Biophysical Journal
Article

Biophysical Society

The Role of Evolutionary Selection in the Dynamics of Protein Structure Evolution

Amy I. Gilson,[1] Ahmee Marshall-Christensen,[1] Jeong-Mo Choi,[1] and Eugene I. Shakhnovich[1,*]
[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts

ABSTRACT Homology modeling is a powerful tool for predicting a protein's structure. This approach is successful because proteins whose sequences are only 30% identical still adopt the same structure, while structure similarity rapidly deteriorates beyond the 30% threshold. By studying the divergence of protein structure as sequence evolves in real proteins and in evolutionary simulations, we show that this nonlinear sequence-structure relationship emerges as a result of selection for protein folding stability in divergent evolution. Fitness constraints prevent the emergence of unstable protein evolutionary intermediates, thereby enforcing evolutionary paths that preserve protein structure despite broad sequence divergence. However, on longer timescales, evolution is punctuated by rare events where the fitness barriers obstructing structure evolution are overcome and discovery of new structures occurs. We outline biophysical and evolutionary rationale for broad variation in protein family sizes, prevalence of compact structures among ancient proteins, and more rapid structure evolution of proteins with lower packing density.

INTRODUCTION

A wide variety of protein structures exist in nature, but 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). Despite this, Wang et al. (4) showed that protein structure evolution is a continual process, proceeding in a molecular clocklike fashion with new folds emerging regularly on a billion-year timescale.

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 identities (or another measure of sequence similarity) and structural similarities can be assessed (5–9). The result is a cusped relationship between sequence and structure divergence: sequences reliably diverge up to 70% without significant protein structure evolution. Below 30% sequence identity, however, 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 is the foundation of one of the most important methods in protein biophysics: structure homology modeling (10,11). Despite the fact that the plateau of high structural similarity above 30% sequence identity has been crucial for homology modeling and that many of the advanced structure prediction methods have been motivated by abrupt onset of the twilight zone, the cusped relationship between sequence and structural similarity has not yet received a detailed biophysical justification (12,13).

Previous work characterized the relationship between sequence and structure similarity by fitting the data empirically with an exponential function, and the adequacy of this model was interpreted as evidence in favor of the local model of protein structure determination, namely, that only a key subset of residues encodes a protein's structure (5,6,8). We are not aware of any experimental evidence favoring the local model such as, for example, showing that mutating a special subset amounting to ~30% of a protein's residues typically causes a protein's structure to evolve to a new structure. Conversely, it is obvious that randomly mutating 70% of a protein's residues will almost surely unfold it, as even a small number of point mutations can destroy a protein's structure (14). Therefore, the twilight zone in and of itself is not strong evidence for a local model of protein structure determination, and it is clear that without evolutionary selection, the range of 100–30% sequence identity could not correspond to nearly identical structures.

Submitted October 31, 2016, and accepted for publication February 22, 2017.

*Correspondence: shakhnovich@chemistry.harvard.edu

Editor: Amedeo Caflisch.
http://dx.doi.org/10.1016/j.bpj.2017.02.029

Purely physical models of structure evolution, without any selection, have explained many fundamental features of the protein universe. Dokholyan et al. (15) 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 evolution via duplication and structural divergence of proteins (16). Similarly, the birth, death, and innovation models developed by Koonin (17) explain the power-law-like distribution of gene family sizes that exists in many genomes. 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 (18–20) but also that structure emergence may often pass through thermodynamically destabilized evolutionary 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 (18,21). Similarly, the structure of a protein naturally adopting a 3α fold was converted into a 4β+α fold via a series of carefully engineered point mutations. Finding a rare sequence of mutations that avoided unfolded evolutionary intermediates was a major achievement of this work (19,22,23). Computational investigations of protein structure evolution using model proteins of Protein Data Bank (PDB) structures also show that structure evolution traverses unstable intermediates and found that less stable proteins are more evolvable at the structural level (24,25).

Here, we explore the evolutionary dynamics of protein structure discovery. Given the fact that most globular proteins must adopt a well-defined three-dimensional structure to function, and the strong evidence that many fold evolution pathways require protein destabilization, we hypothesize that protein structure discovery requires crossing valleys of low fitness on fitness landscapes. The valleys correspond to genetically encoded, evolutionarily transient, unstable proteins. Via this method of protein structure evolution, the strength of selection for folding stability under which a protein evolves would modulate its capacity to evolve a new structure.

We study evolution of protein structures using the data on sequence-structure relationships in natural proteins and a variety of evolutionary models of increasing complexity and realism. We find that the cusped relationship between structure and sequence divergence is a direct consequence of the interplay between evolutionary dynamics and the biophysical constraint for protein folding stability. In both the bioinformatics data and simulation data, the sharpness of the cusp, but not its position (~30% sequence identity for natural proteins), is determined by the compactness of evolving proteins, a proxy for their thermodynamic stability. Rather than fitting these data empirically, we formalized the mechanism underlying structure-sequence divergence into an ab initio analytical model that fits the bioinformatics data for proteins grouped by their compactness. Simulations show that what underlies the negative correlation between protein stability and structure evolution rate is the strength of evolutionary selection for stability under which proteins evolve. Fitness barriers are imposed by selection for thermodynamic stability, and continual sequence evolution degrades sequence identity over the timescales needed to find mutations that overcome these barriers and which encode alternative, stable protein structures. Many long-standing observations in protein biophysics are reinterpreted and unified using this powerful, yet simple, interpretive framework.

MATERIALS AND METHODS

Structure divergence of proteins sharing a common SCOP fold

Following Chothia and Lesk (5), 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.

We use the set of all single domain α, β, $\alpha+\beta$, and α/β protein domains in SCOP including mutants (26). Protein domain compactness, quantified by the number of amino acid contacts normalized by the domain length, i.e. the contact density (CD), was calculated as a predictor of thermodynamic stability. It is crucial to use a biophysical proxy for thermodynamic stability because experimentally measured values are not available for the vast majority of SCOP domains. The connection between contact density and folding free energy has been established through several lines of evidence. Interresidue 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 (27). 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 (28,29). Accordingly, the results of this study hold whether contact density or protein length is used as the proxy for stability (see Fig. S1 in the Supporting Material). When calculating contact density, we consider two residues in contact if any of their nonhydrogen residues are within 4.5 Ångstroms of each other. Pairs of proteins with comparable lengths (within 10 residues) and both belonging to the 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.

Sequence and structural similarities were calculated for each pair of protein domains classified in the same SCOP fold, in the same contact density class (both having high, intermediate, or low contact density) and with comparable length (within 10 amino acids). Thus, in Fig. 1, *B–D*, and the corresponding Data S1, the data are subdivided by contact density alone, not by

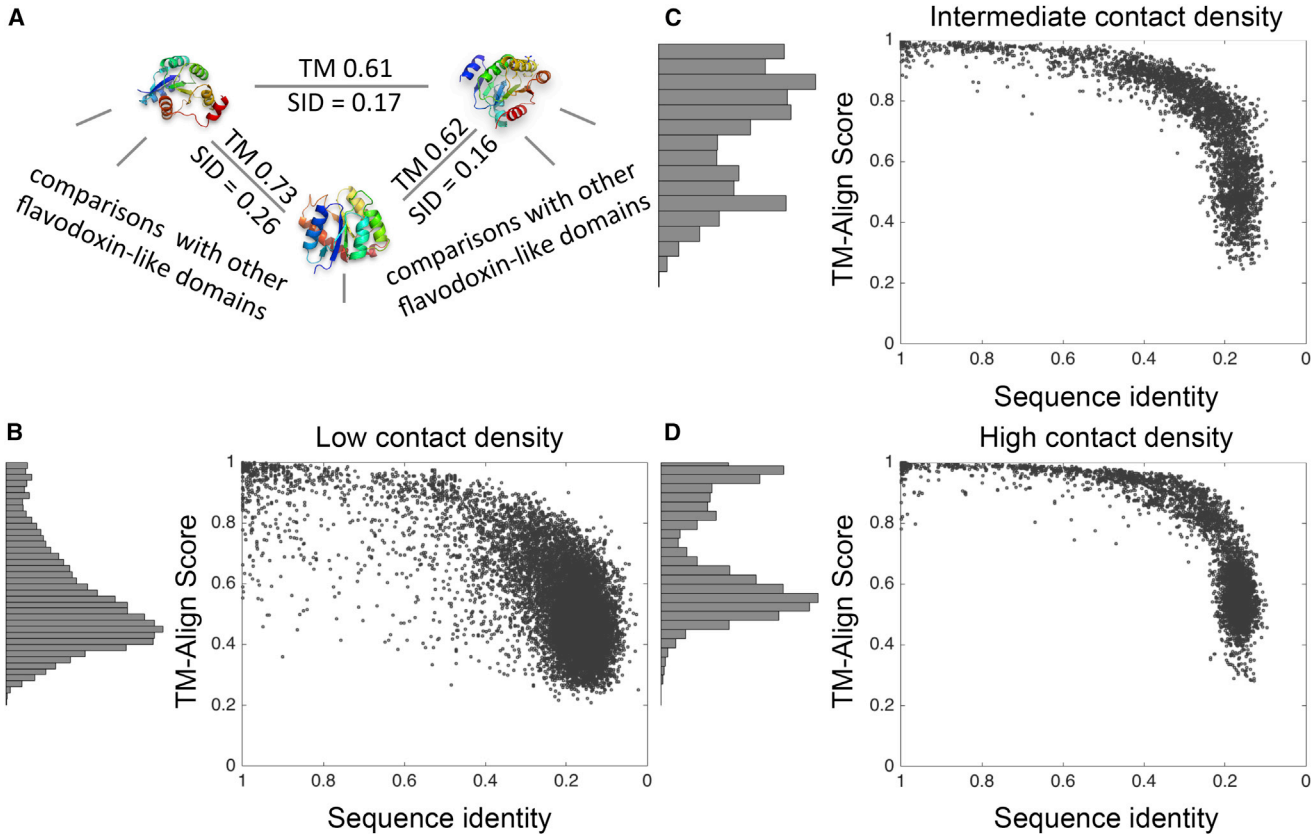

FIGURE 1 Stability and fold evolution of SCOP domains classified as α, β, $\alpha+\beta$, or α/β were used. Sequence identity values and structural similarity (TM-align score) for each pair of domains classified into the same fold and with comparable lengths (within 10 residues) were calculated. (*A*) An illustrative example from the $\alpha+\beta$ flavodoxin-like folds. The SCOP domains and their associated PDB IDs from left to right are d1a04a2 (PDB: 1A04), d1a2oa1 (PDB: 1A2O), and d1eudb1 (PDB: 1EUD). (*B–D*) The relationship between sequence divergence and structure evolution in SCOP domains. Note that sequence identity decreases from left to right. Domains are partitioned by contact density (CD). (*B*) Bottom 10% contact density (<4.13 contacts/residue, average protein length = 83 residues, N = 12,671 data points). (*C*) Middle 10% contact density (4.57 < CD < 4.65 contacts/residue, average protein length = 175 residues, N = 3863 data points). (*D*) Top 10% contact density (>4.93 contacts/residue, average protein length = 337 residues, N = 5672 data points). Histograms at the left of the plots show, in (*B*), that at low contact density, the distribution of TM-align scores is approximately single peaked while (*C*) and (*D*) show that at higher contact densities, the distribution of structure similarity TM-align scores is bimodal. To see this figure in color, go online.

protein class or fold. Proteins classified in different SCOP folds do not share significant sequence homology or structural similarity, and were therefore not compared.

Sequence identities were determined from either global sequence alignment or from within structurally aligned regions. For global alignments, sequences were aligned using the Needleman-Wunsch (NW) algorithm and Blosum30 substitution matrix as implemented in the software MATLAB (The MathWorks, Natick, MA), and sequence identity was defined as the number of identical residues matched in the alignment, normalized by the average length of the aligned proteins (30,31). Structures were aligned using the Template-Modeling (TM) algorithm. From these alignments, sequence identity was defined as the number of identical residues matched within the structurally aligned region, normalized by the length of the structurally aligned region. Structure similarity was quantified by TM-score, ranging from 0 to 1 (32,33). Some alignment scores, including root mean squared deviation, correlate negatively with protein length; however, the TM-score is independent of length (32,34). This is an important feature of TM-scores for our purpose because SCOP domains range from 10 residues to ~1000 residues. Furthermore, because the TM-score weights residues that are close together more highly than distant residue pairs, it is less sensitive to local structural variations than unweighted scores.

Protein model

We used standard 27-mer lattice-model proteins to simulate the structural evolution of proteins (35). Lattice proteins can fold into 103,346 fully compact structures for the 3 × 3 × 3 cubic lattice, and following Heo et al. (36), we use a representative subset of 10,000 randomly chosen structures for computational efficiency. All proteins in this model have 28 contacts and therefore have identical contact densities. 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 (37), 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

$$E_S = \frac{1}{2}\sum_{ij=1}^{27} \mathbf{C}_{ij}^{(S)} \mathbf{M}_{AA(i)AA(j)}, \qquad (1)$$

where $C^{(S)}$ is the contact matrix of structure S, whose elements $C_{ij}^{(S)} = 1$ when residues i and j form a noncovalent contact, and $C_{ij}^{(S)} = 0$ otherwise. $\mathbf{M}$ 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, $P_{\text{nat}}(T)$, of a protein adopting its native state (lowest energy) structure is therefore given by the canonical partition function:

$$P_{\text{nat}}(T) = \frac{e^{-E_{\text{nat}}/T}}{\sum_{i=1}^{10,000} e^{-E_i/T}}, \quad (2)$$

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) (33,38),

$$Q_{\text{a,b}} = \frac{1}{28}\sum_{ij=1}^{27} C_{ij}^{(\text{a})} C_{ij}^{(\text{b})}, \quad (3)$$

where $C^{(\text{a})}$ and $C^{(\text{b})}$ 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 Eq. 1.

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 (39–41). Each simulation of model protein evolution proceeded in two steps. First, each simulation was initialized with a protein stably adopting a particular structure ($P_{\text{nat}} > 0.99$). 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 (42). 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:

$$f = P_{\text{nat}}^{(\text{trial})} - P_{\text{nat}}^{(\text{original})}, \quad (4)$$

where $P_{\text{nat}}^{(\text{trial})}$ is the stability of the protein with the mutation and $P_{\text{nat}}^{(\text{original})}$ is the stability of the protein before the mutation. Any neutral mutation or mutation increasing fitness was accepted while destabilizing mutations were accepted or rejected according to the Metropolis criterion, with $e^{f/\text{T}_{\text{sel}}}$. The selective temperature controls the stringency of selection for protein stability; the more destabilizing a mutation, the less likely it is to be accepted (42). T_{sel}, the selection temperature, directly tunes the equilibrium P_{nat} of evolving proteins (Fig. S1; Table S1). Simulations ran for 1000 mutation attempts, illustrated in Fig. 3 *A* and the structure, sequence, and stability (P_{nat}) values were recorded every 10 generations. This was repeated for ~2750 different starting structures at different strengths of selection for stability P_{nat} (set by T_{sel}).

Monte Carlo algorithm for convergent evolution

We also simulated an alternative, convergent, mechanism of structure discovery. These simulations did not instantiate evolutionary pressures at play in nature but allowed us to test whether a purely biophysical constraint might explain the cusped sequence-structure relationship. These simulations start with many, independently generated proteins that each adopt a different structure with high P_nat. By random selection, one of these proteins is designated as the target sequence, and all the other proteins evolve under selection pressure to maximize their sequence identities with respect to this target sequence. To simulate the regime where selection pressure for folding stability is low (Fig. 4 *A*), any mutation increasing sequence identity with respect to the target sequence identity ($SID_{i,\text{target}}$) 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 $SID_{i,\text{target}} \times P_{\text{nat}}$ (Fig. 4 *C*). A single convergent evolution simulation was run in each selection regime, consisting of 100 proteins converging over 3000 generations in the weak selection regime, while 200 protein converging over 3000 generation were used in the strong selection regime to compensate for slower evolution and to improve statistics. To determine the sequence-structure relationship for convergent evolution, the sequence identity and Q-score values among the converging proteins were calculated every 100 generations within each regime.

Multiscale modeling of protein evolution

We used the multiscale simulation algorithm developed and described in detail in Zeldovich et al. (43). Because this evolutionary procedure yields protein 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 $P_{\text{nat}}^{\text{rand}} = 0.23$ 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 basepair; 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 $d = d_0\left(1 - \min_i P_{\text{nat}}^{(i)}\right)$; or 5) division of the organism into two equivalent daughter cells 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 $N - N_{\text{max}}$ randomly chosen organisms from the population each generation. We modulated the strength of selection in the population by drawing on the longstanding theoretical population genetics finding that the strength of selection increases with population size, and 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 3000 generations or when the population went extinct. Only simulations that ran the full 3000 generations were analyzed, and even among these evolutionary runs, in some cases, the proteins did not evolve high $\overline{P}_{\text{nat}}$(Fig. 5, *B* and *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.

RESULTS

Structure-sequence relationship in the protein universe

A cusplike relationship between structure and sequence divergence has long been observed in the literature (5–7,12) 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 (see Materials and Methods; (44)).

We define thermodynamic stability of a protein as the fraction of molecules in the ensemble that reside 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 (45–50). 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:

$$P_{\text{nat}} = \frac{e^{-\frac{\Delta G}{k_B T}}}{1 + e^{-\frac{\Delta G}{k_B T}}}, \tag{5}$$

where T is temperature and k_B is the Boltzmann constant. Here, we consider only single domains as defined in SCOP and use the contact density as a proxy for the folding free energy ΔG of a protein, as described in Materials and Methods (24). 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. 5.

In Fig. 1, *B–D* (Data 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. 1, *B–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. 1 *D*). However, in the class of low contact density proteins, there is only a peak at low TM-align score. Furthermore, for high and intermediate contact density domains, 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. These results are robust to the exact sequence similarity metric used as shown in Figs. S2 and S3. In Fig. S2, sequence similarity is measured by the NW alignment score using the BLOSOM30 substitution matrix. Because this score takes similarity between aligned residues into account, e.g., a matched alanine and glycine would lead to a higher sequence similarity score than alanine and arginine, and because it penalizes alignment gaps, the NW alignment score attains greater sensitivity than sequence identity. When interpreting the position of the twilight zone, it is important to keep in mind that the NW alignment algorithm produces sequence identities of ~15% between random sequences in the process of optimizing the alignment. Therefore, the exact position of the twilight zone likely reflects both biophysics and artifacts of the alignment program. Our goal was not to explain the absolute value of the twilight zone's position, which is not markedly different between contact density subgroups. However, to control for the possibility that alignment artifacts account for the different behaviors in the onset of the twilight zone that we sought to explain, we examined the sequence identities that arise only within regions that were structurally aligned by the TM-align program. This approach is similar to that used in Chothia and Lesk (5), where sequence identities were determined only for a common structural core shared by the proteins compared. As shown in Fig. S3, this method shifts the twilight zone to a lower sequence identity, as expected, but the shape of the cusp is unchanged. Thus, this analysis robustly 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 contact densities might be enriched in different structural classes so that our findings presented in Fig. 1 reflect differences in evolutionary scenarios for different protein structural classes. However, we can eliminate this possibility because we found that there are no significant differences in the shape of the sequence-structure relationship when the data are subdivided by protein class (Fig. S4). Finally, we note that the results described above hold when protein length is used as a proxy for thermodynamic stability as well as when contact density is used (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 (22,50–53). 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 is not only nonfunctional but also can form toxic aggregates (39). 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. 2 *A*). The model is based on three postulates: First, new structures are discovered in divergent evolution from existing structures. Second, in analogy with chemical

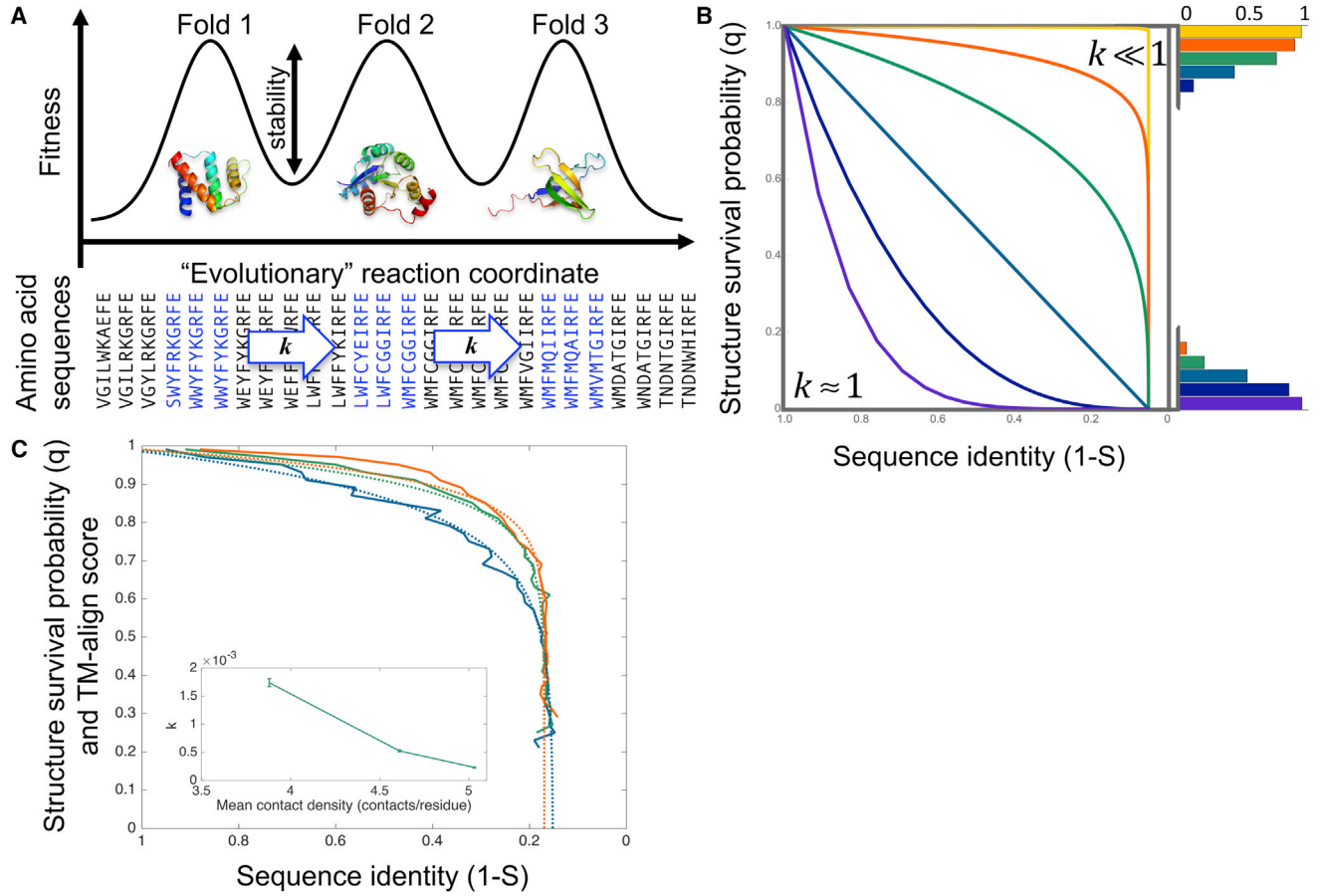

FIGURE 2 Fitness barriers to fold evolution leads to a cusped sequence-structure relationship. (*A*) Schematic representation of the evolutionary process in which fold discovery events occur by overcoming fitness valleys along the evolutionary reaction coordinate. (*Blue*) Sequences that are more fit, i.e., that contribute to high fitness in the organisms they are part of. Families of homologous proteins are formed by sequence evolution between structure evolution events. The rate of structure discovery, k, depends on the thermodynamics stability (a property of protein biophysics) of an evolving protein because in our model's analogy to chemical kinetics, the more stable an evolving protein, the higher the fitness barrier to evolving a new fold. (*B*) The relationship between sequence identity and structural survival probability of an ancestral structure for a protein with length 27 (the length of model proteins used below) at several values of k: from top to bottom, 0.0001 (*yellow*), 0.003 (*orange*), 0.01 (*green*), 0.04 (*light blue*), 0.1 (*dark blue*), and 0.25 (*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. (*Dashed lines*) Model fit to (*Solid lines*) binned bioinformatics data for each contact density subgroup: low contact density (*blue*), intermediate contact density (*green*), and high contact density (*orange*). (*Inset*) Correlation between protein domain contact density and protein structure evolutionary rate, k, predicted analytically. Error bars indicate 1 SD. To see this figure in color, go online.

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 comprising sequences that encode unstable proteins separating two stable protein structures that confer high fitness on their carrier organisms. In this model, organisms encoding thermodynamically stable proteins are fit because we assume proteome stability is a main component of organismal fitness throughout this article (46). Third, we postulate that crossing the fitness valley leads to the discovery of novel folds that are structurally dissimilar to the original fold.

We denote k as the rate of structure evolution. Modeling structure evolution as an activated process, the probability, $q(t)$, that an ancestral structure (the structure at time $t = 0$) is unchanged at time t follows immediately:

$$q(t) = e^{-kt}. \tag{6}$$

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 $t = 0$, normalized by protein length, is given by

$$S(t) = \frac{l(a-1)}{a}\left(1-\left(1-\frac{a}{l(a-1)}\right)^{t}\right), \quad (7)$$

where l is the length of the protein and a is the number of amino acid types, typically 20 (see the Supporting Material 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 structure.

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

$$q(S) = \exp\left(-k\frac{\log\left(1-S\frac{a}{l(a-1)}\right)}{\log\left(1-\frac{a}{l(a-1)}\right)}\right). \quad (8)$$

Fig. 2 *B* shows the dependence of structure survival probability as a function of sequence identity ($q(S)$ versus $1 - S$). When $k \approx 1$, the protein is free to diffuse through structure space as easily as it does through sequence space. As proteins become more stable, the barriers between protein structures begin to retard structure evolution, giving rise to activated dynamics. When $k \ll 1$, 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 minimum free-energy native state. The histograms in Fig. 2 *B* also indicate that differences in stability may explain heterogeneity that has been observed in gene family sizes, defined as the number of nonredundant sequences that adopt highly homologous structures (54). For each of the structure evolution rates 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 slow structure evolution rates ($k \ll 1$) are consistent with the high-contact density class of proteins while those with intermediate values of k are consistent with the low-contact density class of proteins. Unstable proteins move rapidly through structure space, providing little time for gene duplication and sequence divergence to populate a particular family when compared to their thermodynamically stable counterparts.

Finally, we confirm that the analytical model predicts a decreasing rate of structure evolution, k, when fit to the three protein domain subgroups with increasing contact density. For each contact density subgroup, we proceeded by binning the data into 50 bins spanning sequence identity of 0 to 1. 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 occurs between two random sequences, while the bulk of proteins compared in the twilight zone share 20% sequence identity for real proteins. This possibly reflects that sequence alignment algorithms seek to maximize the overlap of sequence pairs being aligned. We added a parameter C to the equation to shift the analytical model twilight zone to the twilight zone of the data. The exact curve that was fitted to the data was

$$\tilde{S}(q) = 1-\frac{1-a}{a}\left(\left(1-\frac{a}{l(a-1)}\right)^{\frac{1-q}{k}}-1\right)+c, \quad (9)$$

where $\tilde{S}(q)$ is the sequence identity as a function of structure survival probability, a rearrangement of Eq. 7 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. Equation 8 was fit to the binned bioinformatics data using the program Igor Pro (WaveMetrics, Lake Oswego, OR), which optimized the parameters k and C for fit.

As shown in Fig. 2 *C*, 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 (Table S1, *inset*). Contact density also correlates negatively with k when the NW alignment score is used to measure sequence similarity (Fig. S5; Table S2). 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 small differences in average contact density. Protein domains in the four structural classes, α, β, α/β, and $\alpha+\beta$, 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. S6; Table S1).

Structure evolution of model proteins

We now turn to explicit modeling of protein structure evolution to test the assumptions of our analytical model and to

get mechanistic insights into the biophysics of fold emergence. Details of the model are provided in the Materials and Methods. In brief, each model protein consists of 27 amino acid residues that fold into a compact 3 × 3 × 3 cube (35,55,56). All 103,346 possible compact structures of such model proteins have been enumerated, and, following Heo et al. (36), we use a subset of randomly selected 10,000 conformations as our space of possible protein structures for computational efficiency throughout this work. Neighboring amino acid residues that are not connected by a covalent bond interact according to a MJ potential (37). In line with previous discussion, the fitness of each model protein is represented by its stability, 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 Materials and Methods).

Hypothesizing that the strength of evolutionary selection under which a protein evolves is the origin of both the protein's stability and its structure evolution rate, we ran many evolutionary simulations where proteins could evolve new structures under stability (P_{nat}) constraints of various stringencies. Each simulation started with a stable protein ($P_{\mathrm{nat}} > 0.99$) and 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 (55,57) (see Materials and Methods, Fig. S7; Data S2). Simulations ran for 1000 generations (mutation attempts, illustrated in Fig. 3 *A*) and 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 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 (i.e., Q-score; see Materials and Methods) (38).

In Fig. 3 *B*, 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, $\overline{Q}$, between a random pair of model proteins is 0.19 ± 0.08, as indicated by the red line in Fig. 3 *B*. For each tested selection pressure, an average Q-score associated with fold evolution events, $\overline{Q}_i$, can be determined from the simulation data by calculating the Q-score between the ancestral and mutant fold each time a new fold arises, and averaging over these values. As depicted in the bottom inset of Fig. 3 *C*, $\overline{Q}_i$ does not differ significantly from $\overline{Q}$ when proteins evolve at the lowest $\overline{P}_{\mathrm{nat}}$ observed ($\overline{P}_{\mathrm{nat}} = 0.23$). The case is very different, however, when proteins evolve at their most stable $\overline{P}_{\mathrm{nat}}$ observed ($\overline{P}_{\mathrm{nat}} = 0.97$). Stable proteins are biased toward discovering new structures similar to ancestral structures such that $\overline{Q}_i = 0.42 \pm 0.17$, almost 3 SD above $\overline{Q}$. The positive correlation between $\overline{Q}_i$ and $\overline{P}_{\mathrm{nat}}$ likely reflects that discovering a mutant fold similar to the wild-type (high Q) avoids destabilizing evolutionary intermediates (Fig. S8). However, even strong selection pressure does not increase $\overline{Q}_i$ above ~0.42 because there exist so few structures with $Q > 0.42$ for any given evolving protein structure (Fig. S9) that their discovery appears unlikely for entropic reasons. 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 $P(Q_{t,t+10})$, the probability distribution of Q-scores between protein structures at time t and at time $t + 10$, which is bimodal with peaks at $Q = 1$ and $Q \approx 0.29$ (see below, Fig. 4, *A–C*), consistent with the bimodal distribution of TM-align scores discussed above as well as with abrupt conformational changes observed in protein switches and evolutionary simulations of using more realistic protein models (58,59).

A key assumption of the analytical model was that fold evolution is an activated process (60). 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. 3 *C*).

Alleviating purifying selection for stability accelerates the rate of structure evolution by allowing proteins to maintain lower stability on average (lower $\overline{P}_{\mathrm{nat}}$). The average wait time between structure discovery events diminishes rapidly as simulations are run at higher T_{sel} values such that the stability of the evolving proteins diminishes. When $\overline{P}_{\mathrm{nat}} \leq 0.74$, selection pressure was so weak that nearly every recorded generation explored a different fold (Fig. 3, *B* and *C*, *top inset*), akin to the $k \approx 1$ regime in the analytical model. The abrupt transition from diffusive to activated dynamics is also apparent in the distribution of fold discovery events (Fig. S10 *A*), which shows that the distribution is Poissonian when $\overline{P}_{\mathrm{nat}} \leq 0.74$, where structure evolution events are rare, and that the mean number of fold discovery events increases when $\overline{P}_{\mathrm{nat}}$ falls below 0.74 (Fig. S10 *A*; Data 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 support 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 -up sequence-structure relationship (Fig. 4 *A*) that is consistent with the $k \approx 1$ regime of the analytical model and not

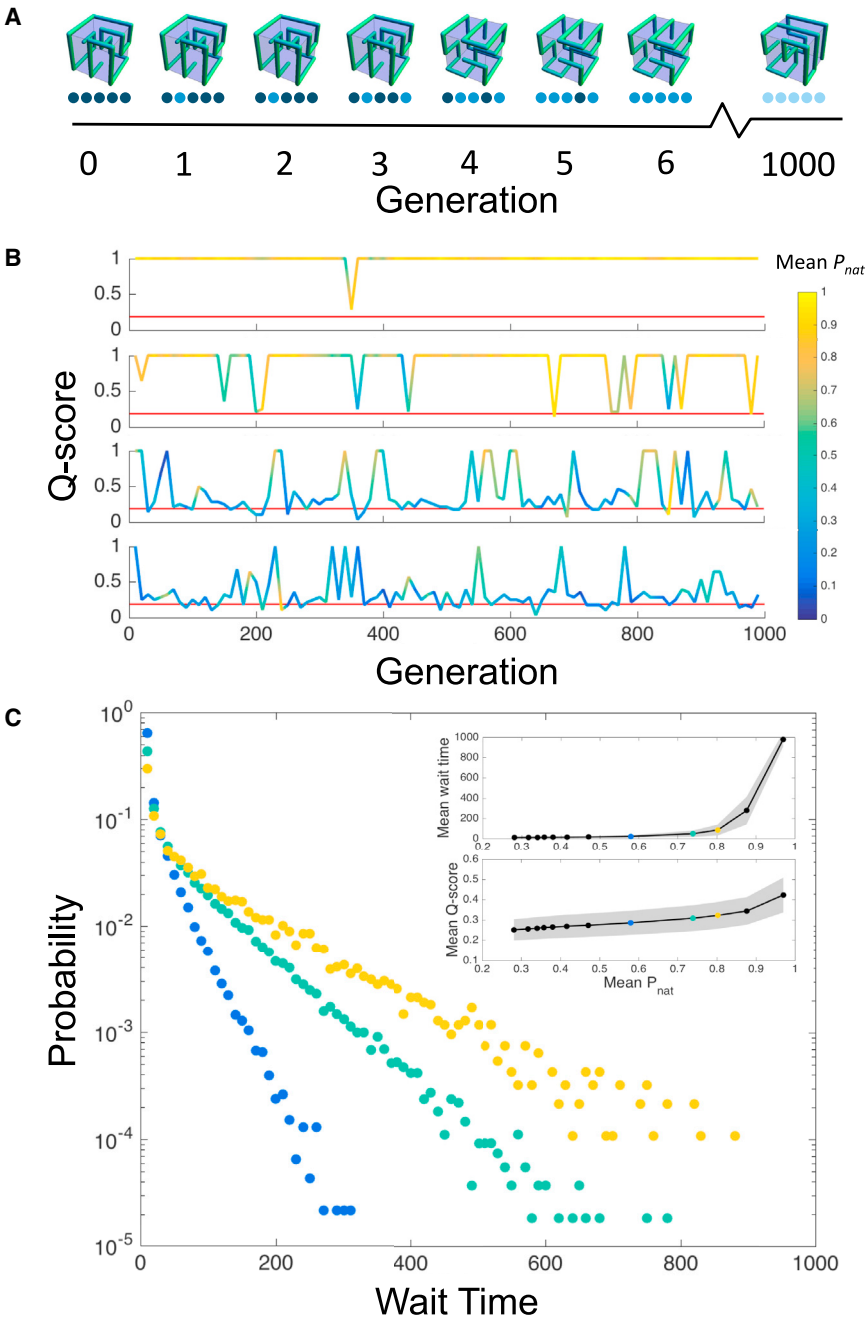

FIGURE 3 Dynamics of fold evolution. (*A*) Schematic representation of Monte Carlo-simulated evolution. Protein sequences (represented by a series of *circles* below each lattice structure; color indicates amino acid type) acquire mutations, periodically causing the protein to transition to a new structure. (*B*) Four representative fold evolution trajectories at different mean thermodynamic stabilities ($\overline{P}_{\mathrm{nat}}$). 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 $\overline{P}_{\mathrm{nat}}$, $\overline{P}_{\mathrm{nat}} = 0.58$ (*blue*), $\overline{P}_{\mathrm{nat}} = 0.74$ (*green*), and $\overline{P}_{\mathrm{nat}} = 0.80$ (*yellow*). In the insets and in (*C*), the colored data points also correspond to these $\overline{P}_{\mathrm{nat}}$ values. (*Top inset*) Average wait time between fold discovery events for the different values of $\overline{P}_{\mathrm{nat}}$ tested. (*Bottom inset*) Average structural similarity (Q-score) between the previous structure and the arising structure for fold discovery events. To see this figure in color, go online.

observed in real proteins. The most interesting cases are at intermediate selection regimes where there is a sigmoidal cusplike transition from high average structural similarity to low structural similarity at higher sequence divergence (Fig. 4, *B* and *C*). In this selection regime, proteins dwell in a particular structure while accumulating sequence mutations, yet are periodically, at longer timescales, 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. 4 *D*). 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 $\overline{P}_{\mathrm{nat}}$ and the mean structure family size (Fig. 4 *E*). Structure family size is equivalent to gene family size, as discussed in Shakhnovich et al. (54), except that amino acid sequences rather than nucleotide sequences are considered here.

The analytical model and simulations indicate that a divergent scenario reproduces the key features of the sequence-structure relationships shown in Fig. 1. In this model, the plateau at high structural similarity arises from

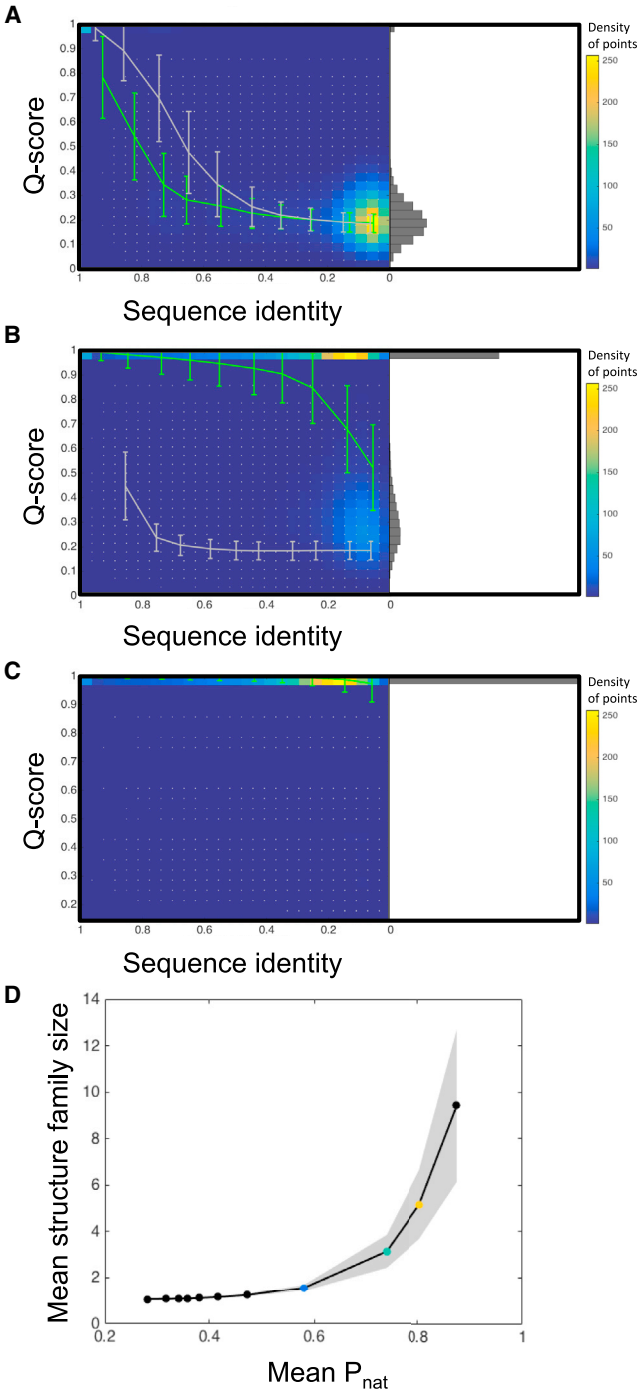

FIGURE 4 The relationship between sequence divergence and structure evolution for model proteins. (*A–C*) The sequence identity and Q-score for each pair of proteins compared. (*White points*) Data points from proteins arising in the same trajectory. The color scale indicates the density of points for each pair of Q-score, SID 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 *gray*) for convergent evolution simulations (see Materials and 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. (*Gray*) Result of convergent evolution simulations without selection for stability. There is no significant evolutionary hysteresis when selection for folding stability is low, as can be seen in the similarity of the green and gray 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 (*gray*) 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 extremely 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 timescale 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 nonredundant sequences (SID < 0.25) evolved in simulations that adopt a particular structure. (*Shaded regions*) 1 SD; (*colored points*) $\overline{P}_{nat} = 0.58$ (*blue*), $\overline{P}_{nat} = 0.74$ (*green*), and $\overline{P}_{nat} = 0.80$ (*yellow*). To see this figure in color, go online.

evolutionary fitness barriers impeding structure evolution, not from the biophysical fact that sequences encoding stable proteins and sharing >30% of their residues must encode the same structure. So far, however, we have not been able to fully exclude the latter possibility. To test whether a purely biophysical constraint might explain the cusped sequence-structure relationship, we simulate an alternative, convergent, mechanism of structure discovery. To simulate convergent evolution, proteins start from sequences that stably fold into randomly chosen structures, but the fitness function favors sequence convergence to another, target protein (see Materials and Methods for details). This scheme was constructed to test the pure biophysics hypothesis of the cusp's origin, described above, and obviously does not reflect actual convergent evolutionary forces found in nature. In Fig. 4, *A* and *C*, we accompany the divergent evolution results described above with the results of convergent evolution simulations, shown in gray. 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. 4 *A*). By contrast, under strong selection for stability, protein structures begin to converge at a much higher sequence identity than where they begin to diverge (~85 and 25%, respectively; Fig. 4 *C*). Proteins converging under the constraint of stability also attain only 69 ± 9% sequence identity with respect to the target sequence after 3000 mutation attempts and only 0.5% of the evolving proteins attained the target structure, compared to 100% in the weak selection regime. Therefore, evolutionary dynamics must play a role in generating the sequence-structure cusp and a pure biophysics explanation is rejected.

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 Alexander et al. (19), in which two proteins with different sequences and structures were subjected to engineered single point mutations that increased their sequence identity up to 88%, yet did not unfold the proteins or change their original structures.

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 Fig. 5 *A* and described in detail in Materials and Methods and Zeldovich et al. (43). 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

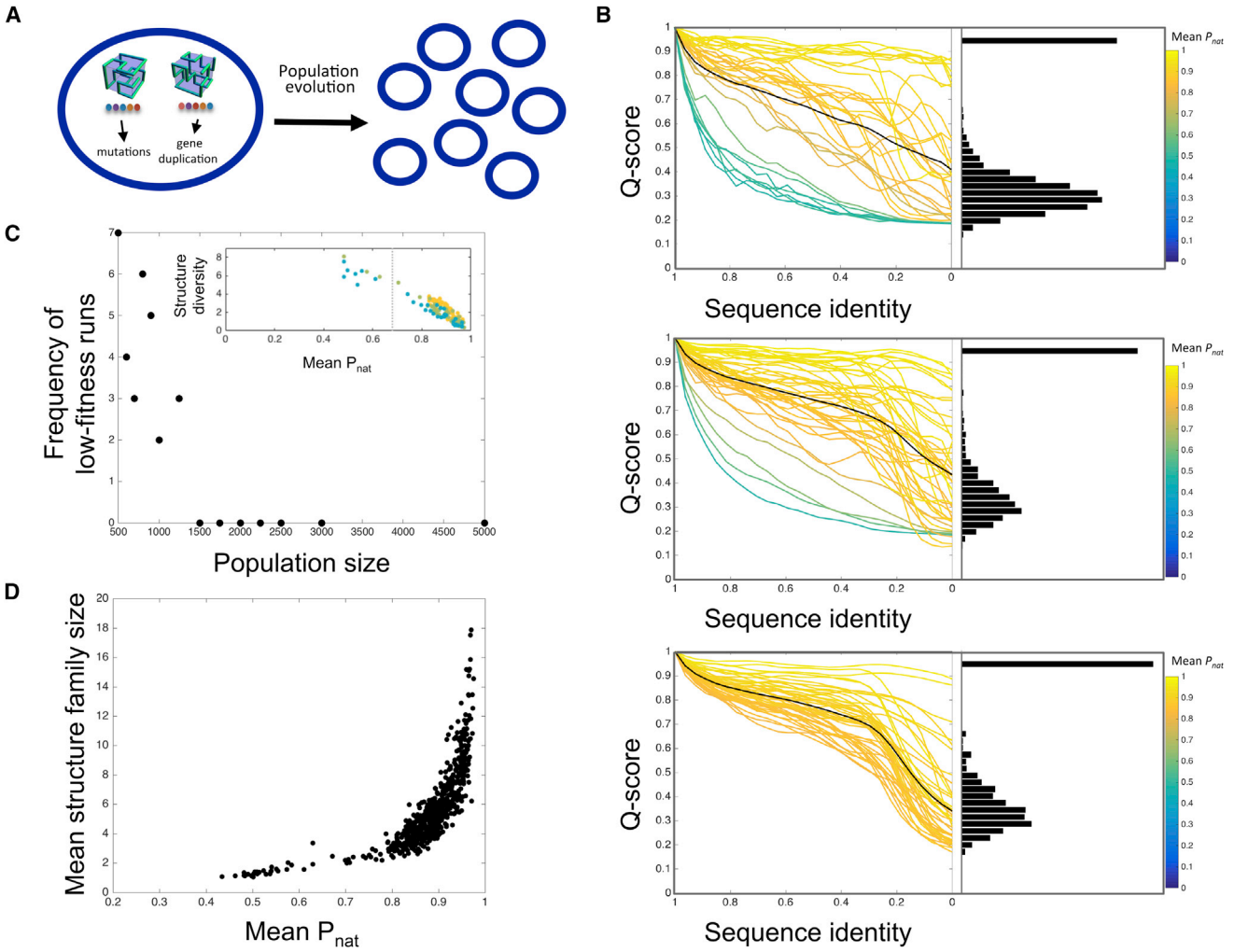

FIGURE 5 Population size modulates selection pressure. (*A*) Schematic representation of the multiscale evolution model. Each organism consists of genes that code for proteins and can be duplicated or acquire point mutations. Organisms evolve under selection pressure for protein folding stability. At the end of the simulation, SID and Q-score are calculated for each pair of extant proteins in the population. (*B*) The number of replicas that evolved the full 3000 generations yet failed to evolve stable proteins ($\overline{P}_{nat} < 0.68$) for each population size. Notably, at population sizes >1250, 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 = 1250; (*yellow*) population size = 5000. (*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. (*Black line*) Average over the individual trajectories. (*Top panel*) Population size = 500; (*middle panel*) population size = 1250; (*bottom panel*) population size = 5000. 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 cusplike dependence of structural similarity (Q-score) on SID only emerges at strong selection pressure. (*D*) Mean structure family size as a function of $\overline{P}_{nat}$ of proteins at the end of each trajectory. To see this figure in color, go online.

superfamily sizes observed in nature (43). 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 nontrivial. The strength of selection (proxied by T_{sel} in the previous model of individual protein evolution) is determined by the population size (61,62) in this more biologically realistic model.

Evolutionary runs were simulated over a range of maximum population sizes, N_{max}, from 500 (weak selection) to 5000 (strong selection) 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 (43).

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 turnover of structures is manifested in a concave-up dependence of average Q-score on sequence identity (*green and blue trajectories* in Fig. 5 *B*), as in the low P_{nat} regime of the Monte Carlo simulations (Fig. 4 *A*). Once the population size passes the critical threshold, of ~1250 organisms, the characteristically cusped sequence-structure relationships become the most probable result of evolution (Fig. 5 *B*, *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} value of proteins at the end of 3000 generations ($\overline{P}_{nat}$) strongly anticorrelates with the diversity of structures observed in the population, as reflected in the structure entropy:

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

where p_i is the probability of finding structure i in the final generation of the simulations (Fig. 5 *C*, *inset*). The correlation between $\overline{P}_{nat}$ 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. 5 *B*, *inset*). Rather, we found that population size modulated the number of evolutionary replicas that failed to evolve stable proteins (Fig. 5 *C*). This is also reflected in the relative peak heights of the Q-scores shown in Fig. 5 *B*.

Finally, we examine the size of the average protein structure family across the $\overline{P}_{nat}$ range explored in the simulations (Fig. 5 *D*). The size of the protein structure family is defined as the number of genes encoding nonredundant protein sequences (SID $<$ 25%) but whose gene products adopt the same native state structure. In datasets of natural proteins, protein structure families of various sizes have been observed and overall, there is a positive correlation between gene family size and protein contact density (54). 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 $\overline{P}_{nat}$ of proteins at the end of an evolutionary replica and the average structure family size (Fig. 5 *D*). 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 $\overline{P}_{nat}$ (Fig. S11). 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.

DISCUSSION

We presented a simple physical analytical model of protein structure evolution that explains why there is a cusped relationship between structure and sequence divergence. Under the constraint of protein folding stability, fitness valleys form barriers that separate sequences encoding stable protein structures. The most stable proteins face formidable fitness valley barriers and therefore, slow structure evolution rates. Continuous sequence evolution degrades sequence identity of diverging proteins over the timescale needed to accumulate the mutations that traverse these valleys. Our simulation results show that protein stabilities and their accompanying rates of structure evolution, k, arise as a result of differential selection pressures for stability. Strong selection for stability causes proteins to evolve high stability and hinders structure evolution, while weak selection permits rapid exploration of structure space, of predominantly unstable proteins. Interestingly, this observation strengthens the analogy to chemical kinetics: just as the ratio of the free energy barrier to temperature controls the rate of barrier crossing in chemical kinetics, the ratio of the fitness barrier to the appropriate measure of strength of evolutionary selection controls the rate of structure discovery in the evolutionary context of the model (61).

Importantly, our bioinformatics analysis of protein domains in SCOP shows that the effect of selection strength, as reflected in proteins' contact densities, is not just theoretical but actually modulates the relationship between structure and sequence divergence, which was previously thought to be universal. By averaging our analysis over hundreds of proteins, features the proteins have in common are amplified, and we are able to extract statistical rules of protein evolution. The analytical model we propose not only

explains the existence of the cusp, it also recapitulates and explains why the transition to the twilight zone becomes sharper as selection pressure increases: high selection decreases the probability that a new fold will emerge before decay of sequence similarity saturates.

Our bioinformatics results were the same when we tested a measure of sequence similarity that accounts for similarity of physical properties between substituted amino acids. While no other study has examined the effect of contact density on the sequence-structure relationship, Wilson et al. (6) rigorously tested multiple methods of scoring sequence similarity (including percent sequence identity, Smith-Waterman Score, and statistical significance of sequence similarity) and confirmed that the nonlinear dependence of structure divergence on sequence divergence is independent of the methods used. Interestingly, several studies that focused on the structure evolution that occurs only within protein families, thereby excluding the distantly related proteins included in Wilson et al. (6) and in our study, reported that the nonlinearity does not arise consistently when percent sequence identity was substituted with other, more advanced, measures of sequence similarity (8,9,63). This apparent contradiction likely arises from the close evolutionary relationship among proteins in a protein family, which share a common ancestry as inferred by significant sequence similarity and very similar structure and function. Sequence identity saturates more rapidly than other measures of sequence similarity so when only closely related proteins are examined, sequence identity begins to saturate, causing apparent nonlinearity, while others such as NW-score or bitscore may or may not begin to saturate depending on the family (9). Therefore, using a measure of sequence similarity such as statistical significance of the similarity generally yields a linear relationship within protein families (8). However, when all proteins sharing a particular SCOP fold are examined (6), the relationship between sequence and structure divergence remains nonlinear, with the cusp present irrespective of the particular definition of sequence and structure divergence (z-scores and p values were used as measures of statistical significance in Wood and Pearson (8) and Wilson et al. (6), respectively). We also note that this finding is consistent with our divergent evolution model, where nonlinearity emerges from rare evolutionary transitions between significantly different protein structures (Fig. 4). Thus, by limiting the analysis to protein families, Wood and Pearson (8) and Wilson et al. (6) excluded significantly diverged structures and, not surprisingly, observed linear relationships between structure and sequence divergence for most families. By contrast, we were able to analyze and explain the nonlinearity in the sequence-structure relationship by expanding the analysis to the level of protein fold.

An underlying motivation of these previous works was to understand how a protein's structure is encoded in its residues, namely, whether structural information is encoded globally across all residues or if it is localized to a subset of gatekeeper residues (64–67). Wood and Pearson (8,9) argued that a nonlinear relationship between sequence and structure divergence was only consistent with the latter mechanism. Indeed, there is a very clear mechanism by which the gatekeeper model would generate a cusped relationship between sequence and structure divergence: if, for example, only 30% of residues determine protein structure, then 70% can evolve without disrupting protein structure, and it is not until these gatekeeper residues accumulate mutations that a new structure emerges. Our analytical model, however, is a global model of residues determining protein structure because we do not define any privileged gatekeeper residues. Thus, we clearly demonstrate that a global model based on fitness barrier crossing is also consistent with the cusp and twilight zone.

To test whether a local model in which a few gatekeeper residues determine protein structure might also consistent with the bioinformatics and simulation data, we constructed a second analytical model based on this mechanism (see the Supporting Material for derivation). This model predicts that the position of the twilight zone depends on the number of gatekeeper residues (Fig. S12). While this result may be intuitive, it is not consistent with the bioinformatics or simulation data, even though the 30% position of the twilight zone is not imposed on these in any way. That we do not observe a moving twilight zone in the data indicates either that 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. Contrary to previous work, therefore, we have shown that the global model is not only consistent with the bioinformatics data but is actually better at explaining the bioinformatics data than the previously favored local model.

Despite its simplicity, the proposed structure evolution mechanism is powerful in placing many longstanding observations in protein biophysical evolution in a single interpretive framework. Along with tuning the rate of structure evolution, another effect of selection is to modulate the size of sequence 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 sequence 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 (defined as the number of nonredundant sequences adopting similar, but different structures, analogous to a SCOP fold) (54,68). The root of these observations 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 (54,68). Our work provides an alternative framework within which to interpret these findings. We suggest that it is pressure for folding stability that causes proteins to evolve high contact density, severely constrains structure evolution; leads to larger gene families due to longer divergence times between structure innovation events; and selects for discovery of structures similar to the ancestral structure, to the extent that it is possible, which leads to larger structural neighborhoods.

It may seem counterintuitive that stability retards the rate of structure evolution because in the context of directed evolution, stability typically enhances evolvability (69). Stability promotes evolvability during directed evolution because engineered stabilizing mutations create a stability buffer that allows the protein to tolerate destabilizing mutations that confer a new function without changing the structure. However, in the context of natural evolution, a protein's stability reflects mutation-selection balance—the point at which selection for protein folding stability is balanced by mutational pressure toward less stable sequences (47). Therefore, proteins that are naturally very stable are such because they are under stronger pressure for stability (e.g., they may be more abundant in the cytoplasm) and for that reason they might not have a reservoir of stability that can be used up and regenerated during subsequent rounds of structure evolution (70). Supporting this point, we observed that structure evolution events were associated with loss of stability for proteins of marginal stability, but not for the most stable proteins (Fig. S13).

Here, we reported a negative correlation between contact density and structure evolution rate. Curiously, when focusing on contact density and sequence evolutionary rate, Zhou et al. (71) reported a positive correlation. The influence of selection on contact density and evolutionary rates therefore apparently depends on the type of evolutionary rate examined (structure versus sequence) and on the timescales (long versus short). Overall, we view contact density and its associated metrics as neither a sign of intrinsic evolvability nor evolution under weak selection, as most previous studies have, but rather as a signature of strong evolutionary selection (54,68,71). These interpretations may not be diametrically opposed but further studies may be needed to clarify their relationship.

We focused on the pace of protein domain structure evolution under selection for folding stability by the mechanism of point mutations. This particular instantiation builds intuition that likely holds more broadly. Like structure evolution by point mutations, structure discovery by insertions, deletions, and gene rearrangements initially generates functionally inferior proteins that are concomitantly or subsequently compensated by point mutations (72) (Figs. 3 *B* and S13). The point mutations that stabilize arising structures likely occur in residue pairs that are making newly formed contacts, and these point mutations may begin to coevolve. In the context of the novel structure (73–75), newly formed contacts in the interresidue contacts residue pairs formed may be detectable by examining which protein residues coevolve.

The evolutionary dynamics we observe when protein domains are under direct selection for stability may arise under different sources of selection as well. For example, proteins often form physical protein-protein interactions and function at the level of macromolecular assemblies rather than individual domains. Selection for assembling correctly would place additional selection pressure to maintain a particular structure and to coevolve with other domains (76,77). Additionally, highly expressed proteins are under strong selection to avoid misfolding and aggregation (39,41). These and other selection pressures could also influence the rate of structure evolution with respect to sequence evolution, and to the extent that sufficient data exist, repeating our analysis by subdividing protein domains by characteristics other than contact density could test this hypothesis. Integrating these various molecular properties could yield even better predictions of structure evolution rate than contact density alone. Combining information about protein stability, abundance, catalytic rate, and organism population size has already proved a fruitful method of predicting the strength of selection and fate of arising mutations (78).

CONCLUSIONS

In this study, we found that the relationship between sequence and structure divergence, once thought to be universal, differs for protein domains with different contact densities. The onset of the twilight zone is gradual for the least compact proteins but is abrupt for compact proteins. Because contact density is a proxy for thermodynamic stability, we hypothesized that protein thermodynamic stability correlates negatively with structure evolution rate. We established the mechanism that would underlie this negative correlation using simulations: strong evolutionary selection for stability erects evolutionary barriers to structure discovery because structure discovery typically passes through unstable evolutionary intermediates. Altogether, this work provides a powerful yet simple interpretive framework that unifies many longstanding observations in protein biophysics.

SUPPORTING MATERIAL

Supporting Materials and Methods, thirteen figures, two tables, and two data files are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(17)30243-6.

AUTHOR CONTRIBUTIONS

A.I.G. and E.I.S. designed the research; A.I.G. carried out the bioinformatics analysis and developed and performed Monte Carlo simulations;

A.M.-C. performed the multiscale evolution simulations and developed the analytical model under the guidance of A.I.G.; data was contributed and analyzed by J.-M.C.; and A.I.G. and E.I.S. prepared the article.

ACKNOWLEDGMENTS

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 discussions, especially with Ned Wingreen and other participants in the discussion group organized by Tal Einav.

This work was supported by NIH grant No. GM068670 and a National Science Foundation Graduate Research Fellowship (awarded to A.I.G.).

REFERENCES

1. Zhang, J., and J.-R. Yang. 2015. Determinants of the rate of protein sequence evolution. *Nat. Rev. Genet.* 16:409–420.
2. Xia, Y., E. A. Franzosa, and M. B. Gerstein. 2009. Integrated assessment of genomic correlates of protein evolutionary rate. *PLOS Comput. Biol.* 5:e1000413.
3. Sadowski, M. I., and W. R. Taylor. 2010. Protein structures, folds and fold spaces. *J. Phys. Condens. Matter.* 22:033103.
4. Wang, M., Y. Y. Jiang, …, G. Caetano-Anollés. 2011. A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation. *Mol. Biol. Evol.* 28:567–582.
5. Chothia, C., and A. M. Lesk. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5:823–826.
6. Wilson, C. A., J. Kreychman, and M. Gerstein. 2000. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* 297:233–249.
7. Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Eng.* 12:85–94.
8. Wood, T. C., and W. R. Pearson. 1999. Evolution of protein sequences and structures. *J. Mol. Biol.* 291:977–995.
9. Panchenko, A. R., Y. I. Wolf, …, T. Madej. 2005. Evolutionary plasticity of protein families: coupling between sequence and structure variation. *Proteins.* 61:535–544.
10. Bordoli, L., F. Kiefer, …, T. Schwede. 2009. Protein structure homology modeling using SWISS-MODEL workspace. *Nat. Protoc.* 4:1–13.
11. Martí-Renom, M. A., A. C. Stuart, …, A. Sali. 2000. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* 29:291–325.
12. Chung, S. Y., and S. Subbiah. 1996. A structural explanation for the twilight zone of protein sequence homology. *Structure.* 4:1123–1127.
13. Khor, B. Y., G. J. Tye, …, Y. S. Choong. 2015. General overview on structure prediction of twilight-zone proteins. *Theor. Biol. Med. Model.* 12:15.
14. Tokuriki, N., F. Stricher, …, D. S. Tawfik. 2007. The stability effects of protein mutations appear to be universally distributed. *J. Mol. Biol.* 369:1318–1332.
15. Dokholyan, N. V., B. Shakhnovich, and E. I. Shakhnovich. 2002. Expanding protein universe and its origin from the biological Big Bang. *Proc. Natl. Acad. Sci. USA.* 99:14132–14136.
16. Deeds, E. J., N. V. Dokholyan, and E. I. Shakhnovich. 2003. Protein evolution within a structural space. *Biophys. J.* 85:2962–2972.
17. Koonin, E. V. 2011. Are there laws of genome evolution? *PLOS Comput. Biol.* 7:e1002173.
18. Roessler, C. G., B. M. Hall, …, M. H. J. Cordes. 2008. Transitive homology-guided structural studies lead to discovery of Cro proteins with 40% sequence identity but different folds. *Proc. Natl. Acad. Sci. USA.* 105:2343–2348.
19. Alexander, P. A., Y. He, …, P. N. Bryan. 2009. A minimal sequence code for switching protein structure and function. *Proc. Natl. Acad. Sci. USA.* 106:21149–21154.
20. Burmann, B. M., S. H. Knauer, …, P. Rösch. 2012. An α helix to β barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell.* 150:291–303.
21. Eaton, K. V., W. J. Anderson, …, M. H. J. Cordes. 2015. Studying protein fold evolution with hybrids of differently folded homologs. *Protein Eng. Des. Sel.* 28:241–250.
22. Alexander, P. A., Y. He, …, P. N. Bryan. 2007. The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proc. Natl. Acad. Sci. USA.* 104:11963–11968.
23. He, Y., Y. Chen, …, J. Orban. 2012. Mutational tipping points for switching protein folds and functions. *Structure.* 20:283–291.
24. Meyerguz, L., J. Kleinberg, and R. Elber. 2007. The network of sequence flow between protein structures. *Proc. Natl. Acad. Sci. USA.* 104:11627–11632.
25. Burke, S., and R. Elber. 2012. Super folds, networks, and barriers. *Proteins.* 80:463–470.
26. Hubbard, T. J. P., B. Ailey, …, C. Chothia. 1999. SCOP: a structural classification of proteins database. *Nucleic Acids Res.* 27:254–256.
27. Berezovsky, I. N., and E. I. Shakhnovich. 2005. Physics and evolution of thermophilic adaptation. *Proc. Natl. Acad. Sci. USA.* 102:12742–12747.
28. Dill, K. A., D. O. V. Alonso, and K. Hutchinson. 1989. Thermal stabilities of globular proteins. *Biochemistry.* 28:5439–5449.
29. Dill, K. A., K. Ghosh, and J. D. Schmit. 2011. Physical limits of cells and proteomes. *Proc. Natl. Acad. Sci. USA.* 108:17876–17882.
30. Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48:443–453.
31. Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA.* 89:10915–10919.
32. Zhang, Y., and J. Skolnick. 2004. Scoring function for automated assessment of protein structure template quality. *Proteins.* 57:702–710.
33. Zhang, Y., and J. Skolnick. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* 33:2302–2309.
34. Reva, B. A., A. V. Finkelstein, and J. Skolnick. 1998. What is the probability of a chance prediction of a protein structure with an RMSD of 6 Å? *Fold. Des.* 3:141–147.
35. Shakhnovich, E. I., and A. M. Gutin. 1990. Enumeration of all compact conformations of copolymers with random sequence of links. *J. Chem. Phys.* 93:5967–5971.
36. Heo, M., S. Maslov, and E. Shakhnovich. 2011. Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. *Proc. Natl. Acad. Sci. USA.* 108:4258–4263.
37. Miyazawa, S., and R. L. Jernigan. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* 256:623–644.
38. Sali, A., E. Shakhnovich, and M. Karplus. 1994. Kinetics of protein folding. A lattice model study of the requirements for folding to the native state. *J. Mol. Biol.* 235:1614–1636.
39. Drummond, D. A., and C. O. Wilke. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell.* 134:341–352.
40. Lobkovsky, A. E., Y. I. Wolf, and E. V. Koonin. 2010. Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proc. Natl. Acad. Sci. USA.* 107:2983–2988.
41. Serohijos, A. W. R., Z. Rimas, and E. I. Shakhnovich. 2012. Protein biophysics explains why highly abundant proteins evolve slowly. *Cell Reports.* 2:249–256.

42. Zeldovich, K. B., I. N. Berezovsky, and E. I. Shakhnovich. 2006. Physical origins of protein superfamilies. *J. Mol. Biol.* 357:1335–1343.

43. Zeldovich, K. B., P. Chen, …, E. I. Shakhnovich. 2007. A first-principles model of early evolution: emergence of gene families, species, and preferred protein folds. *PLOS Comput. Biol.* 3:e139.

44. Fox, N. K., S. E. Brenner, and J. M. Chandonia. 2014. SCOPe: structural classification of proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* 42:D304–D309.

45. Bloom, J. D., A. Raval, and C. O. Wilke. 2007. Thermodynamics of neutral protein evolution. *Genetics.* 175:255–266.

46. Zeldovich, K. B., P. Chen, and E. I. Shakhnovich. 2007. Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc. Natl. Acad. Sci. USA.* 104:16152–16157.

47. Serohijos, A. W., and E. I. Shakhnovich. 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.

48. Bershtein, S., W. Mu, …, E. I. Shakhnovich. 2013. Protein quality control acts on folding intermediates to shape the effects of mutations on organismal fitness. *Mol. Cell.* 49:133–144.

49. Jacquier, H., A. Birgy, …, O. Tenaillon. 2013. Capturing the mutational landscape of the β-lactamase TEM-1. *Proc. Natl. Acad. Sci. USA.* 110:13067–13072.

50. Dokholyan, N. V., and E. I. Shakhnovich. 2001. Understanding hierarchical protein evolution from first principles. *J. Mol. Biol.* 312:289–307.

51. Jones, D. T., C. M. Moody, …, J. M. Thornton. 1996. Towards meeting the Paracelsus Challenge: the design, synthesis, and characterization of paracelsin-43, an α-helical protein with over 50% sequence identity to an all-β protein. *Proteins.* 24:502–513.

52. Dalal, S., S. Balasubramanian, and L. Regan. 1997. Protein alchemy: changing β-sheet into α-helix. *Nat. Struct. Biol.* 4:548–552.

53. Dalal, S., and L. Regan. 2000. Understanding the sequence determinants of conformational switching using protein design. *Protein Sci.* 9:1651–1659.

54. Shakhnovich, B. E., E. Deeds, …, E. Shakhnovich. 2005. Protein structure and evolutionary history determine sequence space topology. *Genome Res.* 15:385–392.

55. Shakhnovich, E. I., and A. M. Gutin. 1993. Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci. USA.* 90:7195–7199.

56. Li, H., R. Helling, …, N. Wingreen. 1996. Emergence of preferred structures in a simple model of protein folding. *Science.* 273:666–669.

57. Ramanathan, S., and E. Shakhnovich. 1994. Statistical mechanics of proteins with "evolutionary selected" sequences. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics.* 50:1303–1312.

58. Holzgräfe, C., and S. Wallin. 2014. Smooth functional transition along a mutational pathway with an abrupt protein fold switch. *Biophys. J.* 107:1217–1225.

59. Chen, S. H., and R. Elber. 2014. The energy landscape of a protein switch. *Phys. Chem. Chem. Phys.* 16:6407–6421.

60. Chandler, D. 1986. Roles of classical dynamics and quantum dynamics on activated processes occurring in liquids. *J. Stat. Phys.* 42:49–67.

61. Sella, G., and A. E. Hirsh. 2005. The application of statistical physics to evolutionary biology. *Proc. Natl. Acad. Sci. USA.* 102:9541–9546.

62. Hartl, D. L., and A. G. Clark. 2006. Principles of Population Genetics. Sinauer Associates, Sunderland, MA.

63. Illergård, K., D. H. Ardell, and A. Elofsson. 2009. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins.* 77:499–508.

64. Stoycheva, A. D., C. L. Brooks, 3rd, and J. N. Onuchic. 2004. Gatekeepers in the ribosomal protein S6: thermodynamics, kinetics, and folding pathways revealed by a minimalist protein model. *J. Mol. Biol.* 340:571–585.

65. Kurnik, M., L. Hedberg, …, M. Oliveberg. 2012. Folding without charges. *Proc. Natl. Acad. Sci. USA.* 109:5705–5710.

66. Mirny, L. A., and E. I. Shakhnovich. 1999. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* 291:177–196.

67. Abkevich, V. I., A. M. Gutin, and E. I. Shakhnovich. 1994. Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry.* 33:10026–10036.

68. Toll-Riera, M., D. Bostick, …, J. B. Plotkin. 2012. Structure and age jointly influence rates of protein evolution. *PLOS Comput. Biol.* 8:e1002542.

69. Bloom, J. D., S. T. Labthavikul, …, F. H. Arnold. 2006. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. USA.* 103:5869–5874.

70. Serohijos, A. W. R., S. Y. R. Lee, and E. I. Shakhnovich. 2013. Highly abundant proteins favor more stable 3D structures in yeast. *Biophys. J.* 104:L1–L3.

71. Zhou, T., D. A. Drummond, and C. O. Wilke. 2008. Contact density affects protein evolutionary rate from bacteria to animals. *J. Mol. Evol.* 66:395–404.

72. Peisajovich, S. G., L. Rockah, and D. S. Tawfik. 2006. Evolution of new protein topologies through multistep gene rearrangements. *Nat. Genet.* 38:168–174.

73. Jacquin, H., A. Gilson, …, R. Monasson. 2016. Benchmarking inverse statistical approaches for protein structure and design with exactly solvable models. *PLOS Comput. Biol.* 12:e1004889.

74. Morcos, F., A. Pagnani, …, M. Weigt. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. USA.* 108:E1293–E1301.

75. Ovchinnikov, S., H. Kamisetty, and D. Baker. 2014. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife.* 3:e02030.

76. Fraser, H. B., A. E. Hirsh, …, M. W. Feldman. 2002. Evolutionary rate in the protein interaction network. *Science.* 296:750–752.

77. Kim, P. M., L. J. Lu, …, M. B. Gerstein. 2006. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science.* 314:1938–1941.

78. Rotem, A., A. W. R. Serohijos, …, E. I. Shakhnovich. 2016. Tuning the course of evolution on the biophysical fitness landscape of an RNA virus. *bioRxiv.* https://doi.org/10.1101/090258.