A thermodynamic model of protein structure evolution explains empirical amino acid substitution matrices

Supplementary information
Supplementary methods

Estimating transition/transversion bias in E. coli

Lee et al.\textsuperscript{1} reported the counting data of spontaneous nucleotide mutations in \textit{E. coli}. The data in their table 3 is the mutation flux. The mutation flux is a function of the instantaneous mutation rates and the equilibrium frequencies: $\Phi_{ij} = q_{ij} \cdot \pi_{ij}$. In accordance with a K80-type Markov model\textsuperscript{2}, we assume that the equilibrium frequency of the nucleotides is 0.25 for all and calculate the transition/transversion rate ratio ($\kappa$):

$$\kappa = \frac{49 + 82}{17 + 38 + 30 + 17} \cdot \frac{4}{2} = 2.57$$

The protein misfolding-fitness function behaves as the sigmoidal stability-fitness function with a modified offset

Another fitness model was proposed by Drummond and Wilke\textsuperscript{3}. They found that a major selection pressure is selection against cytotoxicity caused by protein misfolding. In their model the fraction of unfolded protein has an exponentially decreasing effect on fitness, which depends on a toxicity parameter $c$ and the protein abundance parameter $A$.

$$w_{\text{misfolding}} = \exp \left( \frac{cA}{1 + \exp \left( \frac{s}{s \Delta G} \right)} \right)$$

When $s\Delta G$ is low ($s\Delta G < 3$) the misfolding fitness function can be simplified to the fraction folded fitness function although offset by $\log(cA)$:
\[
\text{misfolding} = \exp\left(\frac{cA}{1 + \exp\left(\frac{s}{G}\right)}\right)
\]
\[
\text{misfolding} = \exp\left(\frac{cA}{\exp\left(\frac{s}{G}\right)}\right)
\]
\[
\text{misfolding} = \exp\left( cA\exp\left(\frac{s}{G}\right)\right)
\]
\[
\text{misfolding} = \frac{1}{\exp\left(cA\exp\left(\frac{s}{G}\right)\right)}
\]
\[
\text{misfolding} = \frac{1}{1 + cA\exp\left(\frac{s}{G}\right)}
\]
\[
\text{misfolding} = \frac{1}{1 + \exp\left(\frac{s}{G + \log(cA)}\right)}
\]
\[
\text{misfolding} = \frac{1}{\frac{1}{\text{folding}\left(\frac{s}{G + \log(cA)}\right)}}
\]

Where \(\omega_{\text{folding}}\) is the fitness function used throughout the paper:

\[
\text{folding}\left(x\right) = \frac{1}{1 + \exp\left(x\right)}
\]

**Estimation of parameter uncertainty by bootstrap analysis**

To estimate the uncertainty of the TSM parameters we performed a bootstrap analysis. Specifically, we compute the phylogenetic log-likelihood (IQ-TREE) for each parameter combination explored in Fig. S1 for each of the 52 sequence alignments individually. Next, using random sampling with replacement (sample size = 52) over the alignment dataset, we took 10,000 bootstrap samples, finding the argmax parameter set for each:

\[
\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \log L\left(MSA_i|Q(\theta, \Delta G)\right)
\]

From these samples, we computed the standard deviation of each parameter.
**Q-matrix estimation error upon noise injection**

Deviations between inferred rates and rates predicted with the TSM-model could originate from ΔΔG prediction errors. The magnitudes of ΔΔG prediction errors were estimated from the correlation between predicted ΔΔG values from Rosetta and known experimental values for the same 52 proteins described in the main text (σ_{err}=1.7 kcal/mol based on 590 measurements)^4. To estimate the impact of such errors, we predicted rates with the TSM model as described in the main text but modified the ΔΔG values by adding errors sampled from the empirical error distribution. The expected correlation between Q-matrix parameters with and without sampled error is r_{t+R^2}=0.79. Thus, ΔΔG prediction error alone likely contributes at least 21% of the variance in the LG matrix that is unexplained by our model. We note that the true error could be greater, as half of the substitutions in the experimental ΔΔG dataset are to alanine, and alanine substitutions are more accurately predicted than other mutations (σ_{err,X→A}=1.4 vs σ_{err,X→!A}=1.9 kcal/mol).
Supplementary figures

Fig. S1. Maximum-likelihood optimization of parameters yields similar amino substitution parameters as direct optimization against the LG matrix. Both amino acid frequencies (A) and exchangeabilities (B) are similar regardless of the optimization method. In direct LG optimization, we select the parameter combination $\theta = \{\Delta G_{nat}, N_e, \kappa, \rho\}$ that maximize the free-parameter weighted correlation to the LG matrix: $r_{\text{weighted}}^2 = (19r_\pi^2 + 189r_{\log R}^2)/208$. For description of how parameters for TMS are optimized see main text.
**Fig. S2.** Coupling and sensitivity-analysis of optimal parameter set for the TSM-model. The likelihood surface is plotted on a grid for pairwise combinations of parameters. Higher likelihood corresponds to darker shades of blue. The strong diagonal likelihood ridge for $\Delta G_{\text{nat}}$ vs $N_e$ indicates a strong correlation between those two parameters, suggesting one underlying factor rather than two independent parameters.
Fig. S3. Amino acid substitution parameters are not affected by coupled variations in $\Delta G_{nat}$ and $N_e$. As described in the main text and shown in figure S1 with respect to log(L), $\Delta G_{nat}$ and log($N_e$) are coupled parameters under a simple stability-fitness model under equilibrium. To understand whether variations of {$\Delta G_{nat}, N_{eff}$} along the max(logL) axis affect amino acid frequencies and exchangeabilities, we selected two parameter sets from our grid with different population sizes. Neither (A) amino acid frequencies nor (B) exchangeabilities are substantially affected.
Fig. S4. The TMS model recapitulates the mean substitution behavior between amino acids.

A) Comparison of the amino acid exchangeability matrix predicted by TSM (black circles) and the WAG global exchangeability matrix (red circles). Inset, correlation between exchangeabilities from TSM (x-axis) vs WAG (y-axis). B) Correlation between amino acid equilibrium probabilities predicted by TSM (x-axis) and values from WAG (y-axis). C) The same as A, but without selection. D) the same as B but without selection.
Fig. S5. Heatmap of correlations between stationary amino acid frequencies for common substitution matrices. Squared Pearson correlations ($r^2$) were computed over the 20 equilibrium frequency parameters. Matrices were clustered based on $r^2$ using hierarchically clustering. 
Fig. S6. Proteins used for calculating substitution energetics are representative of the alignments used to infer LG. A) Comparison of the amino acid exchangeability matrix (LG*) inferred from the 52 alignments in our dataset using the LG methodology (red circles) and the original LG matrix (black circles). Inset, correlation between exchangeabilities. B) Correlation between amino acid equilibrium probabilities inferred from LG* and values from LG.

Fig. S7. Global patterns of amino acid substitutions are affected by %GC content bias. Starting from the optimal parameters of the TMS matrix, we expanded the model by replacing the
K80 nucleotide proposal model with a T92 model. The figure shows that the K80 assumption of no GC-bias (i.e., %GC=0.5) results in exchangeabilities and stationary frequencies that correlate optimally with the LG matrix parameters. Other values for GC-bias correlate less well.

**Table S1**: Log-likelihoods of trees computed for MSAs in the LG Pfam test dataset. The table shows likelihoods for the MSAs with a delta log likelihood of less than 5. Out of 500 MSAs, our Q-matrix achieved a higher likelihood than LG for 8.

| aln          | TSM logL | LG logL | ΔlogL |
|--------------|----------|---------|-------|
| Aln2393.txt-gbiphyml | -577.403 | -590.073 | -12.670 |
| Aln6612.txt-gbiphyml | -2493.606 | -2500.485 | -6.879 |
| Aln3459.txt-gbiphyml | -3494.576 | -3499.601 | -5.025 |
| Aln3653.txt-gbiphyml | -1272.407 | -1276.997 | -4.590 |
| Aln5878.txt-gbiphyml | -950.257 | -954.827 | -4.570 |
| Aln1284.txt-gbiphyml | -634.963 | -638.874 | -3.911 |
| Aln3227.txt-gbiphyml | -756.610 | -758.864 | -2.254 |
| Aln3035.txt-gbiphyml | -978.598 | -980.501 | -1.903 |
| Aln4281.txt-gbiphyml | -754.795 | -753.672 | 1.123 |
| Aln6528.txt-gbiphyml | -490.274 | -489.116 | 1.158 |
| Aln5555.txt-gbiphyml | -392.622 | -390.080 | 2.542 |
| Aln2167.txt-gbiphyml | -798.087 | -795.381 | 2.706 |
| Aln1980.txt-gbiphyml | -1019.706 | -1016.945 | 2.761 |
| Aln2041.txt-gbiphyml | -1754.531 | -1751.638 | 2.893 |
| Aln4313.txt-gbiphyml | -1345.046 | -1341.611 | 3.435 |
| Aln4019.txt-gbiphyml | -635.864 | -632.414 | 3.450 |
| Aln1245.txt-gbiphyml | -850.857 | -847.353 | 3.504 |
| Aln0070.txt-gbiphyml | -433.801 | -429.861 | 3.940 |
| Aln4914.txt-gbiphyml | -847.560 | -843.521 | 4.039 |
| Aln2179.txt-gbiphyml | -546.731 | -542.558 | 4.173 |
| Aln4871.txt-gbiphyml | -389.268 | -385.089 | 4.179 |
| Aln2807.txt-gbiphyml | -1758.808 | -1754.351 | 4.457 |
| Aln5867.txt-gbiphyml | -696.083 | -691.502 | 4.581 |
| aln            | TSM logL | LG logL | ΔlogL |
|---------------|----------|---------|-------|
| Aln4013.txt-gb.phyml | -858.934 | -854.133 | 4.801 |
References

1. Lee H, Popodi E, Tang H, Foster PL (2012) Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. Proc. Natl. Acad. Sci. 109:2774–2783.

2. Kimura M (1980) A Simple Method for Estimating Evolutionary Rates of Base Substitutions Through Comparative Studies of Nucleotide Sequences. J. Mol. Evol. 16:111–120.

3. Drummond DA, Wilke CO (2008) Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. Cell 134:341–352.

4. Park H, Bradley P, Greisen P, Liu Y, Mulligan VK, Kim DE, Baker D, DiMaio F (2016) Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. J. Chem. Theory Comput. 12:6201–6212.

5. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17:261–272.