Beyond Thermodynamic Constraints: Evolutionary Sampling Generates Realistic Protein Sequence Variation

Qian Jiang,*,†,1 Ashley I. Teufel,*,1 Eleisha L. Jackson,* and Claus O. Wilke*,2
*Department of Integrative Biology, The University of Texas at Austin, Texas 78712 and †State Key Laboratory of Crop Genetics and Germplasm Enhancement, College of Horticulture, Nanjing Agricultural University, 210095, China
ORCID IDs: 0000-0002-2877-1179 (Q.J.); 0000-0002-1076-0439 (A.I.T.); 0000-0002-7470-9261 (C.O.W.)

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, which is, in part, captured by the simpler metrics. We compare protein sequences that are computationally designed to 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 conserved on the protein surface relative to natural sequences, whereas evolved sequences are not. Our results suggest that evolutionary simulation produces a realistic sampling of sequence space. By contrast, protein design—at least as currently implemented—does not. Existing energy functions seem to be sufficiently accurate to correctly describe the key thermodynamic constraints acting on protein sequences, but they need to be paired with realistic sampling schemes to generate realistic sequence alignments.

KEYWORDS protein evolution; site specific variability; protein structure

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 sequence. The relative variability of a position in a sequence is often associated with where that position is located in the protein 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

doi: https://doi.org/10.1534/genetics.118.300699
Manuscript received January 4, 2018; accepted for publication January 25, 2018; published Early Online January 30, 2018.
Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.118.300699/-/DC1.

[1] These authors contributed equally to this work.

[2] Corresponding author: Department of Integrative Biology, The University of Texas at Austin, #1 University Station - C0930, Austin, TX 78712. E-mail: wilke@austin.utexas.edu

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, 2016; Shahmoradi *et al.* 2014).

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). At the same time, all-atom models remain limited in important aspects. They generally do not accurately estimate the entropic contribution of the unfolded protein ensemble, and they tend to become increasingly inaccurate as more changes are introduced into the original protein sequence.

In addition to any biophysical limitations exhibited by all-atom models, there is another critical aspect by which protein design methods deviate 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, theoretically can create sequence configurations that are thermodynamically allowed but not easily accessible by evolution (Huang *et al.* 2016). In fact, a number of recent studies have begun to examine how the state of the background sequence affects a the ability of a protein 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 the propensity of a position to tolerate further mutations. Therefore, the evolutionary history of a protein may influence how mutational space is sampled, and the order in which states were sampled, and the length of time any specific state has been resident, may also influence 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.

Methods

Protein structures and alignments

We analyzed a set of 38 proteins 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. Supplemental Material, Table S1 in File 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 via both protein design and protein evolution. In all cases, we first minimized the structures with Rosetta (Leaver-Fay *et al.* 2011). This minimization ensures that the energy difference observed upon subsequent mutation can be attributed to the effect of the mutation and is not simply caused by improved amino-acid side-chain packing. Unless explicitly noted otherwise, we initialized both protein design and protein evolution simulations with the exact structures and amino-acid sequences corresponding to the PDB and chain identifiers listed in Table S1 in File S1.

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 ALLAA.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, nonaccelerated 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 ΔG 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 ΔG_i is given by

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

Here, β is the inverse temperature, ΔG_i is the stability of the mutant protein, and ΔG_{thresh} is the stability value at which fitness has declined to 50% of its maximum. For the majority of our simulations, we set ΔG_{thresh} to the average score obtained for the proteins designed on the same template structure. This was done to ensure 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 ΔG_{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-transformed fitness,

$$x_i = \log(f_i) = -\log\left[e^{\beta(\Delta G_i - \Delta G_{\text{thresh}})} + 1\right]. \tag{2}$$

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

$$\pi(i \to j) \approx \begin{cases} 1 & \text{for } x_j > x_i, \\ e^{-2N_e(x_i - x_j)} & \text{otherwise}, \end{cases} \tag{3}$$

where N_e is the effective population size. In all simulations, we set $N_e = 100$ and $\beta = 1$.

We used a uniform mutation model throughout, *i.e.*, every amino acid was equally likely to mutate into every other amino acid. However, we disallowed mutations to or from cysteines, because cysteine disulfide bonds are not properly handled by the Rosetta energy function. We ran each simulation until 5000 substitutions had occurred. Simulations were run for a fixed number of substitutions to ensure similar amounts of divergence from the each of the starting templates, allowing for a fair comparison to the designed sequences. However, we note that this choice is equivalent to selecting a random substitution rather than selecting a sequence at a random time point, and it enriches for genotypes with high substitution rates.

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

$$H_i = -\sum_j p_{ij} \ln p_{ij}, \tag{4}$$

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,

$$n_{\text{eff}} = \exp(H_i). \tag{5}$$

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

$$D_i^{\text{KL}} = \sum_j p_{ij} \ln(p_{ij}/q_{ij}). \tag{6}$$

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 alignments 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 the buriedness or exposure of a residue, we calculated its 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 surface exposed to solvent and are generally closer to the protein surface, while residues with lower RSA are closer to protein 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 WCN, defined as

$$\text{WCN}_i = \sum_{j \neq i} 1/r_{ij}^2, \tag{7}$$

where i 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 (iWCN = 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_α 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 in a git repository at: https://github.com/reductase4/evol_sim_vs_rosetta.git An archive of this repository has been deposited with Zenodo, and is available at DOI 10.5281/zenodo.1160646.

Results

To examine how evolutionary history affects the emergence 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 in File 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 S1 in File S1).

Amino-acid distributions

We first compared the overall amino-acid frequencies between natural and simulated sequences (Figure S2 in File 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. We also saw a moderate excess of arginines. We note that both are amino acids with a large number of rotamers, which the design algorithm biases toward. Importantly, there were only minor differences in amino-acid frequencies between evolved sequences and natural sequences. Because our evolutionary simulation made all mutations to different amino acids equally likely, we might have expected that amino acids encoded by four- or sixfold degenerate codon families would have been under-represented in comparison to their frequencies in natural alignments, which evolve at the DNA level, and, hence, can be affected by codon degeneracy. Though these amino acids are not under-represented, neither in the evolutionary simulations nor in protein design, this suggests that the energetic interactions between amino acids at different sites impose stronger constraints on amino-acid frequencies than does mutation bias. In summary, while evolutionary simulations outperformed protein design specifically for lysine and arginine, overall the differences between both approaches and natural sequences were minor.

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 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 1). Two exceptions (indicated as outliers in the middle boxplot of Figure 1) 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_{eff} (Equation 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 2), whereas designed sequences had significantly less variability compared to natural sequences (paired t- test, $p < 2.2 \times 10^{-16}$, Figure 2). 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 S1 in File S1). This seemingly paradoxical

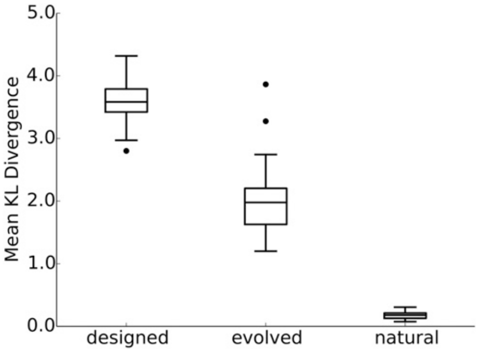

Figure 1 Mean KL divergence of designed, evolved, and natural sequences to natural sequences. Natural sequences were compared to themselves by randomly splitting alignments into two groups and calculating the KL divergence between them. A lower KL divergence indicates that the amino acid distributions at individual sites are closer to that of natural proteins.

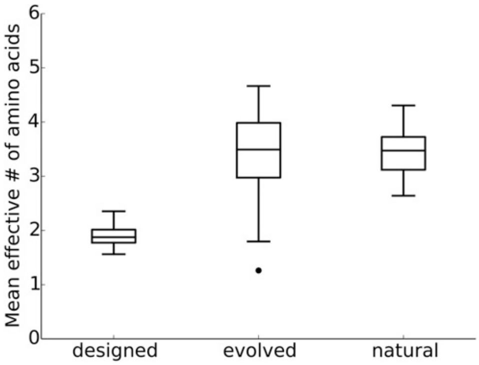

Figure 2 Mean effective number of amino acids for designed, evolved, and natural sequences. Evolved and natural sequences had comparable mean effective numbers of amino acids (paired *t*-test, $p = 0.68$), whereas designed sequences had significantly lower mean effective numbers (paired *t*- test, $p < 2.2 \times 10^{-16}$). 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.

observation implies that designed sequences, relative to the evolved sequences, experienced changes at a larger number of sites but toward a smaller set of different amino acids.

Just because two alignments have a comparable mean n_{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_{eff} among alignments generated by different methods (designed sequences, evolved sequences, and 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 3). The former were all positive and ranged between ~ 0.2 and 0.7, whereas the latter did not exceed ~ 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.

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; Chang *et al.* 2013; Yeh *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 RSA (Figure 4). Correlations between n_{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 4). Results were similar when we used inverse Weighted Contact Number (iWCN) instead of RSA (Figure S3, D–F in File S1).

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 $\Delta G_{\mathrm{thresh}}$ 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 S4 in File S1). 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 S5 in File S1). Yet the observed pattern of RSA–n_{eff} correlations was virtually unchanged from that in the original dataset (Figure S6A in File S1). 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 structures were indeed much less stable than before (Figure S7 in File S1). Yet again, there was virtually no change in the observed pattern of RSA–n_{eff} correlations (Figure S6B in File S1).

Evolving designed sequences

To recap, we have found that, by any metric considered, evolved sequence alignments looked much more realistic than

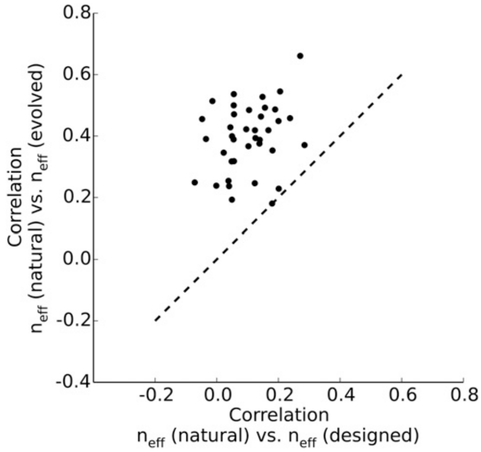

Figure 3 Spearman correlations between n_{eff} for natural and either evolved or designed sequences. Each dot represents the correlation coefficients for one protein. Correlations are substantially higher when comparing evolved to natural sequences than when comparing designed to natural sequences (paired *t*-test, $p < 2.2 \times 10^{-16}$).

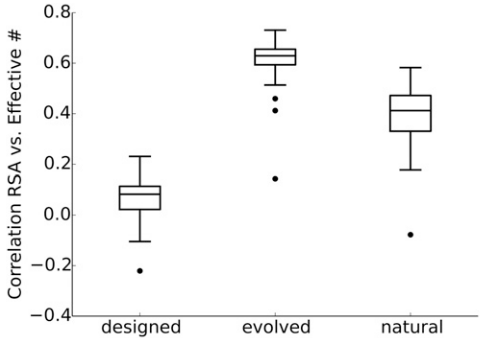

Figure 4 Spearman correlations between RSA and n_{eff} for designed, evolved, and natural sequences. Correlations for designed sequences were significantly lower than for natural sequences (paired *t*-test, $p = 1.5 \times 10^{-14}$). By contrast, correlations for evolved sequences were significantly *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 S3, A–C in File S1.

designed 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 in File 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 S8 in File S1).

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 5). Sequences that were evolved from design had a lower KL divergence than designed sequences (Figure 5A), a higher mean effective number of amino acids (Figure 5B), higher correlations of site-variability patterns to natural alignments (Figure 5C), and higher correlations between n_{eff} and RSA (Figure 5D). However, by all these metrics, sequences evolved from design were intermediate between designed sequences and sequences evolved from a natural sequence. These intermediate metrics reflect that the designed sequences have a site-wise preference for a smaller set of amino acids than do the evolved sequences (Figure 2), and this preference is only partly undone by subsequent evolution. While evolution could expand the set of acceptable amino acids at some sites, other sites remained frozen in the narrow area of sequences space supplied by design, and might require much longer evolutionary simulations to become unfrozen.

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 S9A in File S1). However, the n_{eff} from evolved sequences performed similarly to RSA (Figure S9B in File S1), and n_{eff} from evolved from design sequences displayed intermediate performance (Figure S9C in File S1). Results were similar for the iWCN, except that correlations between n_{eff} and iWCN tended to be somewhat stronger than they were between n_{eff} and RSA (Figure S9, D–F in File S1), consistent with a large body of prior work finding that WCN tends to correlate more strongly with evolutionary variation than does RSA (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 (Jack *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

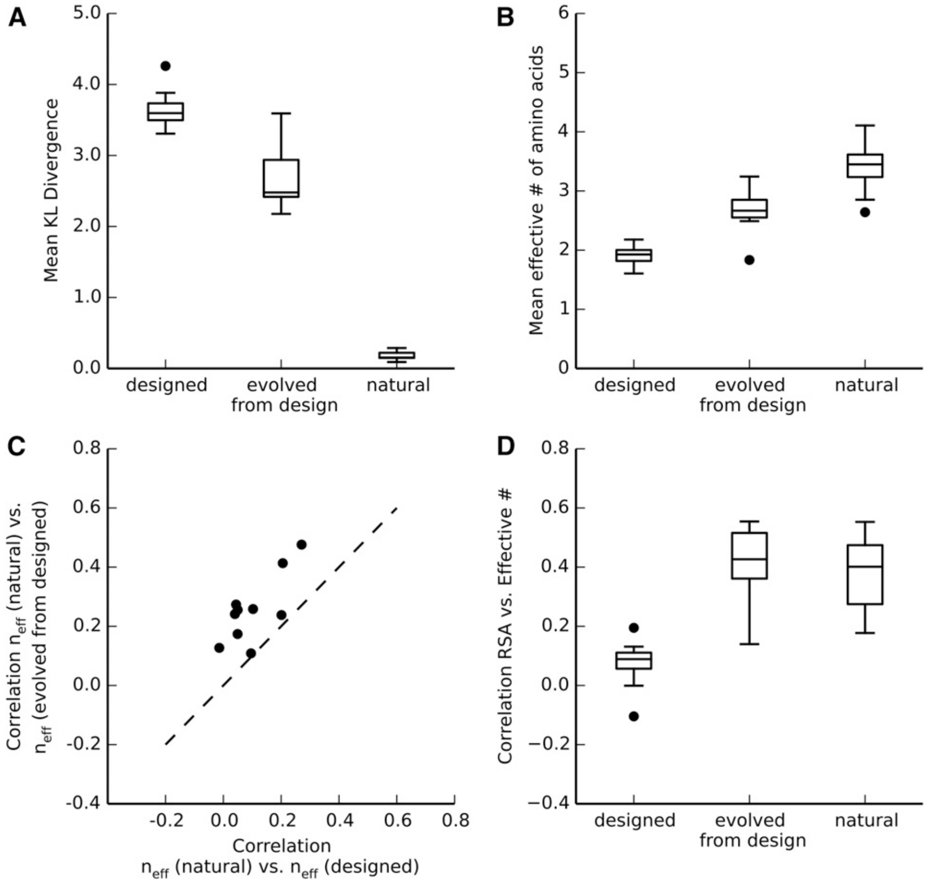

Figure 5 Comparison between designed sequences and sequences evolved from design. (A) Mean KL divergence of designed, evolved from design, and natural sequences. The KL divergence was lower for evolved from design sequences than for designed sequences (paired *t*-test, $p = 2.57 \times 10^{-6}$). (B) Mean effective number of amino acids for designed, evolved from design, and natural sequences. The mean effective number (n_{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_{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_{eff} from natural sequences were significantly higher for n_{eff} from evolved from design sequences than for n_{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, $p = 0.62$).

traditional protein-design algorithm implemented in RosettaDesign (Kuhlman *et al.* 2003). 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). Both methods impose the same thermodynamic constraints on the simulated sequences, but they impose different constraints in their sampling of sequence space. For both methods, we compared the 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 side-chain replacements are made with the goal of maximizing 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.

Protein design is frequently billed as being able to explore a much larger sequence space than does evolution, because the design process starts from scratch with a random sequence, and, thus, in theory should be able to reach any place in sequence space, whereas evolution can only explore near neighbors to already existing sequences (Huang *et al.* 2016). However, we did not find this to be the case in our simulations. Specifically, even though our designed sequences had a higher mean sequence divergence than did evolved or natural sequences (Figure S1 in File S1), designed sequences had, on average, smaller effective numbers of amino acids at each site (Figure 2). Thus, on average, individual sites were less variable under design than they were under evolution, natural or simulated. We were able to explain this unexpected result by investigating how exactly the initial sequences are chosen under design. The initial sampling of side chains is done from a rotamer library (Kaufmann *et al.* 2010), rather than uniformly from the 20 amino acids. Consequently, amino acids with many rotamers are over-represented in the initially chosen sequence. The subsequent optimization algorithm is constrained by the biased pool of initial sequences, and cannot undo the unequal sampling. We can see the remnants of the oversampling of certain amino acids in the excess of lysine and arginine in the overall amino-acid frequencies (Figure S2 in File S1), as both are amino acids with a very large number of rotamers. When we subjected the designed sequences to further evolution, a wider range of substitutions became allowed. Therefore, we saw the evolved-from-design sequences approaching site variabilities similar to those displayed by the original evolved and natural sequences (Figure 5).

Our findings are consistent with recent works on evolutionary entrenchment (Pollock *et al.* 2012; McCandlish *et al.* 2015, 2016; Shah *et al.* 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. Here, entrenchment was visible in particular for designed and evolved-from-design sequences, which were highly and somewhat biased by the initial sampling process of protein design.

Even though our evolutionary simulations considered both structural constraints and a realistic, sequential sampling of sequence space, 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 the simplistic assumption of our model that selection acts exclusively on protein stability. In natural organisms, numerous selection pressures beyond just protein stability act on protein sequences (Thorne 2007; 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; Ollikainen and Kortemme 2013). However, since we used the same fixed backbone model for both protein design and protein evolution, we do not 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, but we expect this reordering to be of little consequence for the metrics considered in this work (Teufel and Wilke 2017). Third, 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. Finally, it is possible that a modified design algorithm that samples initial sequences from a uniform distribution of amino acids rather than from a distribution of rotamers would produce sequence alignments on par with those we generated here under simulated evolution. Testing this hypothesis, however, falls beyond the scope of the present paper.

Acknowledgments

This work was supported by National Institutes of Health grants R01 GM088344 and National Science Foundation (NSF) Cooperative agreement no. DBI-0939454 (BEACON Center). Computational resources were provided by The University of Texas at Austin∈s Center for Computational Biology and Bioinformatics 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

Arenas, M., A. Sánchez-Cobos, and U. Bastolla, 2015 Maximum-likelihood phylogenetic inference with selection on protein folding stability. Mol. Biol. Evol. 32: 2195–2207.

Bastolla, U., M. Porto, H. E. Roman, and M. Vendruscolo, 2007 The structurally constrained neutral model of protein evolution, pp. 75–112 in *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations*. Springer, Berlin.

Chang, C.-M., Y.-W. Huang, C.-H. Shih, and J.-K. Hwang, 2013 On the relationship between the sequence conservation and the packing density profiles of the protein complexes. Proteins 81: 1192–1199.

Chaudhury, S., S. Lyskov, and J. J. Gray, 2010 Pyrosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics 26: 689–691.

Chen, P., and E. I. Shakhnovich, 2009 Lethal mutagenesis in viruses and bacteria. Genetics 183: 639–650.

Chi, P. B., and D. A. Liberles, 2016 Selection on protein structure, interaction, and sequence. Protein Sci. 25: 1168–1178.

Echave, J., and C. O. Wilke, 2017 Biophysical models of protein evolution: understanding the patterns of evolutionary sequence divergence. Annu. Rev. Biophys. 46: 85–103.

Echave, J., S. J. Spielman, and C. O. Wilke, 2016 Causes of evolutionary rate variation among protein sites. Nat. Rev. Genet. 17: 109–121.

Fleishman, S. J., T. A. Whitehead, D. C. Ekiert, C. Dreyfus, J. E. Corn *et al.*, 2011 Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science 332: 816–821.

Franzosa, E. A., and Y. Xia, 2009 Structural determinants of protein evolution are context-sensitive at the residue level. Mol. Biol. Evol. 26: 2387–2395.

Goldstein, R. A., and D. D. Pollock, 2016 The tangled bank of amino acids. Protein Sci. 25: 1354–1362.

Halle, B., 2002 Flexibility and packing in proteins. Proc. Natl. Acad. Sci. USA 99: 1274–1279.

Huang, P.-S., S. E. Boyken, and D. Baker, 2016 The coming of age of de novo protein design. Nature 537: 320–327.

Huang, T.-T., M. L. del Valle Marcos, J.-K. Hwang, and J. Echave, 2014 A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. BMC Evol. Biol. 14: 78.

Jack, B. R., A. G. Meyer, J. Echave, and C. O. Wilke, 2016 Functional sites induce long-range evolutionary constraints in enzymes. PLoS Biol. 14: e1002452.

Jackson, E. L., N. Ollikainen, A. W. Covert, III, T. Kortemme, and C. O. Wilke, 2013 Amino-acid site variability among natural and designed proteins. PeerJ 1: e211.

Jackson, E. L., A. Shahmoradi, S. J. Spielman, B. R. Jack, and C. O. Wilke, 2016 Intermediate divergence levels maximize the strength of structure–sequence correlations in enzymes and viral proteins. Protein Sci. 25: 1341–1353.

Jones, S., and J. M. Thornton, 1996 Principles of protein–protein interactions. Proc. Natl. Acad. Sci. USA 93: 13–20.

Kabsch, W., and C. Sander, 1983 Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577–2637.

Kachroo, A. H., J. M. Laurent, C. M. Yellman, A. G. Meyer, C. O. Wilke *et al.*, 2015 Systematic humanization of yeast genes reveals conserved functions and genetic modularity. Science 348: 921–925.

Kaufmann, K. W., G. H. Lemmon, S. L. DeLuca, J. H. Sheehan, and J. Meiler, 2010 Practically useful: what the Rosetta protein modeling suite can do for you. Biochemistry 49: 2987–2998.

Kimura, M., and T. Ohta, 1974 On some principles governing molecular evolution. Proc. Natl. Acad. Sci. USA 71: 2848–2852.

Kuhlman, B., G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard *et al.*, 2003 Design of a novel globular protein fold with atomic-level accuracy. Science 302: 1364–1368.

Leaver-Fay, A., M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson *et al.*, 2011 ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 487: 545–574.

Liberles, D. A., S. A. Teichmann, I. Bahar, U. Bastolla, J. Bloom *et al.*, 2012 The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci. 21: 769–785.

Marcos, M. L., and J. Echave, 2015 Too packed to change: side-chain packing and site-specific substitution rates in protein evolution. PeerJ 3: e911.

McCandlish, D. M., J. Otwinowski, and J. B. Plotkin, 2015 Detecting epistasis from an ensemble of adapting populations. Evolution 69: 2359–2370.

McCandlish, D. M., P. Shah, and J. B. Plotkin, 2016 Epistasis and the dynamics of reversion in molecular evolution. Genetics 203: 1335–1351.

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.

Ollikainen, N., and T. Kortemme, 2013 Computational protein design quantifies structural constraints on amino acid covariation. PLOS Comput. Biol. 9: e1003313.

Overington, J., D. Donnelly, M. S. Johnson, A. Šali, and T. L. Blundell, 1992 Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. Protein Sci. 1: 216–226.

Pollock, D. D., G. Thiltgen, and R. A. Goldstein, 2012 Amino acid coevolution induces an evolutionary stokes shift. Proc. Natl. Acad. Sci. USA 109: E1352–E1359.

Ramsey, D. C., M. P. Scherrer, T. Zhou, and C. O. Wilke, 2011 The relationship between relative solvent accessibility and evolutionary rate in protein evolution. Genetics 188: 479–488.

Röthlisberger, D., O. Khersonsky, A. M. Wollacott, L. Jiang, J. DeChancie *et al.*, 2008 Kemp elimination catalysts by computational enzyme design. Nature 453: 190–195.

Serohijos, A. W., and E. I. Shakhnovich, 2013 Contribution of selection for protein folding stability in shaping the patterns of polymorphisms in coding regions. Mol. Biol. Evol. 31: 165–176.

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.

Serohijos, A. W., Z. Rimas, and E. I. Shakhnovich, 2012 Protein biophysics explains why highly abundant proteins evolve slowly. Cell Rep. 2: 249–256.

Shah, P., D. M. McCandlish, and J. B. Plotkin, 2015 Contingency and entrenchment in protein evolution under purifying selection. Proc. Natl. Acad. Sci. USA 112: E3226–E3235.

Shahmoradi, A., and C. O. Wilke, 2016 Dissecting the roles of local packing density and longer-range effects in protein sequence evolution. Proteins 84: 841–854.

Shahmoradi, A., D. K. Sydykova, S. J. Spielman, E. L. Jackson, E. T. Dawson *et al.*, 2014 Predicting evolutionary site variability from structure in viral proteins: buriedness, packing, flexibility, and design. J. Mol. Evol. 79: 130–142.

Teufel, A. I., and C. O. Wilke, 2017 Accelerated simulation of evolutionary trajectories in origin-fixation models. J. R. Soc. Interface 14: 20160906.

Teufel, A. I., J. A. Grahnen, and D. A. Liberles, 2012 Modeling proteins at the interface of structure, evolution, and population genetics, pp. 347–361 in *Computational Modeling of Biological Systems*. Springer, Berlin.

Thorne, J. L., 2007 Protein evolution constraints and model-based techniques to study them. Curr. Opin. Struct. Biol. 17: 337–341.

Tien, M. Z., A. G. Meyer, D. K. Sydykova, S. J. Spielman, and C. O. Wilke, 2013 Maximum allowed solvent accessibilities of residues in proteins. PLoS One 8: e80635.

Tóth-Petróczy, Á., and D. S. Tawfik, 2011 Slow protein evolutionary rates are dictated by surface–core association. Proc. Natl. Acad. Sci. USA 108: 11151–11156.

Wylie, C. S., and E. I. Shakhnovich, 2011 A biophysical protein folding model accounts for most mutational fitness effects in viruses. Proc. Natl. Acad. Sci. USA 108: 9916–9921.

Yeh, S.-W., J.-W. Liu, S.-H. Yu, C.-H. Shih, J.-K. Hwang *et al.*, 2013 Site-specific structural constraints on protein sequence evolutionary divergence: local packing density *vs.* solvent exposure. Mol. Biol. Evol. 31: 135–139.

Yeh, S.-W., T.-T. Huang, J.-W. Liu, S.-H. Yu, C.-H. Shih *et al.*, 2014 Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. BioMed Res. Int. 2014: 572409.

Zhou, H., and Y. Zhou, 2004 Quantifying the effect of burial of amino acid residues on protein stability. Proteins 54: 315–322.

Communicating editor: J. Hermisson