rapid exploration of structure space, of predominantly unstable proteins. The bioinformatics and simulation data indicated that selection tunes how gradually the twilight zone begins, and this is consistent with our activated dynamics model of fold discovery. Yet it is curious that selection does not tune the position of the onset of the twilight zone, i.e. below 30% for proteins under strong selection pressure or above 30% for proteins under weak selection. In fact, we found that a different analytical model of fold creation, one based on the "gatekeeping" mechanism of fold determination *does* predict a shift in the twilight zone position (Fig. S8) (51, 52). The gatekeeping model of protein folds posits that several gatekeeping residues determine a protein's structure (53-55). Therefore we posited that mutations in these key residues may be necessary and sufficient for discovering a new fold. That we do not observe a moving twilight zone in the data indicates that either folds have roughly the same number of gatekeeper residues regardless of stability or that structure evolution is mediated globally rather than via a few key residues.

Along with tuning the rate of structure evolution, another effect of selection is to modulate the size of fold families, i.e. families of sequences that fold into the same native structure. This is because strong selection slows the pace of structure evolution more dramatically than sequence evolution, so during the time period a protein is "trapped" in a particular fold by selection, continued duplication and exploration of the sequences adopting that fold generates larger and larger fold families.

It has been clear for years that protein contact density is deeply connected to evolution. Proteins with high contact density tend to be old, part of larger gene families, and part of larger structural neighborhoods (38). The root of these observation has been attributed to intrinsic evolvability conferred upon a protein by its contact density, and it has been

hypothesized that young proteins may evolve quickly because they are under positive selection to evolve stability of new functions (56). Our work provides an alternative framework within which to interpret these findings. We suggest that pressure for folding stability causes proteins to evolve high contact density, severely constraints structure evolution, leading to larger gene families, and to the extent that its possible, selects for discovery of structures similar to the ancestral structure, leading to larger fold neighborhoods. Further bioinformatics and experimental studies may test this interpretation.

Methods

Structure divergence of proteins sharing a common SCOP fold

We use the set of all α, β, α+β, and α/β protein domains in SCOP including mutants (22). The contact density of each domain was calculated as a predictor of thermodynamic stability. When calculating contact density, we consider two residues in contact if any of their non-hydrogen residues are within 4.5 angstroms of each other. Proteins belonging to top (>4.93 contacts/residue), middle (4.57<CD<4.65 contacts/residue), and bottom 10% (<4.13 contacts/residue) with respect to contact density were studied further.

Following Chothia and Lesk (4), who studied the divergence of proteins within protein families, we studied the divergence of proteins classified into the same SCOP fold. By choosing a broader classification than family, we could track divergence over the long timescales over which significant structure evolution takes place. SCOP folds are extremely broadly defined so proteins sharing a fold classification often adopt significantly different structures. The structural similarity of protein domains classified in the same fold, in the same contact density class (both having high, intermediate, or low contact density) and comparable length (within 10 amino acids) were compared using the Template Modeling-align (TM-align) algorithm which yields a number representative of the similarity between two folds, ranging from zero to one (57). Then their percent sequence identity was calculated from their alignment by the Needleman-Wunsch algorithm and Blosum30 substitution matrix, implemented in MATLAB. Proteins classified in different SCOP folds do not share significant sequence homology or structural similarity and were therefore not compared. In Figures 1B-D and the corresponding Dataset 1, the data are subdivided by contact density alone, not by protein class or fold.

Fitting analytical model to bioinformatics data

In order to quantify the differences in structure evolutionary rates of real proteins apparent in Figure 1, we fit the analytical expression for structure evolution, equation (5), to the data presented in Figure 1B-D (where sequence and structure similarity is subdivided by contact density) as well as in Figure S1 (where the data is subdivided by structural class). We proceeded by binning the data into 50 bins spanning sequence identity of zero to one. This step was necessary to achieve a fit to the data because

otherwise the large majority of data points at low sequence identity and low structural similarity dominate and foreclose the possibility of a meaningful fit to the high sequence identity and cusp regions of the data. Because the twilight zone of the analytical model occurs at 5% sequence identity, the average sequence identity between two random sequences, while the bulk of proteins compared in the twilight zone share 20% sequence identity for real proteins, possibly reflecting that sequence alignment algorithms seek to maximize the overlap of sequence pairs being aligned, we added a parameter c to the equation in order to shift the analytical model twilight zone to the twilight zone of the data. The exact curve that was fitted to the data was

$$\tilde{S}(q) = 1 - \frac{1-a}{a}\left(\left(1 - \frac{a}{l(a-1)}\right)^{\frac{1-q}{k}} - 1\right) + c \tag{7}$$

where $\tilde{S}(q)$ is the sequence identity as a function of structure survival probability, a rearrangement of (5) plus c, the twilight zone shift. The parameter a is 20, the number of amino acid types, and l is the protein length, which is set to the average length of proteins in the dataset being fit. $\tilde{S}(q)$ rather than $q(S)$ was fit to the data in order to account for the twilight zone in more detail when using the binning procedure. Equation (7) was fit to the binned bioinformatics data using the program Igor, which optimized the parameters k and c for fit.

Protein model

We use standard 27-mer lattice-model proteins to simulate the structural evolution of proteins (58). Lattice proteins can fold into 103,346 fully compact structures for the 3x3x3 cubic lattice, and following (42), we use a representative subset of 10,000 randomly chosen structures for computational efficiency. There are 20 amino acid types in the model. The energy of a protein in any given structure can be computed from the Miyazawa Jernigan (MJ) potential (43), which contains interaction energies for each spatially proximal pair of any two amino acid types. The energy of a 27-mer sequence adopting a particular structure, S, is given by

$$E_S = \tfrac{1}{2}\sum_{ij=1}^{27} C_{ij}^{(S)} \mathrm{M}_{AA(i)AA(j)} \tag{8}$$

Where $C^{(S)}$ is the contact matrix of structure S, whose elements $C_{ij}^{(S)} = 1$ when residues i and j form a non-covalent contact and $C_{ij}^{(S)} = 0$ otherwise. M is the MJ interaction energy matrix such that the element $M_{AA(i)AA(j)}$ contains the interaction energy of the amino acid types of residues i and j. The Boltzmann probability, $P_{nat}(T)$, of a protein adopting its native state (lowest energy) structure is therefore given by the canonical partition function:

$$P_{nat}(T) = \frac{e^{-E_{nat}/T}}{\sum_{i=1}^{10,000} e^{-E_i/T}} \tag{9}$$

Where E_i is the protein's energy in structure i and T is the temperature in arbitrary units. The structure in which a protein's energy is minimized is considered the protein's native state.

The degree of structural homology between two model proteins is quantified using the number of contacts the structures have in common, normalized by the total number of contacts (28 contacts for all compact 27-mer model proteins) (46, 57).

$$Q_{1,2} = \frac{1}{28}\sum_{ij=1}^{27} C_{ij}^{(1)} C_{ij}^{(2)} \tag{10}$$

Where $C^{(1)}$ and $C^{(2)}$ are the contact matrices of the two structures being compared such that $C_{ij} = 1$ if residues i and j are in contact and 0 otherwise, excluding covalently linked residues as in (8).

Monte Carlo algorithm for divergent evolution

The model proteins were evolved under selection for folding stability, P_{nat}. This constraint captures two biological features. First, most proteins must be folded to carry out their function (the obvious exception is intrinsically disordered proteins). Second, unfolded or misfolded species can be toxic (36, 59, 60). Each simulation of model protein evolution proceeded in two steps. First, each simulation was initialized with a protein stably adopting a particular structure ($P_{nat} > 0.99$). Stable proteins were made by generating a random 27-mer amino acid sequences, introducing random mutations, and accepting the mutations only if they stabilized a predetermined ground state (61). This procedure, rather than a Monte Carlo procedure, is sufficient for generating unique, stable sequences for each structure. Then, evolutionary trajectories at different selection pressures for folding stability were carried out as follows. Each generation, the protein was subjected to a point mutation. The fitness effect of the point mutation was defined as follows:

$$f = P_{nat}^{(trial)} - P_{nat}^{(original)} \tag{11}$$

Where $P_{nat}^{(trial)}$ is the stability of the protein with the mutation and $P_{nat}^{(original)}$ is the stability of the protein prior to the mutation. Any neutral mutation or mutation increasing fitness was accepted while destabilizing mutations were accepted or rejected according to the

Metropolis criterion with f/T_{sel}. The selective temperature, controls the stringency of selection for protein stability; the more destabilizing a mutation, the less likely it is to be accepted (61). T_{sel} directly tunes the equilibrium P_{nat} of evolving proteins (SI, Fig. 1 and Table 1). Simulations ran for 1,000 mutation attempts, illustrated in Fig. 3A and the structure, sequence, and stability (P_{nat}) were recorded every 10 generations. This was repeated for approximately 2,750 different starting structures at different strengths of selection for stability P_{nat} (set by T_{sel}).

Monte Carlo algorithm for convergent evolution

In these simulations, many proteins all initially adopting different structures with high $\overline{P}_{nat}$ evolve under selection pressure to maximize their sequence identities with respect to the same designated "target" sequence. The target sequence is a randomly selected sequence with a thermodynamically stable ground state. To simulate the regime where selection pressure for folding stability is low (Figure 4A), any mutation increasing sequence identity with respect to the target sequence ($SID_{i,\text{target}}$) was accepted. No selection pressure for folding stability was applied. The regime where selection pressure for folding stability is high was simulated by using the fitness function $SID_{i,\text{target}} \times P_{nat}$ (Figure 4C). A single convergent evolution simulation was run in each selection regime, consisting of 100 proteins converging over 3,000 generations in the weak selection regime while 200 protein converging over 3,000 generation were used in the strong selection regime to compensate for slower evolution and improve statistics. To determine the sequence-structure relationship for convergent evolution, the sequence identity and Q-score of the converging proteins were calculated every 100 generations within each regime.

Multi-scale modeling of protein evolution

We used the multi-scale simulation algorithm developed and described in detail in Zeldovich et al. 2007 (48). Because this evolutionary procedure yields fold families of sizes matching the distributions observed in nature, it is an ideal source of comparison with SCOP data. The simulations were initialized with 100 identical organisms, each consisting of a single gene (81-mer DNA) with random sequence and its 27-mer gene product. (Thus, an average initial protein has a $P_{nat}^{rand} = 0.23$ corresponding to average stability of random sequences) At each step in evolutionary time, one of five fates affects each organism with equal probability (1) no event (2) point mutation at a random position in the genome to a random nucleotide at rate $m = 0.3$ per unit time per base pair (3) gene duplication with rate $\mu = 0.03$ of a random gene in the organism's genome (4) death of the organism with rate d as given by $d = d_0(1 - \min_i P_{nat}^{(i)})$, and (5) division of the

organism into two equivalent daughter cell with birthrate $b = 0.15$. Thus, for example, there is a 20% chance that gene duplication is selected as an organism's possible fate and a further probability u that a gene will actually be duplicated. As the population evolves, the population size may increase until it reaches a carrying capacity, $N_{\max}$. Once the maximum population size is attained it is maintained by removing $N - N_{\max}$ randomly chosen organisms from the population each generation. We modulate the strength of selection in the population by simulating populations with size $N_{\max}$ set at 500, 600, 700, 800, 900, 1000, 1250, 1500, 1750, 2000, 2250, 2500, 3000, and 5000.

Fifty evolutionary trajectories were run at each population size. As the simulation progresses, proteins evolve, becoming more stable and periodically changing structure, spawning new protein families. Trajectories were terminated after 3,000 generations or when the population went extinct. Only simulations that ran the full 3,000 generations were analyzed, and even among these evolutionary runs, in some cases, the proteins did not evolve high $\overline{P}_{nat}$ (Fig. 5B-C). For our analysis of structure divergence, we calculate the pairwise sequence identity and Q-score of each extant pair of proteins harvested from a particular evolutionary run.

Acknowledgements

We thank the Kavli Institute for Theoretical Physics Quantitative Biology Summer School, which A.I.G. attended in 2015 and where she had many useful discussion, especially with Ned Wingreen. This work was supported by NIH grant GM068670 and a National Science Foundation Graduate Research Fellowship (awarded to A.I.G.).

References

1. Zhang J & Yang J-R (2015) Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics* 16(7):409-420.10.1038/nrg3950
2. Xia Y, Franzosa EA, & Gerstein MB (2009) Integrated assessment of genomic correlates of protein evolutionary rate. *PLoS computational biology* 5(6):e1000413-e1000413.10.1371/journal.pcbi.1000413
3. Sadowski MI & Taylor WR (2010) Protein structures, folds and fold spaces. *Journal of physics. Condensed matter : an Institute of Physics journal* 22(3):033103-033103.10.1088/0953-8984/22/3/033103
4. Chothia C & Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. *The EMBO journal* 5(4):823-826
5. Wilson CA, Kreychman J, & Gerstein M (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *Journal of molecular biology* 297(1):233-249.10.1006/jmbi.2000.3550
6. Rost B (1999) Twilight zone of protein sequence alignments. *Protein engineering* 12(2):85-94.10.1093/protein/12.2.85
7. Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, & Schwede T (2008) Protein structure homology modeling using SWISS-MODEL workspace. *Nature Protocols* 4(1):1-13.10.1038/nprot.2008.197
8. Marti-Renom MAS, Ashley C.; Fiser, Andras; & Sanchez RM, Francisco ; Sali, Andrej (2000) Comparative Protein Structure Modeling of Genes and Genomes. *Annual Review of Biophysics and Biomolecular Structure* 29:291-325
9. Chung SY & Subbiah S (1996) A structural explanation for the twilight zone of protein sequence homology. *Structure* 4(10):1123-1127.10.1016/S0969-2126(96)00119-0
10. Tokuriki N, Stricher F, Schymkowitz J, Serrano L, & Tawfik DS (2007) The stability effects of protein mutations appear to be universally distributed. *J Mol Biol* 369(5):1318-1332.10.1016/j.jmb.2007.03.069
11. Dokholyan NV, Shakhnovich B, & Shakhnovich EI (2002) Expanding protein universe and its origin from the biological Big Bang. *Proceedings of the National Academy of Sciences of the United States of America* 99(22):14132-14136.10.1073/pnas.202497999
12. Deeds EJ, Dokholyan NV, & Shakhnovich EI (2003) Protein evolution within a structural space. *Biophysical journal* 85(5):2962-2972.10.1016/S0006-3495(03)74716-X
13. Bourne PE & Koonin EV (2011) Are There Laws of Genome Evolution? *PLoS Computational Biology* 7(8):e1002173.10.1371/journal.pcbi.1002173
14. Roessler CG, Hall BM, Anderson WJ, Ingram WM, Roberts Sa, Montfort WR, & Cordes MHJ (2008) Transitive homology-guided structural studies lead to discovery of Cro proteins with 40% sequence identity but different folds. *Proceedings of the National Academy of Sciences of the United States of America* 105(7):2343-2348.10.1073/pnas.0711589105

15. Alexander Pa, He Y, Chen Y, Orban J, & Bryan PN (2009) A minimal sequence code for switching protein structure and function. *Proceedings of the National Academy of Sciences of the United States of America* 106(50):21149-21154.10.1073/pnas.0906408106
16. Burmann BM, Knauer SH, Sevostyanova A, Schweimer K, Mooney RA, Landick R, Artsimovitch I, & Rösch P (2012) An α helix to β barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell* 150(2):291-303.10.1016/j.cell.2012.05.042
17. Eaton KV, Anderson WJ, Dubrava MS, Kumirov VK, Dykstra EM, & Cordes MHJ (2015) Studying protein fold evolution with hybrids of differently folded homologs. *Protein Engineering Design and Selection* 28(8):241-250.10.1093/protein/gzv027
18. Alexander Pa, He Y, Chen Y, Orban J, & Bryan PN (2007) The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proceedings of the National Academy of Sciences of the United States of America* 104(29):11963-11968.10.1073/pnas.0700922104
19. He Y, Chen Y, Alexander PA, Bryan PN, & Orban J (2012) Mutational tipping points for switching protein folds and functions. *Structure* 20(2):283-291.10.1016/j.str.2011.11.018
20. Meyerguz L, Kleinberg J, & Elber R (2007) The network of sequence flow between protein structures. *Proceedings of the National Academy of Sciences of the United States of America* 104(28):11627-11632.10.1073/pnas.0701393104
21. Burke S & Elber R (2012) Super folds, networks, and barriers. *Proteins* 80(2):463-470.10.1002/prot.23212
22. Hubbard TJP, Ailey B, Brenner SE, Murzin AG, & Chothia C (1999) SCOP: A structural classification of proteins database. *Nucleic Acids Research* 27(1):254-256.10.1093/nar/27.1.254
23. Zhang Y & Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function and Genetics* 57(4):702-710.10.1002/prot.20264
24. Bloom JD, Raval A, & Wilke CO (2007) Thermodynamics of neutral protein evolution. *Genetics* 175(1):255-266
25. Zeldovich KB, Chen P, & Shakhnovich EI (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc Natl Acad Sci U S A* 104(41):16152-16157
26. Serohijos AW & Shakhnovich EI (2014) Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics. *Curr Opin Struct Biol* 26:84-91.10.1016/j.sbi.2014.05.005
27. Bershtein S, Mu W, Serohijos AW, Zhou J, & Shakhnovich EI (2013) Protein quality control acts on folding intermediates to shape the effects of mutations on organismal fitness. *Molecular cell* 49(1):133-144.10.1016/j.molcel.2012.11.004
28. Jacquier H, Birgy A, Le Nagard H, Mechulam Y, Schmitt E, Glodt J, Bercot B, Petit E, Poulain J, Barnaud G, Gros PA, & Tenaillon O (2013) Capturing the mutational landscape of the beta-lactamase TEM-1. *Proc Natl Acad Sci U S A*.10.1073/pnas.1215206110

29. Berezovsky IN & Shakhnovich EI (2005) Physics and evolution of thermophilic adaptation. *Proceedings of the National Academy of Sciences of the United States of America* 102(36):12742-12747.10.1073/pnas.0503890102
30. Dill KA (1990) Dominant forces in protein folding. *Biochemistry* 29(31):7133-7155.10.1021/bi00483a001
31. Dill KAA, Darwin O.V.; Hutchinson, Karen (1989) Thermal Stabilities of Globular Proteins. *Biochemistry* 28:5439-5449
32. Jones DT, Moody CM, Uppenbrink J, Viles JH, Doyle PM, Harris CJ, Pearl LH, Sadler PJ, & Thornton JM (1996) Towards meeting the Paracelsus Challenge: The design, synthesis, and characterization of paracelsin-43, an alpha-helical protein with over 50% sequence identity to an all-beta protein. *Proteins* 24(4):502-513.10.1002/(SICI)1097-0134(199604)24:4<502::AID-PROT9>3.0.CO;2-F
33. Dalal S, Balasubramanian S, & Regan L (1997) Protein alchemy: Changing β-sheet into α-helix. *Nature Structural Biology* 4(7):548-552.10.1038/nsb0797-548
34. Dalal S & Regan L (2000) Understanding the sequence determinants of conformational switching using protein design. *Protein Science* 9(09):1651-1659.10.1110/ps.9.9.1651
35. Dokholyan NV & Shakhnovich EI (2001) Understanding hierarchical protein evolution from first principles. *Journal of molecular biology* 312(1):289-307.10.1006/jmbi.2001.4949
36. Drummond DA & Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134(2):341-352.10.1016/j.cell.2008.05.042
37. Sella G & Hirsh AE (2005) The application of statistical physics to evolutionary biology. *Proceedings of the National Academy of ...* 102(27)
38. Shakhnovich BE, Deeds E, Delisi C, & Shakhnovich E (2005) Protein structure and evolutionary history determine sequence space topology Protein structure and evolutionary history determine sequence space topology. *Genome Research*:385-392.10.1101/gr.3133605
39. Shakhnovich E & Gutin A (1990) Enumeration of all compact conformations of copolymers with random sequence of links. *The Journal of Chemical Physics* 93(8):5967-5967.10.1063/1.459480
40. Shakhnovich EI & Gutin aM (1993) Engineering of stable and fast-folding sequences of model proteins. *Proceedings of the National Academy of Sciences of the United States of America* 90(15):7195-7199.10.1073/pnas.90.15.7195
41. Li H, Helling R, Tang C, & Wingreen N (1996) Emergence of Preferred Structures in a Simple Model of Protein Folding. *Science, New Series* 273(5275):666-669.10.1126/science.273.5275.666
42. Heo M, Maslov S, & Shakhnovich E (2011) Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. *Proceedings of the National Academy of Sciences of the United States of America* 108(10):4258-4263.10.1073/pnas.1009392108
43. Miyazawa S & Jernigan RL (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of molecular biology* 256(3):623-644.10.1006/jmbi.1996.0114

44. Shakhnovich EI & Gutin AM (1993) Engineering of stable and fast-folding sequences of model proteins. *Proc Natl Acad Sci U S A* 90(15):7195-7199
45. Ramanathan S & Shakhnovich E (1994) Statistical mechanics of proteins with "evolutionary selected" sequences. *Physical Review. E. Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* 50(2):1303-1312
46. Sali A, Shakhnovich E, & Karplus M (1994) Kinetics of Protein Folding: A Lattice Model Study of Requirements for Folding to the Native State. *Journal of molecular biology* 235:1614-1636
47. Chandler D (1986) Roles of Classical Dynamics and Quantum Dynamics on Activated Processes Occuring in Liquids. *Journal of Statistical Phyics* 42
48. Zeldovich KB, Chen P, Shakhnovich BE, & Shakhnovich EI (2007) A first-principles model of early evolution: emergence of gene families, species, and preferred protein folds. *PLoS computational biology* 3(7):e139-e139.10.1371/journal.pcbi.0030139
49. Daniel L. Hartl AGC (2006) *Principles of Population Genetics* (Sinauer Associates, Sunderland, Massachusetts) Fourth Edition Ed.
50. Shakhnovich BEDEDCSEI (2005) Protein structure and evolutionary history determine sequence space topology. *Genome Research* 15(3):385-392.10.1101/gr.3133605
51. Vendruscolo M, Paci E, Dobson CM, & Karplus M (2001) Three key residues form a critical contact network in a protein folding transition state. *Nature* 409(6820):641-645.10.1038/35054591
52. Vendruscolo M, Dokholyan NV, Paci E, & Karplus M (2002) Small-world view of the amino acids that play a key role in protein folding. *Phys Rev E Stat Nonlin Soft Matter Phys* 65(6 Pt 1):61910-61910.10.1103/PhysRevE.65.061910
53. Stoycheva AD, Brooks CL, 3rd, & Onuchic JN (2004) Gatekeepers in the ribosomal protein s6: thermodynamics, kinetics, and folding pathways revealed by a minimalist protein model. *J Mol Biol* 340(3):571-585.10.1016/j.jmb.2004.04.073
54. Kurnik M, Hedberg L, Danielsson J, & Oliveberg M (2012) Folding without charges. *Proc Natl Acad Sci U S A* 109(15):5705-5710.10.1073/pnas.1118640109
55. Mirny LA & Shakhnovich EI (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J Mol Biol* 291(1):177-196
56. Toll-Riera M, Bostick D, Albà MM, & Plotkin JB (2012) Structure and age jointly influence rates of protein evolution. *PLoS computational biology* 8(5):e1002542-e1002542.10.1371/journal.pcbi.1002542
57. Zhang Y & Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* 33(7):2302-2309.10.1093/nar/gki524
58. Shakhnovich EI & Gutin AM (1990) Enumeration of all compact conformations of copolymers with random sequence of links. *Journal of Chemical Physics* 93(8):5967-5971
59. Lobkovsky AE, Wolf YI, & Koonin EV (2010) Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proceedings of the*

National Academy of Sciences of the United States of America 107(7):2983-2988.10.1073/pnas.0910445107
60. Serohijos AWR, Rimas Z, & Shakhnovich EI (2012) Protein biophysics explains why highly abundant proteins evolve slowly. *Cell reports* 2(2):249-256.10.1016/j.celrep.2012.06.022
61. Zeldovich KB, Berezovsky IN, & Shakhnovich EI (2006) Physical Origins of Protein Superfamilies. *Journal of Molecular Biology* 357(4):1335-1343.10.1016/j.jmb.2006.01.081
62. Mirny LA & Shakhnovich EI (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *Journal of molecular biology* 291(1):177-196.10.1006/jmbi.1999.2911
63. Abkevich VI, Gutin AM, & Shakhnovich EI (1994) Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* 33:10026-10036.10.1021/bi00199a029

Supporting Information

The role of evolutionary selection in the dynamics of protein structure evolution

Amy I. Gilson[1], Ahmee Marshall-Christensen[1], Jeong-Mo Choi[1], Eugene I. Shakhnovich[1,*]

[1]Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford St., Cambridge MA, USA

* Correspondence should be addressed to E.I.S. shakhnovich@chemistry.harvard.edu

Derivation of the analytical model of protein evolution with structure evolution barriers

We begin by deriving an expression for the Hamming distance, $S(t)$, between a protein at time t and its ancestral sequence at time $t=0$, i.e. the number of derived rather than ancestral amino acid types comprising the protein at time t. A relationship for the expected Hamming distance between the evolving protein at time $t+1/\mu$ with respect to the ancestral protein is given by

$$S(t+1/\mu)=\frac{l-S(t)}{l}(S(t)+1)+\frac{S(t)}{l}\left(\frac{1}{a-1}(S(t)-1)+\frac{a-2}{a-1}S(t)\right) \qquad \text{(S1)}$$

where l is the length of the protein and a is the number of amino acid types (typically 20) and μ is the mutation rate of the protein. Each of the three terms in (S1) captures one of the three possible effects of a mutation. First, a mutation may increase $S(t)$ by one if it substitutes a residue in its ancestral state with a different amino acid type. Second, a mutation may revert a residue from a derived amino acid type back to its ancestral type, decreasing $S(t)$ by one. Finally, $S(t)$ remains constant if a mutation substitutes a residue with a derived amino acid type with another non-ancestral amino acid type. This expression simplifies to

$$S(t+1/\mu)=1+\frac{(a(l-1)-1)S(t)}{al} \qquad \text{(S2)}$$

Setting $a_k=S(k/\mu)$, we obtained a mathematic sequence, a closed form for which can be found by using generating functions. Define $A(x):=\sum_{k=0}^{\infty}a_k x^k$. Noting that $a_0=S(0)=0$, we have

$$A(x)=\sum_{k=1}^{\infty}a_k x^k=\sum_{k=1}^{\infty}\left(1+\frac{(a(l-1)-1)a_{k-1}}{al}\right)x^k \tag{S3}$$

$$A(x)=\sum_{k=1}^{\infty}x^k+\sum_{k=1}^{\infty}\left(\frac{(a(l-1)-1)a_{k-1}}{al}\right)x^k \tag{S4}$$

$$A(x)=\sum_{k=1}^{\infty}x^k+\sum_{k=1}^{\infty}\left(\frac{(a(l-1)-1)}{al}\right)x\sum_{k=1}^{\infty}a_{k-1}x^{k-1} \tag{S5}$$

$$A(x)=\frac{x}{1-x}+\frac{a(l-1)-1}{al}xA(x) \tag{S6}$$

This implied the closed form $A(x)=\left(\frac{x}{1-x}\right)/\left(1-\frac{a(l-1)-1}{al}x\right)$.

Expanding the function into a power series yields

$$A(x)=\sum_{k=0}^{\infty}\frac{al}{1+a}\left(1-\left(1-\frac{a+1}{al}\right)^k\right)x^k \tag{S7}$$

This implied $a_k=\frac{al}{1+a}\left(1-\left(1-\frac{a+1}{al}\right)^k\right)$, and recalling that $S(k/\mu)=S(t)$, we concluded

$$S(t)=\frac{al}{1+a}\left(1-\left(1-\frac{a+1}{al}\right)^{\mu t}\right) \tag{S8}$$

Rearranging and setting $\mu=1$,

$$S(t)=\frac{l(a-1)}{a}\left(1-\left(1-\frac{a}{l(a-1)}\right)^t\right) \tag{S9}$$

From the Arrhenius equation, we have that the rate of structure change

$$k \propto e^{-\Delta G_f / T_{sel}} \tag{S10}$$

where ΔG_f is the "free energy" barrier to evolution of a new structure and T_{sel} is the evolutionary selection temperature. This gave us an expression for probability that the structure will not yet have changed at time t (i.e. the structure survival probability).

$$q(t) = e^{-kt} \tag{S11}$$

The expression for $S(t)$ can be rearranged to solve for t which will permit us to write a parametric equation for structure evolution versus sequence evolution.

$$t = \frac{\log\left(1 - S(t)\frac{a}{l(a-1)}\right)}{\log\left(1 - \frac{a}{l(a-1)}\right)} \tag{S12}$$

Thus, the expression for structure evolution with respect to sequence divergence is

$$q(S) = \exp\left(-k\frac{\log\left(1 - S\frac{a}{l(a-1)}\right)}{\log\left(1 - \frac{a}{l(a-1)}\right)}\right) \tag{S13}$$

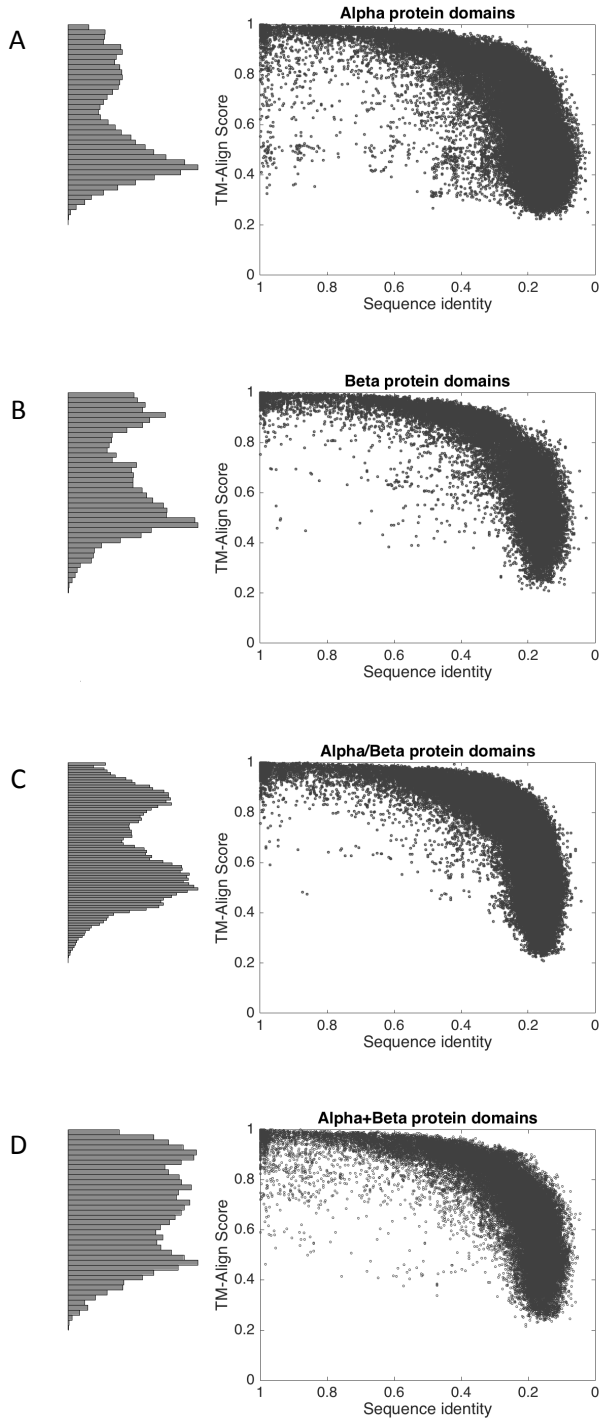

Fig. S1. Structural divergence of α, β, α/β, and α+β protein domains from SCOP. Sequence identity and structural homology (TM-align score) for all the domains classified into the same fold and with similar length (within 10 residues) were calculated and are plotted for (A) α (N=59,903 data points) (B) β (N=25,854 data points) (C) α/β (N=89,730 data points) and (D) α+β (N=41,264 data points) domains. Histograms at left are projections of the data onto the y-axis and are markedly bimodal across classes except, perhaps, for α+β domains. Furthermore, there is a pronounced cusped relationship between sequence and structure divergence in each class.

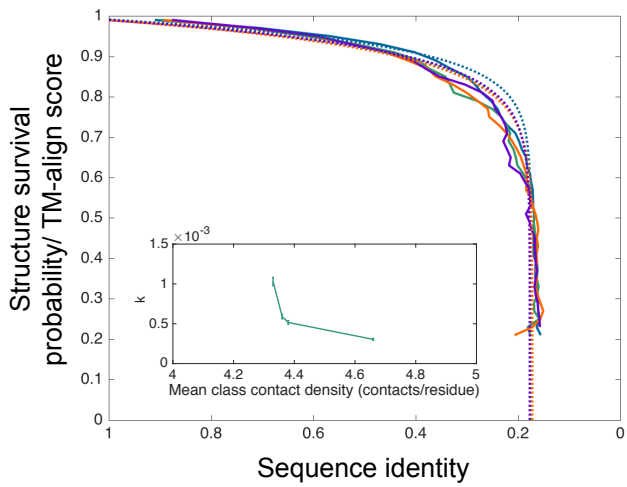

Fig. S2. Fit of the analytical model to data from each protein domain classes: α (purple), β (green), α/β (blue), and α+β (orange). The model fit to the data is presented in dashed lines while binned bioinformatics data is presented in solid lines. Even though there is not a clearly discernable difference among classes in the rate of structure divergence with respect to sequence divergence, the inset show that there are in fact different rates of structure evolution, k, across protein domains and that these differences can be explained by different mean contact densities of each class. Error bars indicate one standard deviation.

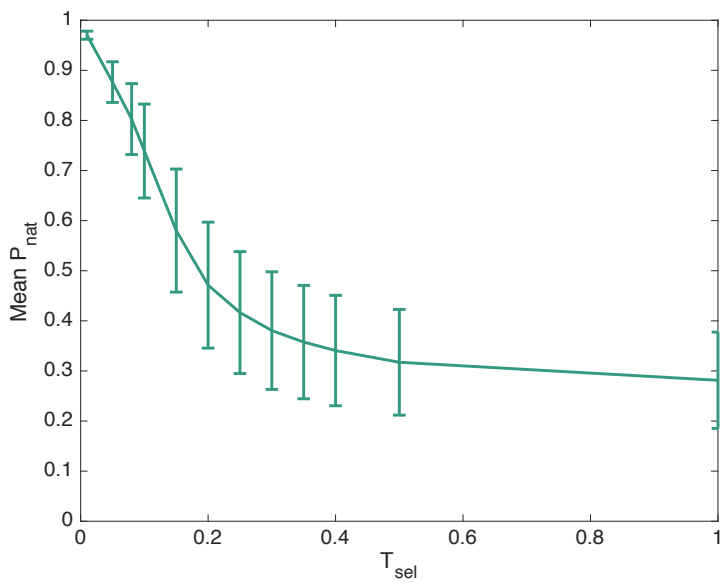

Fig. S3. The Monte Carlo simulation parameter, T_{sel}, sets the probability that a deleterious, destabilizing mutation is accepted. This adjusts the mean P_{nat} at which proteins evolve. Error bars span one standard deviation in recorded P_{nat} values, where the standard deviation has been averaged over runs at each T_{sel}. See also Dataset S2.

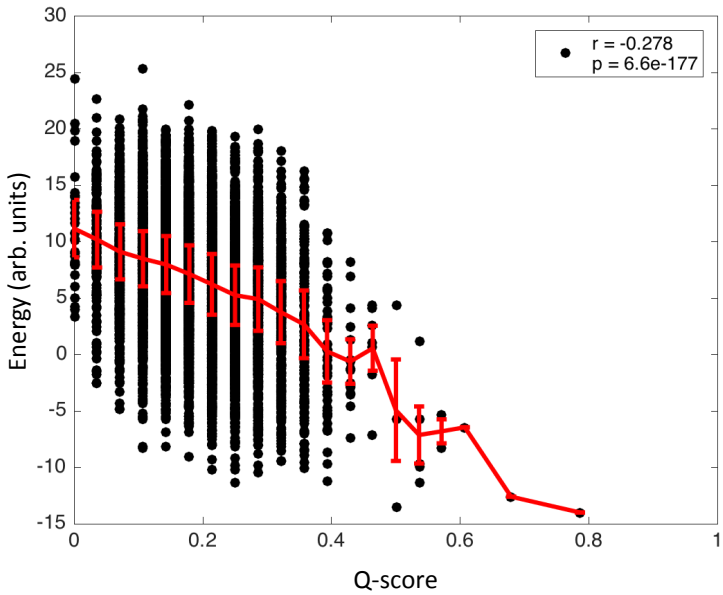

Fig. S4. Low-energy conformations of a sequence with a stable native structure tend to be similar (high Q-score) to the sequence's native structure. However note a broad distribution of energies with low Q-score indicating the existence of low energy misfolds that are dissimilar to the native structure.

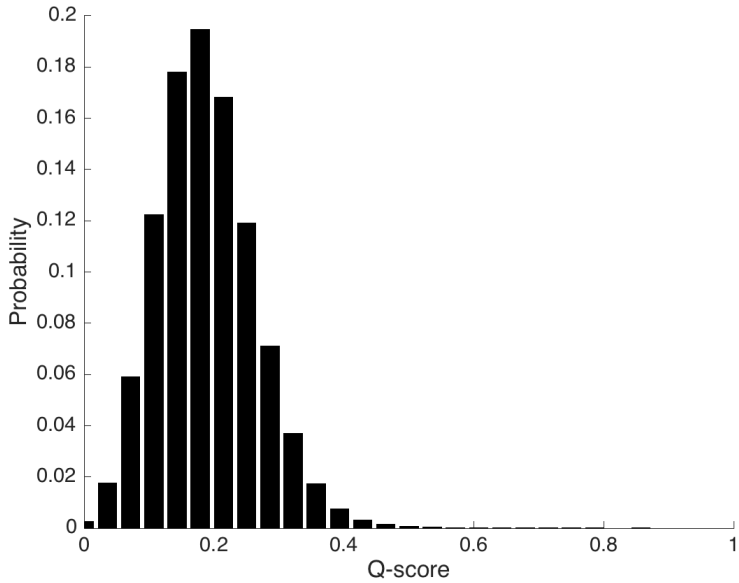

Fig. S5. Distribution of Q-scores for all pairs of the 10,000 structures used in this study.

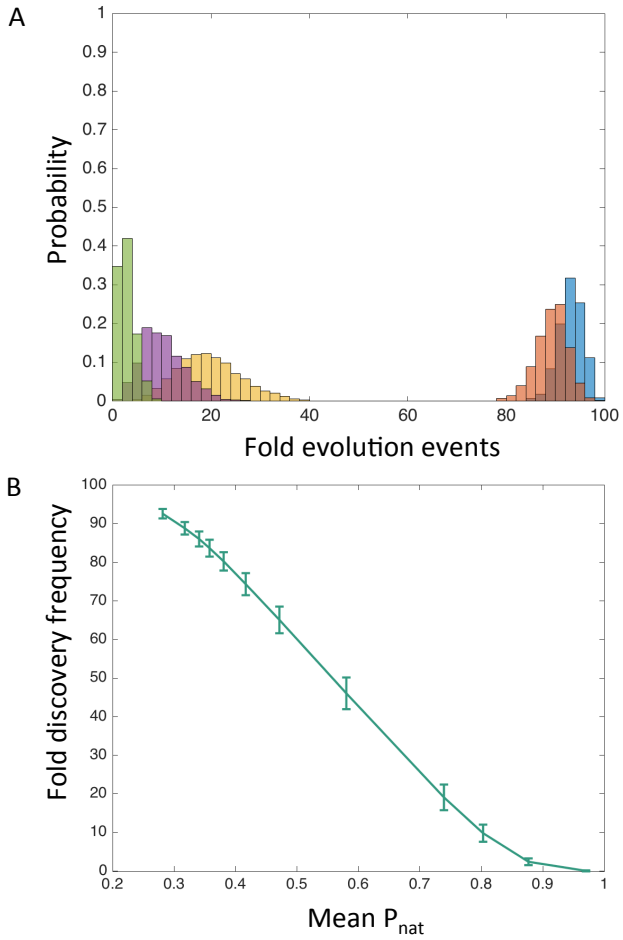

Fig S6. Fold Discovery Events (A) The distribution of fold evolution events over hundreds of different evolutionary runs at five different $\overline{P}_{nat}$, $\overline{P}_{nat} = 0.88$ (green), $\overline{P}_{nat} = 0.80$ (purple), $\overline{P}_{nat} = 0.74$ (yellow), $\overline{P}_{nat} = 0.32$ (orange), $\overline{P}_{nat} = 0.28$ (blue). The maximum number of fold evolution events recorded is 100 because every ten generations for 1,000 generations was recorded. (B) The number of fold discovery events averaged over evolutionary simulations run at the same selection strength. Error bars span one standard deviation.

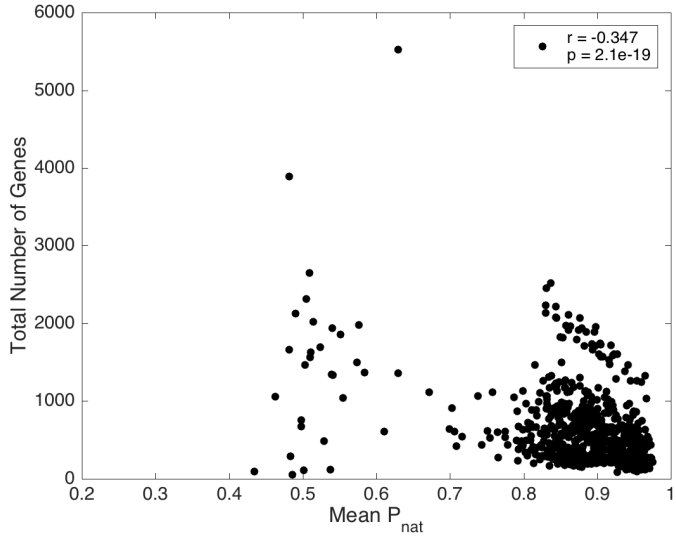

Fig. S7. The total number of genes in a population at the end of a population evolution simulation correlates negatively with the mean stability ($\overline{P}_{nat}$) of proteins in the population at the end of the simulation. This indicates that the positive correlation between stability and family size presented in Figure 5D is not due to differing numbers of overall proteins across simulations.

Derivation of an analytical model of protein evolution with key gatekeeper residues

Here, we build a phenomenological theory of protein structure evolution with additional structural, mechanistic detail. Consider a protein of length l residues where each amino acid can mutate to any of the other $m=19$ amino acids. The protein's structure is determined by n key residues (62, 63). A structure evolution event occurs each time that all n key residues mutate to any of v amino acid types that encode a different fold, as illustrated in Fig S8A. The parameters n and v reflect the degree of selection pressure under which a protein is evolving. As before, the sequence evolution is captured by equation (S9).

Rigidly fixing the number of key residues sets a hard lower bound on the minimum number of mutations before a fold evolution event can occur. Assuming that fold evolution is a Poisson process governed by mutation of key residues, we derived the "survival" probability that a fold evolution event has not occurred after time T (below).
With this model, both the shape and position of the twilight zone are adjusted by the strength of evolutionary selection. The requirement that all key residues must mutate before the structure evolves reflects that epistasis among the structures n key residues and that extensive exploration of sequence space may be necessary before a sequence coding for a stable mutant fold is discovered.

Derivation of structure survival probability time ($F(t)$)

An analytic solution for the probability that the protein has undergone a structural change is derived here. For this, we will first need to calculate the characteristic wait time between fold evolution events. After i "good" mutations (mutations of designated residues to any of the v other amino acid types that encode another fold), the probability of a mutation being a good one is given by $P_i(good) = \frac{(n-1)v}{(l-1)m}$. The time for the protein to change structure is $T = (1/\mu)\sum_{i=1}^{n} 1/P_i$. If $H_n = \sum_{i=1}^{n} 1/i$ is the n th harmonic number, then

$$T = \frac{mn}{\mu v}(1 + H_n p/n - H_n) \tag{S14}$$

Now assume that after a protein changes structure, the clock "resets." That is, after a structural change, the model is applied to the protein as if it were an ancestral state. Looking over some interval of time Δt, we expect that the protein has undergone a structural change event $\Delta t/T$ times. Therefore, after cT time has elapsed, the expected number of structural changes is c. Now, letting the time domain be infinite, we can divide it into intervals of length cT, and the number of changes in each interval should be random, roughly independent, and on average c. The lack of independence is because there is a minimum number of time steps between structural transitions, so a transition that happens at the end of one interval will prohibit a transition at the beginning of the next. However if selection is strong, the probability distribution of the number of structural change events in an interval cT should be roughly Poisson

$$P(k \text{ changes in } cT) \approx \frac{c^k}{k!}e^{-c} \tag{S15}$$

If there are k changes, the average wait time t between transitions is well-approximated by cT/k, so we have that

$$P(\text{average wait time } cT/k) \approx \frac{c^{(cT/t)}}{(cT/t)!}e^{-c} \tag{S16}$$

Assume now, that cT/t is continuous. If we choose c such that $ce^{-c} \ll 1$, there will be only a negligible probability that k is between 0 and 1, which implies that there is a negligible chance that τ is greater than cT. From this we can obtain a continuous cumulative distribution function (Fig S8B).

$$F(\text{structure change after time } t) = \frac{\int_0^t \frac{c^{cT/\tau}}{\Gamma(cT/\tau+1)}e^{-c}\,d\tau}{\int_0^{ct} \frac{c^{cT/\tau}}{\Gamma(cT/\tau+1)}e^{-c}\,d\tau} \tag{S17}$$

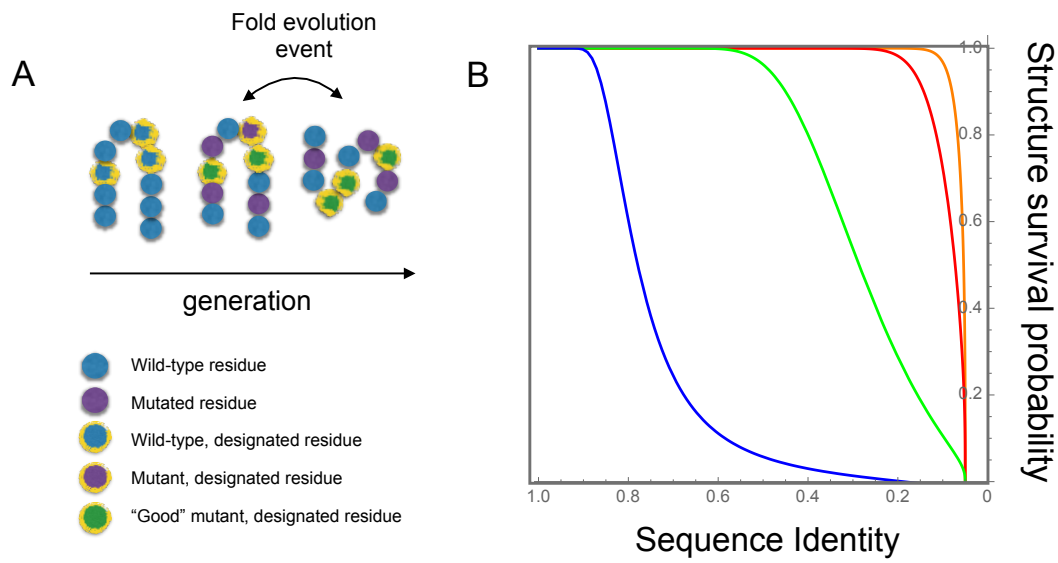

Fig. S8. Protein evolution with key gatekeeper residues (A) No structure or ordered sequence is explicitly modeled. Rather, in a protein of length p, n residues are designated as residues crucial for fold stability (indicated with yellow edge). ν amino acid types out of 19 are considered "good." Once all designated residues acquire a "good" mutation (shown in green), we consider a fold discovery event to have taken place. In this model, n and ν tunes the selection pressure on fold evolution. From weakest to strongest selection pressure: $n = 1$, $\nu = 19$ (blue), $n = 10$, $\nu = 10$ (green), $n = 15$, $\nu = 5$ (red), $n = 27$, $\nu = 5$ (orange).

Table S1. Physical properties and analytical fitting of protein domains

Protein domain category	Average domain length (number of residues)	Mean contact density (contacts/length)	k (s^{-1})	c (percent sequence identity)
Low Contact Density (<4.13 contacts/residue)	83.06	3.88	1.74E-03	0.10
Intermediate Contact Density (4.57<CD<4.65 contacts/residue)	174.67	4.61	5.27E-04	0.12
High Contact Density (>4.93 contacts/residue)	336.67	5.03	2.31E-04	0.12
α class domains	74.96	4.33	1.03E-03	0.13
β class domains	154.29	4.38	5.16E-04	0.12
α / β class domains	221.29	4.66	3.04E-04	0.12
α + β class domains	139.65	4.36	5.90E-04	0.12

The average domain length and contact density of protein in each subgrouping considered along with the parameters of the analytical model fit to the datasets. The parameter k is the rate of protein structure evolution and c modulates the position of the twilight zone.