Beyond thermodynamic constraints: Evolutionary history shapes protein sequence variation.

Qian Jiang∗†, Ashley I. Teufel∗, Eleisha L. Jackson∗ and Claus O. Wilke∗†

∗Department of Integrative Biology, The University of Texas at Austin, Austin, TX, USA, 78712, †State Key Laboratory of Crop Genetics and Germplasm Enhancement, College of Horticulture, Nanjing Agricultural University, Nanjing, China, 210095

ABSTRACT Biological evolution generates a surprising amount of site-specific variability in protein sequences. Yet attempts at modeling this process have been only moderately successful, and current models based on protein structural metrics explain, at best, 60% of the observed variation. Surprisingly, simple measures of protein structure, such as solvent accessibility, are often better predictors of site-specific variability than more complex models employing all-atom energy functions and detailed structural modeling. We suggest here that these more complex models perform poorly because they lack consideration of the evolutionary process that is in part captured by the simpler metrics. We compare protein sequences that are computationally evolved using the same protein-design energy function and to homologous natural sequences. We find that by a wide variety of metrics, evolved sequences are much more similar to natural sequences than are designed sequences. In particular, designed sequences are too variable in the protein core relative to natural sequences whereas evolved sequences are not. We attribute these differences to the context-dependent nature of the substitution process under evolution. The substitution of a specific amino acid is often not possible due to conflicts with the surrounding sequence background, even if wholesale rearrangements of many amino acids in the sequence could be made to accommodate the specific change. In aggregate, thus, the observation in natural sequences that buried sites are much more conserved than exposed sites may not primarily be driven by global thermodynamic constraints imposed by the protein structure but rather by the entrenchment of specific amino acids in the context of their surrounding sequence and structural background.

KEYWORDS protein evolution, site specific variability, protein structure

Introduction
Alignments of protein sequences can display large amounts of position-specific variability. This variability exists because selective pressures for proteins to fold and function are not experienced uniformly across a protein’s sequence. The relative variability of a position in a sequence is often associated with where that position is located in the protein’s structure (Thorne 2007; Liberles et al. 2012; Arenas et al. 2015; Echave et al. 2016). For example, patterns of sequence variation differ between the core and surface of proteins. The core tends to evolve more slowly and be more conserved than sites on the surface (Kimura and Ohta 1974; Overington et al. 1992; Mirny and Shakhnovich 1999; Franzosa and Xia 2009; Ramsey et al. 2011; Tóth-Petróczy and Tawfik 2011). These core sites also tend to have more hydrophobic residues while sites on the surface tend to have more polar residues (Jones and Thornton 1996; Bastolla et al. 2007; Jackson et al. 2013). In general, sites that are conserved are assumed to be more important for protein structure, stability, and ultimately function, even though most sites in a protein are not directly involved in function.

While a significant body of literature exists examining how protein structure and thermodynamic constraints interact with evolutionary processes to shape site-specific variability, current models explain, at best, 60% of variability (Echave et al. 2016). Existing models span from ones using very simple metrics, such as relative solvent accessibility (RSA) or weighted contact number (WCN), to ones using complicated all-atom energy functions and protein design. One surprising result from these efforts is that
the simple models tend to perform as well as or better than the complicated models. In particular, protein design has worked particularly poorly in predicting site variability observed in natural sequence alignments (Jackson et al. 2013; Shahmoradi et al. 2014; Jackson et al. 2016).

Why the more complex models fail to explain observed sequence variation is unknown. All-atom stability models should account for the same properties reflected in RSA and WCN, as well as an array of other features. For example, the mean interaction energy of a site with the rest of the protein is a function of a site’s contacts, reflected in WCN (Halle 2002). And thermodynamic stability is related to RSA through the energetic cost of burying a side chain in the protein core (Zhou and Zhou 2004).

However, protein design methods deviate in one crucial aspect from the process of protein evolution as it plays out in living organisms: Under evolution, mutations are introduced one by one and either fix or are lost to drift. As a consequence, a mutation that is deleterious in the current genetic background will rarely fix, even if it were acceptable in other genetic backgrounds. Protein design, by contrast, allows the replacement of multiple amino acids at once, and thus can create sequence configuration that are thermodynamically allowed but not easily accessible by evolution. In fact, a number of recent studies have begun to examine how the state of the background sequence affects a protein’s ability to accumulate substitutions (Pollock et al. 2012; Shah et al. 2015; Goldstein and Pollock 2016; McCandlish et al. 2016). These studies suggest that the order in which changes to a sequence occur can affect a position’s propensity to tolerate further mutations. Therefore, a protein’s evolutionary history influences how mutational space is sampled, and the order in which states were sampled and the length of time any specific state has been resident influences sampling behavior. Hence, the effect the evolutionary process has on shaping site-specific variability could be considerable.

Here we explore the influence of evolutionary history in the generation of site-specific variability. By comparing protein sequences produced under a model that combines evolutionary history and stability constraints to sequences produced by a model that considers stability constraints alone, we can begin to tease apart the relationship between evolutionary history, protein stability, and sequence variability. To generate sequences in the absence of evolutionary history, we employ protein design as implemented in the RosettaDesign suite (Kuhlman et al. 2003). In this method, all of the residues are replaced simultaneously. To add evolutionary history, we use the same Rosetta force field but introduce substitutions sequentially, in a process that mimics evolution under natural selection (Teufel and Wilke 2017). We then compare the sequences produced by each of these methods to one another and to natural sequences. We find that proteins generated by simulating evolution display similar site-specific variability to natural proteins whereas designed sequences do not. This relationship highlights that a significant portion of site-specific variability is due to a combination of evolutionary history and stability constraints rather than to stability constraints alone.

Methods

Protein structures and alignments

We analyzed a set of 38 protein with available structures from Saccharomyces cerevisiae. This dataset had previously been studied in the context of protein design by Jackson et al. (2013), and it had originally been assembled by Ramsey et al. (2011). For each structure, Ramsey et al. (2011) had also assembled alignments of homologous sequences containing at least 50 sequences each. Table S1 lists the protein data bank identifier (PDB ID) for each of the template structures and the number of homologous sequences available in the respective alignment.

Generation of protein sequences

For each of the 38 protein structures, we computationally generated alignments of 500 variant sequences both via protein design and via protein evolution. In all cases, we first minimized the structures with Rosetta (Leaver-Fay et al. 2011) to assure that any changes in stability during the evolutionary simulations is due to the introduced mutations rather than to repacking of amino-acid side chains.

For protein design, we used the fixed-backbone method implemented in RosettaDesign (Kuhlman et al. 2003). This method only allows for movement of the side chains, while the backbone is held fixed. The following command was used:

```
./fixbb.linuxgccrelease -database rosetta_database -s input.pdb -resfile ALLA.res -ex1 -ex2 -extrachi_cutoff 0 -nstruct 1 -overwrite -minimize_sidechains -linmem_ig 10
```

For evolutionary simulations, we used an accelerated origin-fixation algorithm (Kachroo et al. 2015; Teufel and Wilke 2017). In an origin-fixation model, mutations are sequentially introduced and either accepted or rejected based on their effect on fitness. Under the accelerated version of this model, beneficial mutations are always fixed while deleterious mutations are exponentially suppressed. We have previously shown that this accelerated algorithm produces the same steady-state distribution of genotypes visited as the regular, non-accelerated algorithm (Teufel and Wilke 2017).

To calculate stabilities for proteins during simulated evolution, we used the get_fa_scorefxn function in the PyRosetta (Chaudhury et al. 2010), and interpreted its result as the stability $\Delta G$ of each proposed mutation. To convert protein stability into fitness, we use a soft-threshold model (Chen and Shakhnovich 2009; Wylie and Shakhnovich 2011; Serohijos et al. 2012), where the fitness $f_i$ of a protein $i$ with stability $\Delta G_i$ is given by

$$f_i = \frac{1}{e^{\beta(\Delta G_i - \Delta G_{\text{thresh}})} + 1}.$$

(1)

Here, $\beta$ is the inverse temperature, $\Delta G_i$ is the stability of the mutant protein, and $\Delta G_{\text{thresh}}$ is the stability value at which fitness has declined to 50% of its maximum. For the majority of our simulations, we set $\Delta G_{\text{thresh}}$ to the average score obtained for the proteins designed on the same template structure. This was done to insure that proteins generated by the evolutionary model have similar stabilities as those produced by the design method. In an additional set of simulations, we set $\Delta G_{\text{thresh}}$ to the maximum score (i.e., corresponding to the least stable protein) obtained for the designed proteins.

To calculate the probability of fixation of a new mutation, we first log-transform fitness,

$$x_i = \log(f_i) = -\log(e^{\beta(\Delta G_i - \Delta G_{\text{thresh}})} + 1).$$

(2)

We then calculate the probability of fixation of a new mutation $j$ in a background population of genotypes $i$ as

$$\pi(i \rightarrow j) \approx \begin{cases} 1 & \text{for } x_j > x_i, \\ e^{-2N_e(x_i - x_j)} & \text{otherwise}, \end{cases}$$

(3)
where \( N_e \) is the effective population size. In all simulations, we set \( N_e = 100 \) and \( \beta = 1 \), and we ran each simulation until 5000 substitutions had occurred.

**Data analysis**

**Site-specific variability and amino acid distributions:** We separately aligned the 500 resulting sequences produced by each method for each of the 38 structures. To quantify the variability of sites in these alignments, we calculated the site entropy

\[
H_i = - \sum p_{ij} \ln p_{ij},
\]

where \( p_{ij} \) is the frequency of amino acid \( j \) at column \( i \) in the alignment. Exponentiating \( H_i \), we obtain the effective number of amino acids,

\[
n_{eff} = \exp(H_i).
\]

This number falls between 1 and 20 and can be interpreted as the number of different amino-acid types present at a given site.

To compare an amino-acid distribution to a reference distribution (e.g., to compare the amino-acid distribution of designed sequences to that of natural sequences), we used the Kullback-Leibler (KL) divergence, defined as

\[
D_{KL} = \sum p_{ij} \ln(p_{ij}/q_{ij}).
\]

Here \( q_{ij} \) is the frequency of amino acid \( j \) in column \( i \) of the sequence alignment to compare, and \( p_{ij} \) is the relative frequency in the reference alignment. If any \( q_{ij} \) or \( p_{ij} \) were zero, we added 1/20 to each amino acid count before calculating the frequencies.

To compare natural alignment to themselves, we randomly split each alignment into two equal-sized sets of sequences and then calculated the KL divergence of the first half against the second.

**Residue buriedness:** To estimate a residue’s burial or exposure we calculated its relative solvent accessibility (RSA). RSA ranges from 0 for completely buried residues to 1 for completely exposed ones (Tien et al. 2013). We first calculated the Accessible Surface Area (ASA) of each residue in each structure, using the software DSSP (Kabsch and Sander 1983). ASA indicates the surface area of a given residue that is accessible to water. These ASA values were then normalized by the maximum ASA value for a given amino acid to obtain RSA (Tien et al. 2013). Residues at sites with higher RSA have a larger part of the residue’s surface exposed to solvent and are generally closer to the protein’s surface while residues with lower RSA are closer to protein’s core. We defined sites with RSA ≤ 0.05 as buried sites and sites with RSA > 0.05 as exposed.

**Packing density:** We estimated residue packing density via the side-chain weighted contact number (WCN), defined as

\[
WCN_i = \sum_{j \neq i} 1/r_{ij}^2
\]

where \( j \) indicates the focal residue, \( r_{ij} \) is the distance between the geometric centers of the side chains of the focal residue \( i \) and of residue \( j \), and the sum runs over all residues \( j \) in the protein. Since packing density tends to have a negative correlation with site entropy and RSA, here we use the inverse of WCN (1/WCN) for all correlation calculations, as was done previously by Shahmoradi et al. (2014).

We used side-chain WCN as defined above rather than WCN calculated from distances of \( C_\alpha \) atoms because side-chain WCN provides the more robust determinant of evolutionary variation (Marcos and Echave 2015; Shahmoradi and Wilke 2016).

**Score matching:** To verify that our results were not caused by differences in stability between the designed and evolved sequences, we generated, for each template protein structure, subset alignments with matched stability scores between the designed and evolved sequences. We carried out this matching as follows. We first identified the intersection of the stability ranges between designed and evolved proteins. This intersection generally coincided with the stability range of the evolved proteins, i.e., evolved proteins had a narrower stability range than the corresponding designed proteins. We then identified, for each designed protein in that range, the evolved protein with the most similar score. Each designed protein was matched with exactly one unique evolved protein, and we stopped the matching step when there were no more designed or evolved proteins available for matching in the stability-range intersection. The resulting matched alignments consisted of between 19 and 433 sequences, with a median of 70.

**Data availability**

All data and analysis scripts are available at: https://github.com/reductase4/evol_sim_vs_rosetta.git

**Results**

To examine how evolutionary history affects the emergences of sequence variability, we analyzed two distinct sets of protein sequences, one produced by protein design (no evolutionary history, all amino acids are replaced at once) and one produced by evolutionary simulation (amino acids are introduced one at a time, and the protein need to remain viable at all times). Importantly, for both approaches we used the same methods to replace amino acids and evaluate the energy of the resulting protein structures, based on the fixed-backbone protein-design algorithm implemented in Rosetta (Kuhlman et al. 2003; Leaver-Fay et al. 2011). For each approach, we generated 500 sequences each from 38 template structures. To compare these sequences to natural sequences, we analyzed alignments of at least 50 homologs for each of the 38 protein structures, taken from Jackson et al. (2013) (see also Table S1). We found that the sequence divergence in the simulated alignments was comparable to that of the natural sequences, even though divergence in designed sequences was somewhat larger than divergence in evolved sequences (Figure 1).

**Amino-acid distributions**

We first compared the overall amino-acid frequencies between natural and simulated sequences (Figure S1), because prior work comparing designed sequences to the same alignments of natural sequences had shown significant discrepancies, in particular for hydrophobic residues in buried sites (Jackson et al. 2013). We found that these discrepancies had mostly disappeared in our newly generated dataset. This difference may be due to an additional round of energy minimization we performed here, or (more likely) to recent updates to the energy function in Rosetta (Leaver-Fay et al. 2011). However, there was now a substantial discrepancy for Lysine, in that it was much more prevalent at exposed sites in designed sequences than in natural sequences. Notably, there were only minor differences in amino-acid frequencies between evolved sequences and natural sequences. In summary, while evolutionary simulations outperformed protein design specifically for Lysine, overall the differences between both approaches and natural sequences were minor.
Figure 1 Mean sequence divergence from the starting sequence template in designed, evolved, and natural sequences. For natural sequences, we considered the sequence of the focal protein structure as the starting sequence template. The mean sequence divergence was higher in designed sequences than in both evolved (paired \( t \)-test, \( p < 2.2 \times 10^{-16} \)) and natural (paired \( t \)-test, \( p = 5.58 \times 10^{-13} \)) sequences, but it was comparable between evolved and natural sequences (paired \( t \)-test, \( p = 0.37 \)).

While the analysis of aggregate amino-acid frequencies is useful as a first sanity check, it does not address whether either simulation approach places the correct amino acids at individual sites. Therefore, we next calculated the average Kullback-Leibler (KL) divergence for each protein, which quantifies the extent to which site-specific amino-acid distributions of simulated sequences are comparable to those of natural sequences. The lower the KL divergence, the more similar the distributions. As a control, we compared natural sequence to themselves, by randomly dividing each alignment into two groups and then comparing one to the other. We found that the evolutionary simulation produced sequences that were more similar to natural sequences than were the designed sequences (paired \( t \)-test, \( p < 2.2 \times 10^{-16} \), Figure 2). Two exceptions (indicated as outliers in the middle boxplot of Figure 2) were chain D from the Sfi1p/Cdc31p complex (PDB ID: 2GV5) and initiation factor 5a (PDB ID:1XTD). This analysis suggests that accounting for evolutionary history is an important component in simulating realistic protein sequences.

**Patterns of site variability**

To investigate site-specific sequence variability we calculated the effective number of amino acids \( n_{\text{eff}} \) (Eq. 5) for each site in each protein. We found that sequences produced by simulating protein evolution showed similar overall variability to natural sequences (paired \( t \)-test, \( p = 0.68 \), Figure 3), whereas designed sequences had significantly less variability compared to natural sequences (paired \( t \)-test, \( p < 2.2 \times 10^{-16} \), Figure 3). Importantly, designed sequences had a lower mean effective number of amino acids even though overall sequence divergence was higher in designed sequences than in evolved sequences (Figure 1). This seemingly paradoxical observation implies that designed sequences, relative to the evolved sequences, experienced changes at a larger number of sites but towards a smaller set of different amino acids.

Figure 2 Mean Kullback-Leibler (KL) divergence of designed, evolved, and natural sequences to natural sequences. Natural sequences were compared to themselves by randomly splitting alignments into two groups and calculating the KL divergence between them. A lower KL divergence indicates that the amino acid distributions at individual sites are closer to that of natural proteins.

Figure 3 Mean effective number of amino acids for designed, evolved, and natural sequences. Evolved and natural sequences had comparable mean effective numbers of amino acids (paired \( t \)-test, \( p = 0.68 \)) whereas designed sequences had significantly lower mean effective numbers (paired \( t \)-test, \( p < 2.2 \times 10^{-16} \)). One exception was chain D of the Sfi1p/Cdc31p complex (PDB ID: 2GV5, outlying data point in the boxplot for evolved sequences), for which evolutionary simulation yielded a much smaller mean effective number of amino acids relative to all other cases.
Just because two alignments have a comparable mean $n_{\text{eff}}$ does not mean that the same sites are more variable or more conserved in the two alignments. Therefore, we next calculated the correlations between $n_{\text{eff}}$ among alignments generated by different methods (designed sequences, evolved sequences, natural sequences). We found that the correlations in site variability between evolved and natural sequences were much higher than those between designed and natural sequences (Figure 4). The former were all positive and ranged between $\sim 0.2$ and $0.7$, whereas the latter did not exceed $\sim 0.3$ and several fell below zero. Thus variable and conserved sites in evolved sequences tend to coincide with the same types of sites in natural sequence alignments, but the same is not true for designed sequences.

![Figure 4](image)

**Figure 4** Spearman correlations between $n_{\text{eff}}$ for natural and either evolved or designed sequences. Each dot represents the correlation coefficients for one protein. Correlations are substantially higher when comparing evolved to natural sequences than when comparing designed to natural sequences (paired $t$ test, $p < 2.2 \times 10^{-16}$).

**Site variability in the context of structural features**

Site-specific variability tends to correlate with features in the protein structure, most notably solvent exposure and packing density (Kimura and Ohta 1974; Yeh et al. 2013; Chang et al. 2013; Huang et al. 2014; Echave et al. 2016). Since these variability patterns are likely driven by biophysical constraints generated from the protein structure, we would expect that they would be recapitulated in computationally designed proteins. Yet, prior work had shown that site-variability patterns in designed proteins did not recapitulate the patterns observed in natural sequences—buried sites in designed sequences were not sufficiently conserved, or alternatively, exposed sites not sufficiently variable (Jackson et al. 2013).

We found here again that site variability in designed proteins did not appreciably correlate with relative solvent accessibility (RSA) (Figure 5). Correlations between $n_{\text{eff}}$ and RSA ranged between $-0.2$ and $0.2$. In comparison, for natural sequences these correlations fell mostly between $0.2$ and $0.6$. For evolved sequences, we observed even higher correlations, with most values falling between $0.5$ and $0.7$ (Figure 5). Results were similar when we used inverse Weighted Contact Number (iWCN) instead of RSA (Figure S2).

![Figure 5](image)

**Figure 5** Spearman correlations between RSA and $n_{\text{eff}}$ for designed, evolved, and natural sequences. Correlations for designed sequences were significantly lower than for natural sequences (paired $t$-test, $p = 1.5 \times 10^{-14}$). By contrast, correlations for evolved sequences were significantly higher than for natural sequences (paired $t$-test, $p = 6.48 \times 10^{-12}$). The individual correlation coefficients for each structure are shown in Figure S2.

One potential caveat to these findings is that designed and evolved sequences may fall into a different range of thermodynamic stability, assessed in our model via the Rosetta score. Indeed, even though we calibrated the stability threshold $\Delta G_{\text{thresh}}$ used during evolution to the mean stability for proteins designed to the same template (see Methods), we found that evolved proteins generally had a narrower range of stabilities than designed proteins and were, on average, more stable (Figure S3). In principle, these differences in stability distributions could be the cause for the other observed differences between designed and evolved sequences.

We addressed this caveat in two ways. First, we generated alignment subsets for each template structure such that each designed sequence retained in an alignment was matched one-to-one to a unique evolved sequence with comparable stability score and homologous structure. This procedure yielded near-identical stability distributions in most cases (Figure S4). Yet the observed pattern of RSA-$n_{\text{eff}}$ correlations was virtually unchanged from that in the original dataset (Figure S5A). Second, for five arbitrarily chosen structures, we ran evolutionary simulations where we used the stability of the least stable designed structure (i.e., the maximum observed score among the designed structures) as stability threshold value. In those simulations, evolved sequences were indeed much less stable than before (Figure S6). Yet again, there was virtually no change in the observed pattern of RSA-$n_{\text{eff}}$ correlations (Figure S5B).

**Evolving designed sequences**

To recap, we have found that by any metric considered, evolved sequence alignment looked much more realistic than designed...
Figure 6 Comparison between designed sequences and sequences evolved from design. (A) Mean Kullback-Leibler (KL) divergence of designed, evolved from design, and natural sequences. The KL divergence was lower for evolved from design sequences than for designed sequences (paired t-test, $p = 2.57 \times 10^{-6}$). (B) Mean effective number of amino acids for designed, evolved from design, and natural sequences. The mean effective number $n_{\text{eff}}$ was higher for evolved from design sequences than for designed sequences (paired t-test, $p = 1.91 \times 10^{-5}$). (C) Spearman correlations of $n_{\text{eff}}$ between natural and evolved from design sequences and natural and designed sequences. Each dot represents one correlation coefficient value for one protein. Correlations with $n_{\text{eff}}$ from natural sequences were significantly higher for $n_{\text{eff}}$ from evolved from design sequences than for $n_{\text{eff}}$ from designed sequences (paired t-test, $p = 6.23 \times 10^{-5}$). (D) Spearman correlations between site entropy and RSA of designed, evolved from design, and natural sequences. Correlations were significantly higher for evolved from design sequences than designed sequences (paired t-test, $p = 1.7 \times 10^{-7}$), however correlations for evolved from design and natural sequences were comparable (paired t-test, $p = 0.62$).
Figure 7 Spearman correlations between $n_{\text{eff}}$ for natural sequences and $n_{\text{eff}}$ for simulated sequences values versus Spearman correlations between $n_{\text{eff}}$ for natural sequences structural measures RSA (A), (B), (C) or iWCN (D), (E), (F). These comparisons allow us to assess whether $n_{\text{eff}}$ in natural sequences correlates better with $n_{\text{eff}}$ in simulated sequences or with structural measures. We find that evolved sequences perform similarly to RSA and iWCN, but designed sequences do not.
sequence alignments. Further, we have seen that these differences between evolved and designed sequences are not caused by differences in protein stability. However, there do remain two potential reasons why we may have obtained these results: (i) evolutionary simulation adds an important element to sequence generation, one that is missed in the currently used protein design algorithm; (ii) evolved sequences look more similar to natural sequences simply because they have not diverged as much and thus retain much historical information about the natural ancestral sequence. To distinguish between these two scenarios, we ran an additional set of evolutionary simulations where we started each replicate evolutionary trajectory from one of the previously designed sequences. We performed this additional set of simulations for a subset of 10 arbitrarily chosen structures (Table S1). We refer to sequences generated in this manner as “evolved from design.” Importantly, the mean sequence divergence in these sequences was significantly higher than in the corresponding natural or evolved sequences, and almost as high as in the designed sequences (Figure S7).

From these additional simulations, we found that the evolutionary process produced more naturally-looking alignments even when designed proteins were used as starting points (Figure 6). Sequences that were evolved from design had a lower KL divergence than designed sequences (Figure 6A), a higher mean effective number of amino acids (Figure 6B), higher correlations of site-variability patterns to natural alignments (Figure 6C), and higher correlations between $n_{eff}$ and RSA (Figure 6D).

However, by all these metrics, sequences evolved from design were intermediate between designed sequences and sequences evolved from a natural sequence.

**Using simulated sequences as a predictor of site variability**

Finally, we asked how well $n_{eff}$ from simulated sequences performed as predictor of natural site variability relative to the other commonly used predictors RSA and WCN. In agreement with prior work (Shahmoradi et al. 2014; Jackson et al. 2016), we saw that the $n_{eff}$ from designed sequences performed poorly relative to RSA (Figure 7A). However, the $n_{eff}$ from evolved sequences performed similarly to RSA (Figure 7B), and $n_{eff}$ from evolved from design sequences displayed intermediate performance (Figure 7C). Results were similar for the inverse weighted contact number (iWCN), except that correlations between $n_{eff}$ and iWCN tended to be somewhat stronger than they were between $n_{eff}$ and RSA (Figure 7D, E, F), consistent with a large body of prior work finding that WCN tends to correlate more strongly with evolutionary variation than RSA does (Yeh et al. 2013, 2014; Shahmoradi et al. 2014; Marcos and Echave 2015; Echave et al. 2016). Unlike RSA, WCN measures both the local structural constraint around a residue and the global arrangement of amino acids in the entire structure Shahmoradi and Wilke (2016). The elevated correlations seen with WCN relative to RSA may thus in part be driven by factors other than folding constraints, such as the location of active sites in enzymes (Jackson et al. 2016).

**Discussion**

We have examined how evolutionary history affects the emergence of sequence variability, by comparing protein alignments generated by two different methods. The first was a traditional protein-design algorithm implemented in RosettaDesign (Kuhlman et al. 2003). This method imposes only thermodynamic constraints on the simulated sequences. The other was an evolutionary method that used the Rosetta energy function to simulate protein evolution according to population-genetics principles (Teufel and Wilke 2017). This method imposes the same thermodynamic constraints on the simulated sequences but adds evolutionary constraints as well. We then compared these simulated alignments to homologous alignments of natural sequences. We found that sequences generated by simulated evolution displayed site-specific variation quite similar to that of natural sequences, whereas patterns of variation were substantially different in designed sequences. Our evolved sequences also showed correlations between sequence variation and residue burial or packing similar to that of natural sequences, whereas the designed sequences did not. Finally, by simulating additional evolution with designed sequences as starting points, we demonstrated that the improvements in site-variability metrics were caused by the evolutionary process itself, and not by insufficient divergence from the starting sequence during the evolutionary simulations.

While both evolutionary and protein-design methods can be used to generate proteins that fold and function, they differ in how they explore sequence space. In particular, the widely used RosettaDesign (Kuhlman et al. 2003) takes a template structure, strips all residue side chains, and then simultaneously replaces them by different side chains. Subsequently additional sidechain replacements are made with the goal to maximize protein stability. This process of designing a sequence can be repeated a number of times to generate a set of diverse sequences. This and similar approaches have proven fruitful for protein engineering. For example, protein design methods have been used to engineer proteins that bind an influenza virus (Fleishman et al. 2011), to create enzymes (Röthlisberger et al. 2008), and to develop novel protein folds (Kuhlman et al. 2003). However, this method of generating functional proteins is very unlike how natural systems generate them. Natural proteins accumulate changes mostly sequentially, one mutation at a time, rather than having their entire sequence modified in a single instance. This sequential change means that not only the current state but also previous states may substantially shape which parts of the sequence space are visited during evolution (Pollock et al. 2012; Shah et al. 2015).

Nevertheless, one would expect that the sequence variability in particular at buried sites is determined by thermodynamic constraints imposed through the protein structure, and that these constraints are substantially similar for designed and evolved sequences. Instead, what seems to be happening is that the same constraints are much more limiting for single amino-acid replacements than for the wholesale replacement of many amino-acids in the structure at the same time. As a consequence, the most buried sites, which experience the strongest structural constraints, are rarely mutated during evolution. The amino acids at these sites remain conserved even if their replacement would be possible in principle as long as other, compensating replacements were made at the same time.

Our findings are consistent with recent works on evolutionary entrenchment (Pollock et al. 2012; McCandlish et al. 2016, 2015; Goldstein and Pollock 2016), which argue that the propensity of a protein to acquire position-specific substitutions varies over time, as previously accumulated mutations become entrenched in the protein structure and slowly alter the constraints imposed on other amino acids in the structure. We have shown here that this evolutionary entrenchment can be a significant driving force behind site-specific variability, contributing pos-
sibly as much to variability as the influence of thermodynamic constraints.

Even though our evolutionary simulations considered both structural constraints and entrenchment of prior mutations, we found that they did not fully capture the relationship between sequence variability, buriedness, and packing density observed in natural sequences. In particular, site-specific variability was correlated more strongly with RSA in our evolved sequences than is observed in nature. This exaggerated relationship between sequence variability and RSA may be caused by our model’s simplistic assumption that selection acts exclusively on protein stability. In natural organisms, numerous selection pressures beyond just protein stability act on protein sequences (Thorner et al.; Teufel et al., 2012; Serohijos and Shakhnovich 2013, 2014; Chi and Liberles 2016; Echave and Wilke 2017), and these other selection pressures should weaken the observed relationship between RSA and sequence variability.

There are several caveats to our work. First, throughout this project, we have used a fixed backbone model, even though allowing for a flexible backbone during design may produce more natural-looking sequences (Jackson et al., 2013; Olikainen and Korhut 2013). However, since we used the same fixed backbone model for both protein design and protein evolution, we don’t expect it to have much bearing on the observed differences between designed and evolved sequences. Moreover, Jackson et al. (2013) had previously shown that even with flexible backbone design the correlation between solvent accessibility and site variability was lower than observed in natural sequences. Second, the accelerated origin-fixation model we used for evolution changes the order in which substitutions are accumulated. However, Teufel and Wilke (2017) have argued that this effect should be relatively minor and should have no bearing on the effect of amino-acid entrenchment which we believe is the root cause for the difference between designed and evolved sequences. Finally, the amount of divergence generated during simulated evolution depends on the length of time for which the simulations are run, and thus is highly dependent on the parameter choices made for the simulations. We chose to run the evolutionary simulations for 5000 substitutions in each trajectory (for proteins of at most several hundred amino acids in length), to ensure that the amount of divergence generated during simulation would exceed that observed in typical natural homologs.

Acknowledgments

This work was supported by National Institutes of Health Grants R01 GM088344 and National Science Foundation Cooperative agreement no. DBI-0939454 (BEACON Center). Computational resources were provided by The University of Texas at Austin’s Center for Computational Biology and Bio-informatics and the Stampede cluster at the Texas Advanced Computing Center. Q.J. was a visiting scholar at UT Austin (September 2015–September 2017) sponsored by the China Scholarship Council (CSC). E.L.J. was supported by NSF GRFP DGE-1110007.

Literature Cited


Table S1  PDB structures used in this study. For each structure, we show the PDB identifier (“PDB ID”), the chain, the length of the corresponding sequence in amino acids, and the number of sequences in the alignment of homologous natural sequences (“alignment size”). Structures that were subjected to the additional “evolved from design” protocol are highlighted in bold.

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>Chain</th>
<th>Sequence length</th>
<th>Alignment size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1CI0</td>
<td>A</td>
<td>205</td>
<td>86</td>
</tr>
<tr>
<td>1G58</td>
<td>B</td>
<td>196</td>
<td>210</td>
</tr>
<tr>
<td>1HUJ</td>
<td>A</td>
<td>281</td>
<td>58</td>
</tr>
<tr>
<td>1IBS</td>
<td>A</td>
<td>297</td>
<td>235</td>
</tr>
<tr>
<td>1JLW</td>
<td>A</td>
<td>217</td>
<td>68</td>
</tr>
<tr>
<td>1KZL</td>
<td>A</td>
<td>202</td>
<td>63</td>
</tr>
<tr>
<td>1M3U</td>
<td>A</td>
<td>262</td>
<td>86</td>
</tr>
<tr>
<td>1MOZ</td>
<td>A</td>
<td>178</td>
<td>85</td>
</tr>
<tr>
<td>1PV1</td>
<td>A</td>
<td>290</td>
<td>121</td>
</tr>
<tr>
<td>1QMV</td>
<td>A</td>
<td>195</td>
<td>138</td>
</tr>
<tr>
<td>1RII</td>
<td>A</td>
<td>243</td>
<td>194</td>
</tr>
<tr>
<td>1V9S</td>
<td>B</td>
<td>201</td>
<td>166</td>
</tr>
<tr>
<td>1W7W</td>
<td>B</td>
<td>150</td>
<td>160</td>
</tr>
<tr>
<td>1XIO</td>
<td>B</td>
<td>276</td>
<td>136</td>
</tr>
<tr>
<td>1YPI</td>
<td>A</td>
<td>247</td>
<td>114</td>
</tr>
<tr>
<td>1ZNN</td>
<td>A</td>
<td>245</td>
<td>99</td>
</tr>
<tr>
<td>2A84</td>
<td>A</td>
<td>277</td>
<td>124</td>
</tr>
<tr>
<td>2BCG</td>
<td>Y</td>
<td>194</td>
<td>167</td>
</tr>
<tr>
<td>2BR9</td>
<td>A</td>
<td>230</td>
<td>95</td>
</tr>
<tr>
<td>2CJM</td>
<td>C</td>
<td>294</td>
<td>110</td>
</tr>
<tr>
<td>2ESF</td>
<td>A</td>
<td>213</td>
<td>98</td>
</tr>
<tr>
<td>2FLI</td>
<td>A</td>
<td>217</td>
<td>199</td>
</tr>
<tr>
<td>2GV5</td>
<td>D</td>
<td>142</td>
<td>58</td>
</tr>
<tr>
<td>1B4T</td>
<td>A</td>
<td>153</td>
<td>159</td>
</tr>
<tr>
<td>1EFV</td>
<td>B</td>
<td>252</td>
<td>83</td>
</tr>
<tr>
<td>1GV3</td>
<td>A</td>
<td>213</td>
<td>180</td>
</tr>
<tr>
<td>1HUR</td>
<td>A</td>
<td>180</td>
<td>106</td>
</tr>
<tr>
<td>1KY2</td>
<td>A</td>
<td>180</td>
<td>54</td>
</tr>
<tr>
<td>1OKC</td>
<td>A</td>
<td>292</td>
<td>60</td>
</tr>
<tr>
<td>1R6M</td>
<td>A</td>
<td>236</td>
<td>251</td>
</tr>
<tr>
<td>1XTD</td>
<td>A</td>
<td>145</td>
<td>65</td>
</tr>
<tr>
<td>1YSB</td>
<td>A</td>
<td>156</td>
<td>50</td>
</tr>
<tr>
<td>1ZWK</td>
<td>A</td>
<td>169</td>
<td>69</td>
</tr>
<tr>
<td>2AIU</td>
<td>A</td>
<td>104</td>
<td>72</td>
</tr>
<tr>
<td>2CFE</td>
<td>A</td>
<td>162</td>
<td>311</td>
</tr>
<tr>
<td>2CNV</td>
<td>A</td>
<td>299</td>
<td>126</td>
</tr>
<tr>
<td>2EU8</td>
<td>A</td>
<td>216</td>
<td>76</td>
</tr>
<tr>
<td>2G0N</td>
<td>B</td>
<td>178</td>
<td>113</td>
</tr>
</tbody>
</table>
Figure S1 Amino-acid frequencies in designed, evolved, and natural sequences. (A) Aggregate frequencies of residues at all sites in all alignments. (B) Aggregate frequencies of residues at exposed sites (defined as sites with RSA > 0.05) in all alignments. (C) Aggregate frequencies of residues at buried sites (defined as sites with RSA ≤ 0.05) in all alignments.
Figure S2 Spearman correlations between site entropy and RSA or iWCN. (A), (B), (C) Spearman correlations between site entropy and RSA. (D), (E), (F) Spearman correlations between site entropy and iWCN. Each dot represents one pair of correlation coefficients for one protein.
Figure S3 Density distributions of scores calculated from designed and evolved structures. The evolved structures were evolved using the mean score from the designed structures as the threshold value $G_{thesh}$. On average, evolved structures occupied a narrower range of stability scores and were more stable than designed structures.
Figure S4 Density distributions of scores calculated from designed and evolved structures after a subset of designed structures was matched to evolved structures with most similar scores. In most cases, the matching produced distributions that were very similar or nearly identical. Notable exceptions were 1RII_A, 1W7W_B, 1XTD_A, and 2CFE_A.
Figure S5 Spearman correlations between RSA and $n_{\text{eff}}$. (A) Spearman correlations between RSA and $n_{\text{eff}}$ for proteins with matched score distributions. See Fig. S3 for score distributions before matching and Fig. S4 for score distributions after matching. Evolved sequences had significantly higher correlations than designed sequences (paired t-test, $p = 2.2 \times 10^{-16}$). (B) Spearman correlations between RSA and $n_{\text{eff}}$ for five proteins (1B4T_A, 1EFV_B, 1GV3_A, 1HUR_A, 1KY2_A) for which we reran the evolutionary simulations using the maximum (least stable) score from protein design as the threshold value $G_{\text{thresh}}$. Shown are the distributions of correlation coefficients for the designed proteins, the proteins evolved using the mean score as threshold value, and the proteins evolved using the maximum score as threshold value. The score distributions from the latter simulations are shown in Fig. S6. Sequences evolved with either threshold value had significantly higher correlations than designed sequences (paired t-test, $p = 2.34 \times 10^{-6}$ and $p = 3.4 \times 10^{-6}$, respectively). There was no detectable difference in correlation strengths between the two datasets of evolved sequences (paired t-test, $p = 0.36$).
Figure S6 Density distribution of scores calculated from designed and evolved structures. The evolved structures were evolved with the maximum score from the designed structures used as the threshold value $G_{\text{thresh}}$. 
Figure S7 Mean sequence divergence from the starting sequence template in designed, evolved from designed, evolved, and natural sequences. Results are shown only for the 10 structures subjected to the additional “evolved from design” protocol (see Table S1). Designed and evolved-from-design sequences had significantly higher mean sequence divergence than natural sequences ($p = 0.00014$ and $p = 0.00018$, respectively, paired $t$-tests). However, designed sequences also had significantly higher mean sequence divergence than evolved-from-design sequences (paired $t$-test, $p = 0.0019$).