PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Atomistic simulation of protein evolution reveals sequence covariation and time-dependent fluctuations of site-specific substitution rates

Christoffer Norn, Ingemar André*

Biochemistry and Structural Biology, Lund University, Lund, Sweden

* ingemar.andre@biochemistry.lu.se

OPEN ACCESS

Citation: Norn C, André I (2023) Atomistic simulation of protein evolution reveals sequence covariation and time-dependent fluctuations of site-specific substitution rates. PLoS Comput Biol 19(3): e1010262. https://doi.org/10.1371/journal.pcbi.1010262

Editor: Roger Dimitri Kouyos, University of Zurich, SWITZERLAND

Received: May 31, 2022

Accepted: March 1, 2023

Published: March 24, 2023

Data Availability Statement: RosettaEvolve is available through a release of Rosetta, which can be downloaded at rosettacommons.org. Additional scripts and running information can be found at https://github.com/Andre-lab/RosettaEvolve.

Funding: This work was funded by a grant from the Swedish research council (grant number 2015-04203) to IA. The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC), partially funded by the Swedish Research Council through

Abstract

Thermodynamic stability is a crucial fitness constraint in protein evolution and is a central factor in shaping the sequence landscapes of proteins. The correlation between stability and molecular fitness depends on the mechanism that relates the biophysical property with biological function. In the simplest case, stability and fitness are related by the amount of folded protein. However, when proteins are toxic in the unfolded state, the fitness function shifts, resulting in higher stability under mutation-selection balance. Likewise, a higher population size results in a similar change in protein stability, as it magnifies the effect of the selection pressure in evolutionary dynamics. This study investigates how such factors affect the evolution of protein stability, site-specific mutation rates, and residue-residue covariation. To simulate evolutionary trajectories with realistic modeling of protein energetics, we develop an all-atom simulator of protein evolution, RosettaEvolve. By evolving proteins under different fitness functions, we can study how the fitness function affects the distribution of proposed and accepted mutations, site-specific rates, and the prevalence of correlated amino acid substitutions. We demonstrate that fitness pressure affects the proposal distribution of mutational effects, that changes in stability can largely explain variations in site-specific substitution rates in evolutionary trajectories, and that increased fitness pressure results in a stronger covariation signal. Our results give mechanistic insight into the evolutionary consequences of variation in protein stability and provide a basis to rationalize the strong covariation signal observed in natural sequence alignments.

Author summary

Modern-day proteins are the result of the process of evolution. The fate of random substitutions at the nucleotide level is dependent on the fitness of the new gene variant. One of the strongest fitness pressures shaping the sequences of protein is thermodynamic stability; proteins must typically be stable to carry out its function and misfolded proteins can be toxic. To understand the importance of thermodynamic stability in protein evolution and to what extent it can explain natural sequence variation we have developed a method

grant agreement no. 2020:5-308 to IA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

for simulating protein evolution using a three-dimensional structure and structure-based stability calculations. In the simulations, the strength of selection can be varied, and complete phylogenetic trees of a protein family can be generated. Using these simulations, we demonstrate how mutation rates at individual sites in a protein are coupled to the overall stability of the protein, and how the spectrum of accepted mutations is shaped by stability, and how strong interactions between residues in a protein can result in sequence covariation.

Introduction

The sequences of natural proteins result from evolutionary processes guided by various fitness constraints. Sequence variation can impact expression levels, functions like binding and catalysis, protein aggregation, and protein stability. There is ample evidence to suggest that the stability fitness constraint is an important factor controlling the evolution of protein sequences. Proteins are marginally stable, with folding free energies of 5–10 kcal/mol [1], meaning that single amino acid mutations can have a significant impact on thermodynamic stability. The impact of thermodynamic stability on molecular fitness depends on the mechanism that couples stability with fitness. Destabilization can lead to a reduction of the concentration of actively folded protein, modifying functions associated with that protein as well as increasing the cost of protein synthesis. It can also have an indirect effect, increasing the concentration of unfolded proteins. The presence of unfolded protein induces a response of the protein quality control system, which may be associated with fitness costs [2]. If destabilization leads to the presence of cytotoxic misfolded species, that can have a direct impact on organismal fitness. Geiler-Samerotte et al. have shown that overexpression of destabilized proteins leads to deleterious fitness effects in yeast [3]. Results like this have been taken as evidence for the misfolding avoidance hypothesis that states that highly abundant proteins evolve slower largely because of selection against toxic misfolded protein [4,5].

Simulations of protein sequence evolution trajectories with fold stability fitness functions [6] have been able to explain the marginal of proteins as a consequence of a balance between selection and mutation [7–10]. They can also demonstrate that the models can show a correlation between the abundance and evolutionary rates of proteins by accounting for the correlation between the stability of a protein (ΔG) and the effect of mutations ($\Delta\Delta G$) [8,11]. Stability-based Markov state models can also account for the variation of evolutionary rates at sites in proteins [5,12,13] and the global substitution patterns captured in amino acid substitution matrices [14].

With evolutionary dynamics simulations guided by stability or cytotoxicity fitness functions, it is possible to get detailed insights into the sequence-structure correlations within a protein or protein family and characterize the effects of selection on the fixation of gene variants [7,8,10,15–20]. The effects of random mutations on stability are well approximated by a bi-Gaussian distribution [21], enabling fast evolutionary dynamics simulations at the level of cellular populations [8,19,20]. For detailed studies of sequence-structure correlations, an energy function is required. Energy functions range from simple contact-based versions [7,17], to fully atomistic [9,16,18]. The choice depends on the questions at hand. Our interest here is to study how the mutation rates vary at individual sites in proteins, characterize the emergence of residue-residue covariation in protein sequence evolution, and evaluate the relationship between protein energetics and sequence variation. This requires the use of an atomistic energy function that can capture the detailed consequences of mutations in proteins. The

use of atomistic energy functions like FoldX [22], Rosetta [23], and ERIS [24] in evolutionary dynamics simulations come with a considerable computational cost. By assuming that the effects of mutations are additive and independent of the structure and sequence context, ΔΔG can be precomputed leading to a very large speed-up of the simulations [15,20,25]. However, to study epistatic interactions between sites in a protein [26], amino acid entrenchment [18], and temporal fluctuations in amino acid propensities at sites [17] it is necessary to update the stability of the protein as the protein sequence evolves. Because of the computational costs involved in these calculations, such simulations have rarely been attempted. One exception is Jiang et al. that simulated evolutionary trajectories with the Rosetta energy function [27] as the fitness function, in which the sequence context evolved to study sequence variation at sites in proteins [16].

To enable studies of how detailed residue-residue interactions in proteins shape the sequence landscape in evolution we extend the approach by Jiang et al. and develop an all-atom evolution simulator, RosettaEvolve (Fig 1), to simulate evolutionary trajectories using the Rosetta macromolecular modeling package [28]. Mutations are evaluated at the level of DNA to account for the structure of the genetic code. The effects of mutations are evaluated on-the-fly using structure-based ΔΔG calculations that account for sidechain flexibility and minor backbone changes [27]. A population genetic framework is used to decide the fate of mutations. Simulations are carried out as a function of evolutionary pressure to study how variations in fitness pressure impact the sequence evolution of proteins.

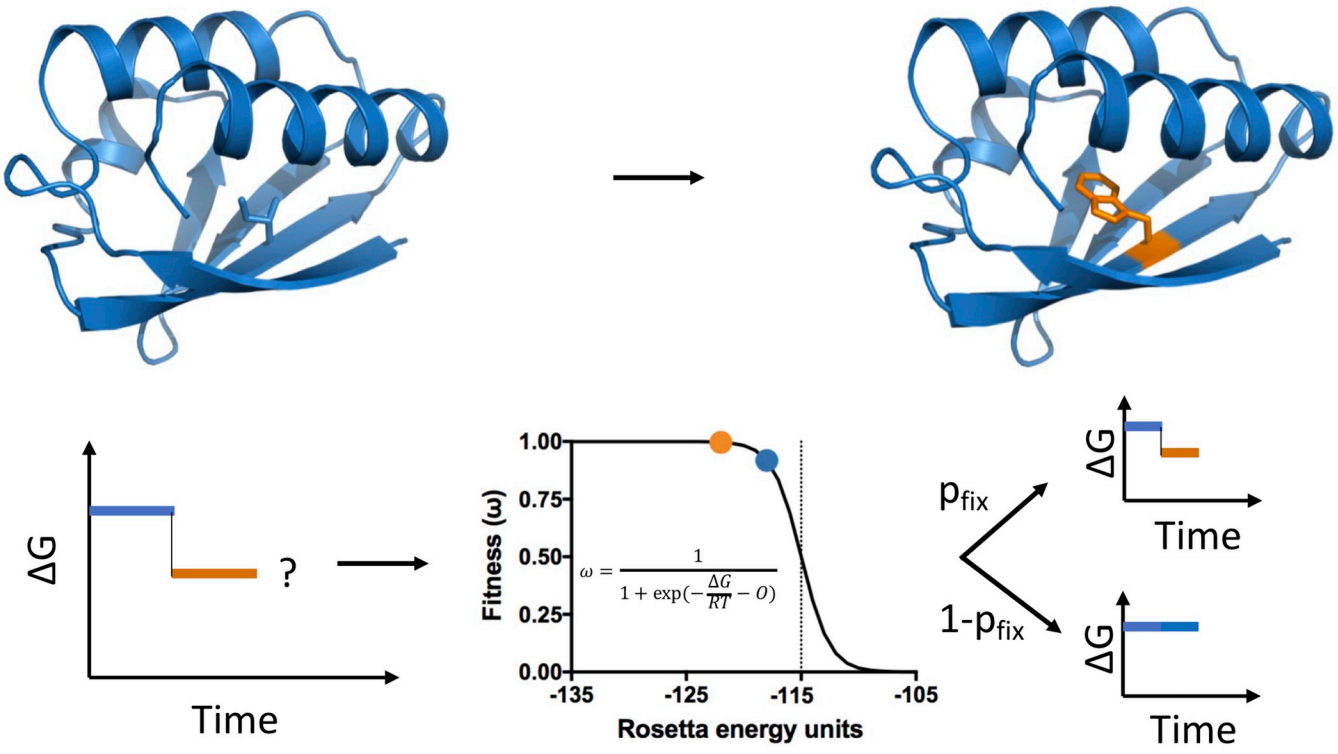

Fig 1. Simulation of evolutionary trajectories with RosettaEvolve. Mutations are proposed at the nucleotide level and nucleotide changes are translated into amino acid substitutions. The fitness of a mutation is estimated by calculating the change in stability of the protein with a ΔΔG prediction method using the Rosetta all-atom energy function. Based on the change in fitness (selection coefficient) the probability of fixating the proposed nucleotide/amino acid is evaluated.

https://doi.org/10.1371/journal.pcbi.1010262.g001

Using RosettaEvolve we demonstrate how variation in the molecular fitness parameters–such as cytotoxicity, required thermodynamic stability, and population size–affects both the proposal and fixated distribution mutational effects. We show that site-specific mutational rates fluctuate over trajectories largely dependent on fluctuations in the stability of the protein. We also show that phylogenetic trees generated by RosetteEvolve result in a robust residue-residue covariation signal which depends on selection pressure.

Results

Simulation of evolutionary trajectories with RosettaEvolve

RosettaEvolve simulates evolution at the nucleotide level (Fig 1). Differences in the chemical properties of nucleotides result in different rates for transitions and transversions [29,30]. This bias is controlled in the simulation by specifying the transition/transversion rate ratio. Multi-nucleotide or whole-codon changes are also observed in nature due to a multitude of genomic processes such as insertions, deletions, UV damage, and tandem mutations [31,32]. These nucleotide changes are captured by a multi-codon mutation rate.

To evaluate the probability that the introduced mutation will be fixated, we first have to evaluate the fitness of the mutation. Several fitness models based on protein stabilities have been described [5,6,12,13], and RosettaEvolve can easily be extended to use alternative fitness expressions. In this study, we use a fitness model that assumes that a protein's contribution to fitness is proportional to the fraction of the protein folded in its native conformation [6]. As described in Norn et al. [33], for stable proteins this is mathematically equivalent to a cytotoxicity fitness model [5], where fitness depends on the concentration of unfolded protein, but with an offset ΔG. Equating fitness to the fraction folded, the expression for fitness becomes

$$\omega_{i,folding} = \frac{1}{1 + exp(\Delta G_i / RT)} \tag{1}$$

Where $\omega_{i,folding}$ is the fitness of sequence i and $\Delta G = G_{folded} - G_{unfolded}$ is the free energy of folding. When $\Delta G <$ -3 kcal/mol, the cytotoxicity fitness model has the same mathematical form as the stability fitness function [33]

$$\omega_{i,misfolding} = \frac{1}{1 + exp\left(\frac{\Delta G_i}{RT} + \log(cA)\right)} \tag{2}$$

Where c is a toxicity parameter and A is the protein abundancy [5].

There are currently no methods that can accurately compute ΔG values with an energy function or force field. However, ΔΔG prediction methods can reach useful correlations between computed and experimental values ($r^2 = 0.56$ reported for the method used in this study [27]). The stability of a protein sequence after each mutation is evaluated

$$\Delta G_j = \Delta G_i + \Delta\Delta G_{i \to j}$$

For a trajectory started from the native sequence, we must assign a stability to the native state. This is done by subtracting an offset (E_{ref}) from the energy of the native sequence $E_{rosetta}$ so that $\Delta G = E_{rosetta} - E_{ref}$. Analogously, as seen in Eq 2, the cytotoxicity and abundance parameters offset ΔG in the fitness function. Furthermore, changes in effective population size (N) have a similar effect as offsetting ΔG, as $\Delta G \sim -\log N$[19, 33]. Hence, we can model the fitness of

a sequence as

$$\omega_i = \frac{1}{1 + exp\left(\frac{E_{rosetta,i}}{RT} - \mathrm{O}\right)} \tag{3}$$

where O is the linear offset. Setting $O = E_{ref}$ converts the fitness function into the fraction folded model in Eq 1. O is treated as a parameter in our simulations, while $E_{rosetta}$ is calculated from the structure with the Rosetta energy function. The value of O controls the offset of the fitness function (through the cytotoxicity/abundance parameters or effective population size) and anchors the computed energy on the free energy scale. Setting a low value of the offset assigns a low fitness to the native sequence, which forces the introduction of stabilizing mutations to increase the fitness of protein and decrease the energy of the protein. Conversely, setting a high value for the offset assigns a high fitness of the native sequence, which facilitates the introduction of destabilizing mutations since such mutations carry little fitness cost, leading to an increase in the energy of the protein until the mutation-selection balance is reached. By sliding the value of the offset parameter, we change the effective selection pressure. From here on, we refer to the negative of the offset O as the selection pressure.

The simulations presented in this study were carried out using a ΔΔG prediction method with limited backbone flexibility, a slight variation of the ΔΔG prediction approach in Rosetta presented by Park et. al. [27]. The method involves repacking residues that are energetically coupled to the mutated amino acid and backbone energy minimization of the focal site and the nearest neighbors in the sequence.

If populations evolve under sufficient strong selection pressure and at a sufficiently low mutation rate, the fixation probability can be estimated using Kimura's fixation probability equation [10]. For diploid organisms

$$f_{i\to j} = \frac{1 - exp(-2s_{i\to j})}{1 - exp(-4Ns_{i\to j})} \tag{4}$$

where $f_{i\to j}$ is the probability of fixation of a mutation i to j, $s_{i\to j} = \omega_j/\omega_i - 1$ is the selection coefficient, and N is the effective population size. In the simulations presented here, we set $N = 10^{4.2}$, a value we previously found to optimize correlations between computed and empirical amino acid substitution rates [19,33]. In experiments where we modeled changes to the selection pressure, we kept N constant and instead modified the offset parameter O. The simulations were carried out using the weak-mutation strong-selection model [34] and the assumption that the proteins evolve in a diploid cell and within a clonal population. We further assume that the protein is a two-state folder without competing alternative states.

In the following sections, we apply RosettaEvolve to study a specific protein, Azurin of *P. aeruginosa [35]*. Azurin is a 128-residue protein with an immunoglobulin-like fold. The protein has a copper-binding site and a single disulfide bond. Before generating evolutionary trajectories with azurin the protein was adapted to the Rosetta energy function with a structure refinement calculation.

Equilibration of trajectories

Before analyzing the dynamics of sequence evolution, the simulations must be equilibrated so that the recorded trajectory is under mutation-selection balance. The fitness equilibria shift depending on the assumed stability of the protein. In our approach, the selection pressure is controlled by the offset value. A separate equilibration is required for each selection pressure. Mutations are evaluated using a ΔΔG prediction protocol that involves structure remodeling

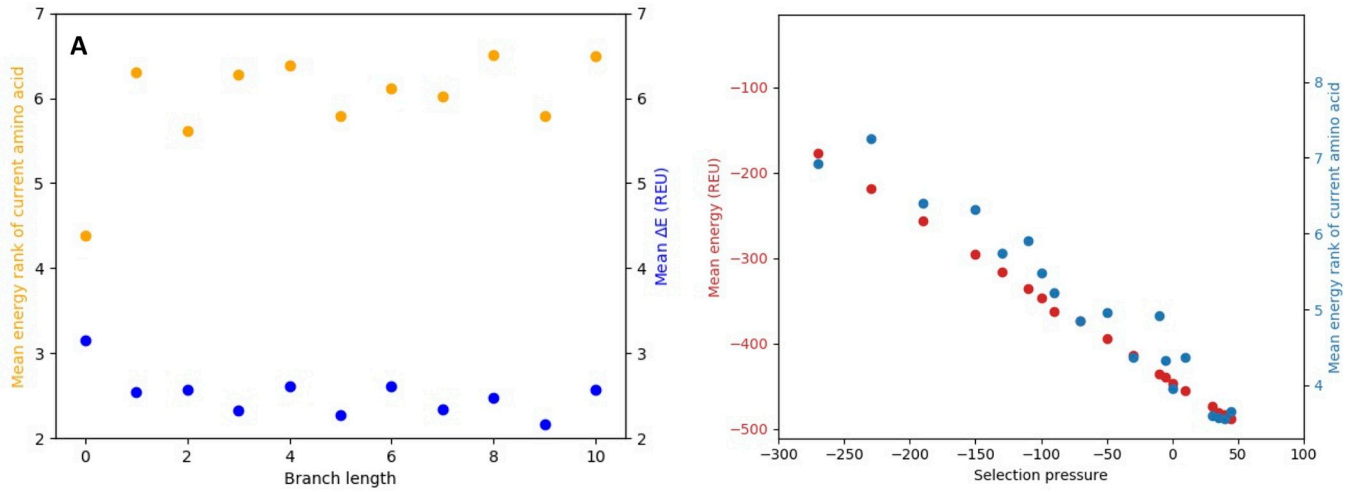

Fig 2. Equilibration of azurin. A) Mean change in energy (blue) relative to the lowest energy amino acid choice and mean energy rank (orange) as a function of branch length for a simulation with selection pressure set to -162. The standard deviation for the ΔE values is on the order of 2.5 REU, while the error for the mean of energy is around 0.2 REU. **B)** Dependence of mean energy of accepted sequences (red) and mean rank (blue) on the selection pressure (fitness function offset).

https://doi.org/10.1371/journal.pcbi.1010262.g002

and energy minimization. This means that structural changes across the trajectory and the effect of this flexibility must be equilibrated. In principle, one could return to the starting structure after each mutation, but this requires more extensive backbone sampling and leads to a far noisier energy estimation.

To follow the progress towards equilibrium, we measure the energy, mean change in energy of accepted mutations, and the average energy rank of the amino acid selected at sites. To calculate the rank, the amino acid variants at sites are sorted according to their relative energy. The energy rank is the position in the list (1–20) for the currently selected amino acid. We ran equilibration trajectories for 11 different selection pressure values corresponding to 10 mutations/site branch length. At every integer branch length, the average change in energy relative to the best choice amino acid and the average rank was evaluated. Fig 2A shows the result for a selection pressure value corresponding to lower stability than the native sequence. Destabilizing interactions are initially introduced into the protein, which increases the mean energy rank of the current amino acid. The fixated amino acids also have higher energy compared to the optimal choice for stability. Trajectories with varying selection pressure values will equilibrate at different protein stabilities. This is observed in Fig 2B where the average energy value for accepted sequences is plotted against selection pressure values. The mean sequence energy is linearly dependent on the selection pressure. At high selection pressure values, sequences have increased stability relative to the native sequence and the mean rank is low because the energetically best choice amino acid occurs very frequently at sites in the protein. At low selection pressure values, the mean rank is close to 10, which is the value expected with a completely random distribution of amino acids at sites. Strong selection pressures–high cytotoxicity/high abundance of the unfolded protein and/or large effective population size–thus result in proteins with increased thermodynamic stability. The sequence identity to the starting sequence range between 12% to 51% for the proteins equilibrated with different selection pressures, with higher identity for sequence equilibrated with higher selection pressure.

The selection pressure impacts the probability distribution over proposed and accepted ΔΔG values

Evolutionary trajectories at different selection pressure values were generated based on the final structure at the end of the equilibration runs. From these trajectories, we summarized the probability distribution over proposed (often referred to as Distribution of Fitness Effects, DFE) and accepted ΔΔG values as a function of the selection pressure value (Fig 3). With increasing selection pressure (resulting in higher protein stability), the mean energetic effect of mutation (ΔΔG) increases (Fig 3C). In other words, mutations become more detrimental when the protein stability increases. The distribution of mutational effects in real proteins behaves the same way albeit the increase of the detrimental effect is about 10 times higher [4].

At very low selection pressure values (resulting in low protein stabilities) the probability distribution over ΔΔG for proposed mutations is symmetric and centered around 0, with an

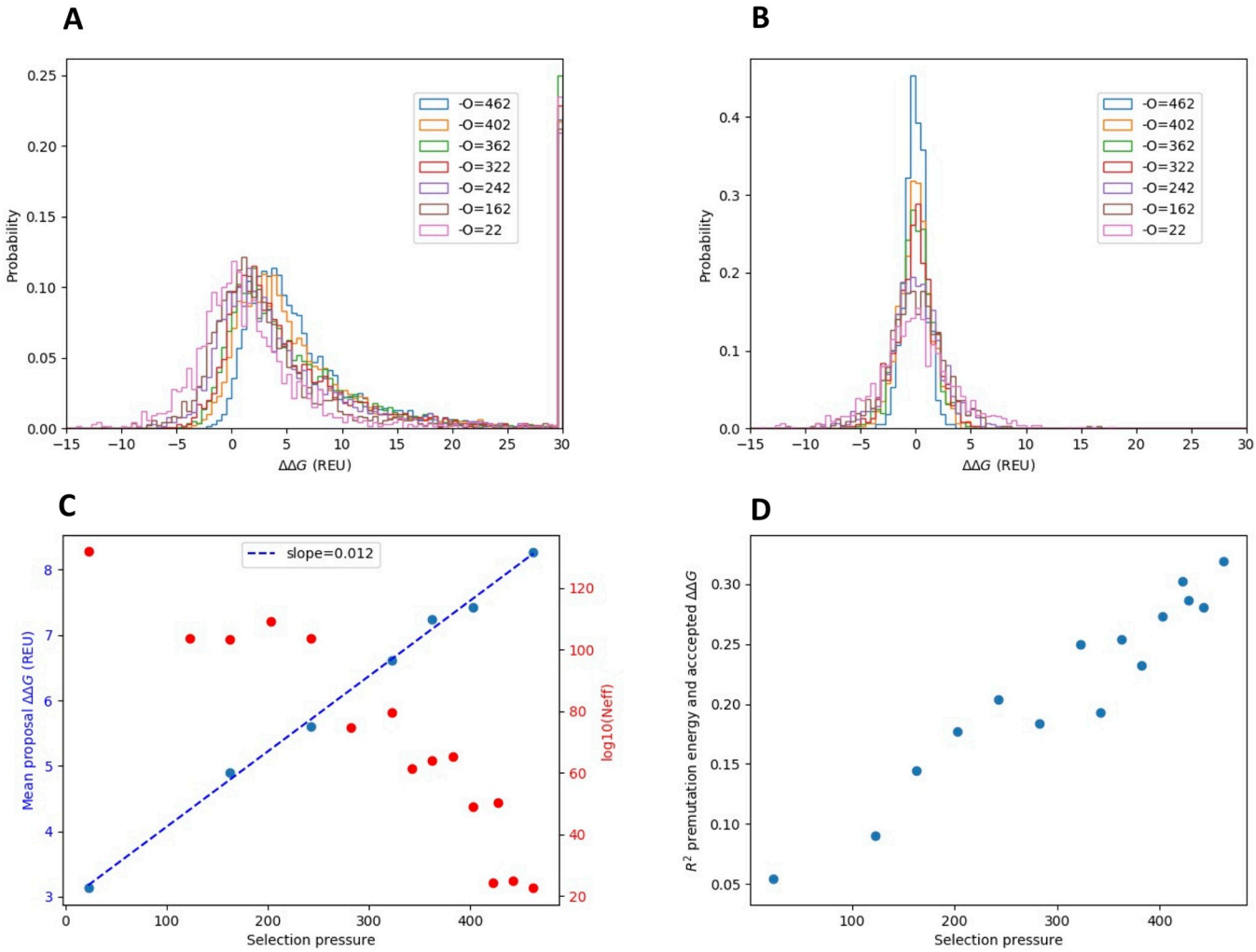

Fig 3. The selection pressure affects the ΔΔG proposal and acceptance probability distribution. A) Proposal ΔΔG probability distribution as function of selection pressures. Values above 30 energy units (corresponding to severe atomic clashes) were placed in the highest bin. **B)** Accepted ΔΔG probability distribution as function of selection pressure. **C)** Mean proposal ΔΔG value (correspond to the distribution < 100 energy units) as a function of selection pressure(blue), with a fitted line (blue). The logarithm of number of available sequences as a function of selection pressure (red) calculated with an assumption of independent sites in the protein. Energies in Rosetta Energy Units (REU). **D)** Correlation between premutation energy and accepted ΔΔG values as a function of selection pressure. Correlations are measured as squared Pearson correlation coefficients.

https://doi.org/10.1371/journal.pcbi.1010262.g003

equal probability of proposing stabilizing and destabilizing mutations. Under these conditions, the distribution over proposed and accepted ΔΔG values are almost identical. At higher selection pressures (resulting in higher protein stabilities), highly stabilizing mutations are much less likely to be proposed, and the probability distribution is shifted towards more destabilizing mutations (Fig 3A). The probability distribution over accepted values are symmetric around 0 for all selection pressure values but becomes more peaked as the stability increase (Fig 3B).

The mean of the proposed ΔΔG values linearly depends on selection pressure (and therefore on the mean stability of the protein, see Fig 3C). Why does the proposal probability change with the stability of the protein? At high stability, there are few accessible mutations that can stabilize the protein since the best choice amino acid is often already selected at many sites in the protein. We can estimate the space of accessible sequence at different reference stability values by multiplying together the effective number of amino acids at each site in the protein (assuming independent sites) calculated from the equilibrium amino acid frequency distribution at each site. As shown in Fig 3C, the sequence space is much smaller for more stable proteins evolved under higher selection pressure. This reduction in sequence space is likely to explain the shift of proposal ΔΔG values towards more destabilizing mutations at higher selection pressures.

The consequence of protein stability on the probability distributions over proposed and accepted ΔΔG -values have previously been studied by Goldstein using a contact-based energy model [7]. They found that the stability of the protein before mutation and the ΔΔG of accepted mutations correlated. We observe the same correlation with the all-atom simulations as seen in Fig 3D: Mutations accepted in a stable protein will generally be less stabilizing than those accepted in a protein with lower stability. The correlation between premutation stability and ΔΔG reduces with decreasing selection pressure. The influence of ΔG on the effect of mutations (ΔΔG) has also been demonstrated in other evolutionary simulations [9].

A strong covariation signal is found when phylogenetic trees are simulated by RosettaEvolve

The success of covariation analysis in identifying residue-residue contacts suggests that epistasis and coevolution are pervasive elements of evolution [25]. Yet, covariation, as measured by statistical coupling methods [26–29], is not necessarily the same as coevolution [30]. Statistical coupling methods are based on sequence alignments and do not consider that substitution has occurred along branches of phylogenetic trees. Tree-based methods to detect coevolution based on evolutionary theory have been developed [31,32], but their high computational cost hampers their use. A method developed to detect coevolving sites [33] does not identify contacting residues in evolutionary trajectories simulated by RosettaEvolve. The evolutionary basis for sequence covariation is therefore not fully understood. Talavera et al. [30] argued that coordinated sequence changes require very high selective pressures to occur, which results in rates so slow that coevolution would not be measurable. They argue that covariation is the consequence of sites with slow evolutionary rates rather than coevolution. Given the practical importance of statistical coupling methods in bioinformatics, it is of great interest to understand the relationship between covariation, coevolution, and protein energetics.

In this study, we investigate whether covariation signals emerge in sequences simulated from a phylogenetic tree using a detailed atomistic simulation of protein energetics and how the strength of the selection pressure affects the covariation signal. To address these questions, we inferred a phylogenetic tree from an alignment of the natural sequence of azurin and used it as the basis for evolutionary simulations with RosettaEvolve. Simulated phylogenetic trees were generated at different selection pressures, starting from equilibrated sequences at each

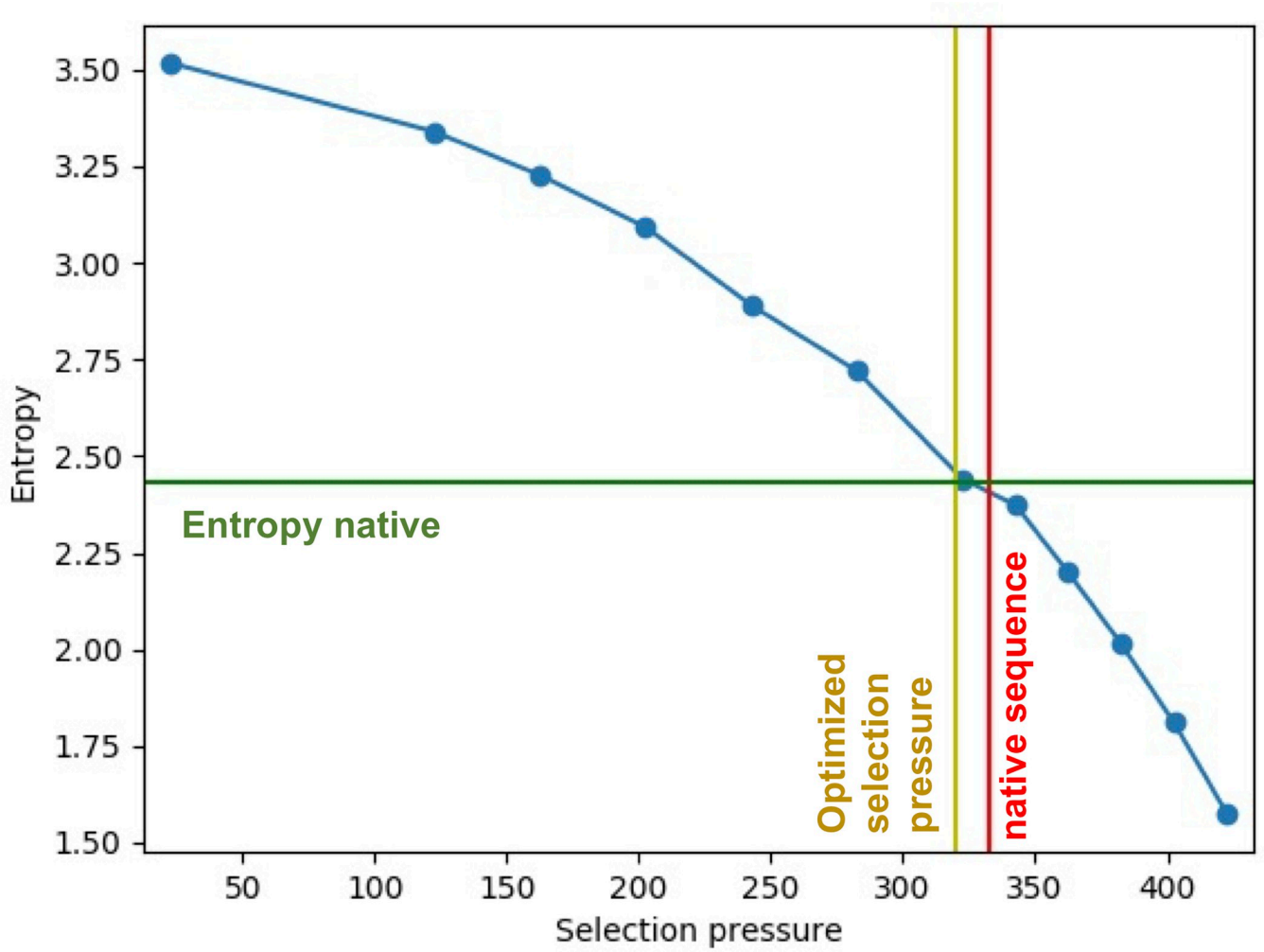

Fig 4. Simulation of phylogenetic trees of azurin with RosettaEvolve. Dependence of sequence entropy of leaf sequences with reference energies. The green line corresponds to the entropy in the natural sequences used to infer the azurin tree. The red line corresponds to the energy of the native sequence of azurin. The yellow line corresponds to the selection pressure that maximizes the correlation between computed and empirical amino acid substitution rates in Norn et al. [7].

https://doi.org/10.1371/journal.pcbi.1010262.g004

given selection pressure. We developed a recursive algorithm that generates evolutionary trajectories over a given tree topology and branch lengths. We populated the tree 11 different times with variable selection pressure.

The sequence entropy at the leaves of simulated azurin trees depends strongly on the selection pressure (Fig 4). Using parameter values we previously found to explain natural amino acid substitution patterns [19,33], we see similar position-specific sequence entropies between our simulated proteins and their natural counterparts. Jiang et al. have studied amino acid diversity in evolved sequences relative to sequences generated by protein design and found that evolved sequences are more similar to natural sequences [16]. Our results demonstrate that native-like sequence distributions can be achieved by controlling the applied selection pressure.

The leaf sequences generated by RosettaEvolve trajectories simulated at different selection pressure values were analyzed for covariation signal statistical coupling score using Gremlin [34]. The ability of Gremlin in predicting residue-residue contacts was summarized in

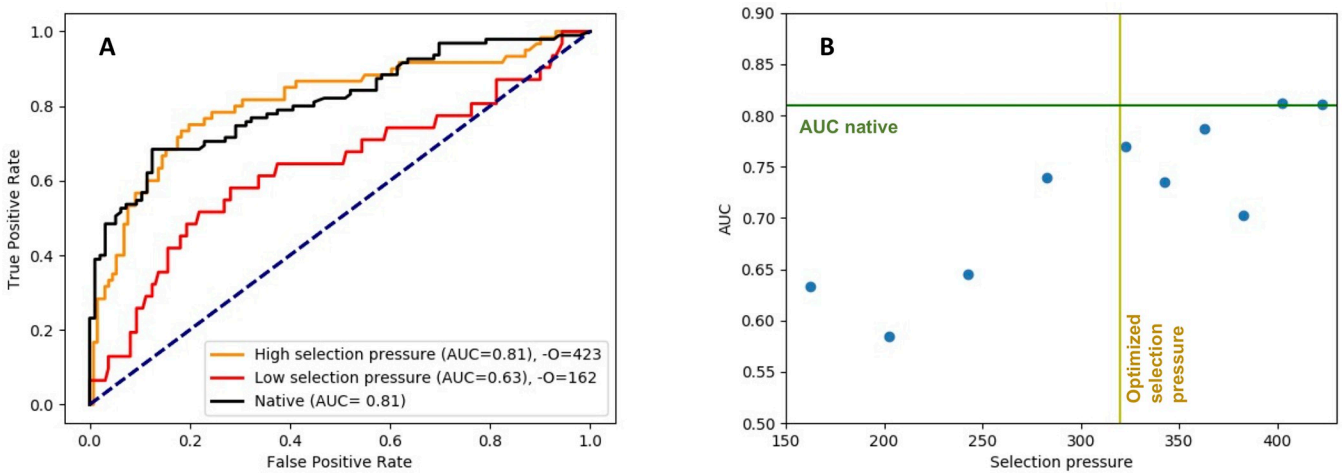

Fig 5. ROC curve for residue-residue contact prediction. A) Comparison of ROC curve for natural sequences (black) and two simulated alignments (red and orange) at two different reference energies. The blue line shows the diagonal. B) Dependence of contact-prediction accuracy (AUC) on selection pressure. The green line corresponds to AUC for the natural sequence. The yellow line corresponds to the selection pressure that gives the optimal correspondence between predicted and empirical amino acid substitution rates in a Rosetta-based rate prediction method [14,33].

https://doi.org/10.1371/journal.pcbi.1010262.g005

Receiver Operator Characteristic (ROC) curves, where the true positive rate is plotted against false positive rate. In Fig 5, the ROC curve for the natural sequences is compared to sequences simulated at two different selection pressures, one corresponding to low (red curve) and one to high selection pressure (orange curve). The area under the curve (AUC) is a metric for the overall performance. For the high selection pressure simulations, the AUC reaches the same values as the natural sequences, while sequences evolved with low selection pressure provide a considerably worse basis for predicting residue-residue contacts.

Even though similar AUC values are found for some simulated sequences, the early enrichment is nonetheless better for the natural sequence, resulting in higher prediction accuracy in the range relevant for structure prediction. The overall predictive power (characterized by AUC) is highly dependent on the selection pressure. In Fig 5B, the AUC is plotted against selection pressure. The ability of GREMLIN to identify true residue-residue contacts drops with decreasing protein stability (as controlled by the selection pressure). For reference stabilities corresponding to most stable proteins, up to 78% (43 out of 55 contacts above the threshold used by Gremlin to predict contacts) of the predicted residue-contacts contacts are validated in the structure corresponding to the native sequence of azurin.

We calculated residue-residue pair energies from the crystal structure of azurin using Rosetta, see Fig 6. The average pair energy between residues predicted to be in contact by Gremlin (blue distribution) has pair interaction values that are considerably stronger than contacts in general in the protein (grey distribution). The mean interaction energy is around -1 REU for sequences simulated with stabilizing reference energies for the predicted Gremlin contacts, compared to -0.15 REU for all contacts in the protein. So, for the most stable proteins, contacts detected by statistical coupling analysis correspond to pair interactions among the most stabilizing contacts in the protein.

We further analyzed the emergence of contacts detected by the statistical coupling analysis during the evolutionary trajectory. Starting from the leaf nodes, we identify the branch point where a residue pair found in the leaf node was first introduced and characterize the change in energy on the evolutionary path towards the leaf node. We find that the average change in energy for the two residues in the predicted contact (in the context of the entire structure) is

favorable, but only slightly so (-0.15 REU). Thus, when the pair was formed, there did not appear to be a large energetic gain in forming the contact. However, the selected pair may become energetically entrenched after initially appearing (evolutionary Stokes shift [17]), or there may be special conditions before it was inserted. Further analysis of the fluctuation in selection coefficients over time will have to be carried out to fully understand the mechanism behind covariation signals for these residue pairs.

Fluctuations in protein stability result in fluctuations in site rates

During the evolutionary trajectory, sites in the protein will experience a fluctuating structural environment (and therefor changes in intermolecular interactions). How much do site rates fluctuate during a mutational trajectory? How much of this variation can be explained by fluctuations in protein stability during an evolutionary trajectory? We calculated site-specific rates across an evolutionary trajectory corresponding to a branch length of 1 mutation per site to address these questions. After each mutation, we calculated the energy, site-rates and compared site rates to empirical values predicted by rate4site [35], a method to infer site rates from sequence alignments, from the sequence alignment of azurin. Fig 7A shows the energy and correlation with empirical site rates fluctuate across the trajectory. The Pearson correlation between calculated and empirical site-specific rates fluctuates considerably during the trajectory, ranging from 0.48 to 0.61. Fluctuations in the stability of the protein (Fig 6A, red line) will result in an overall change in substitution rate, with less stable proteins having higher

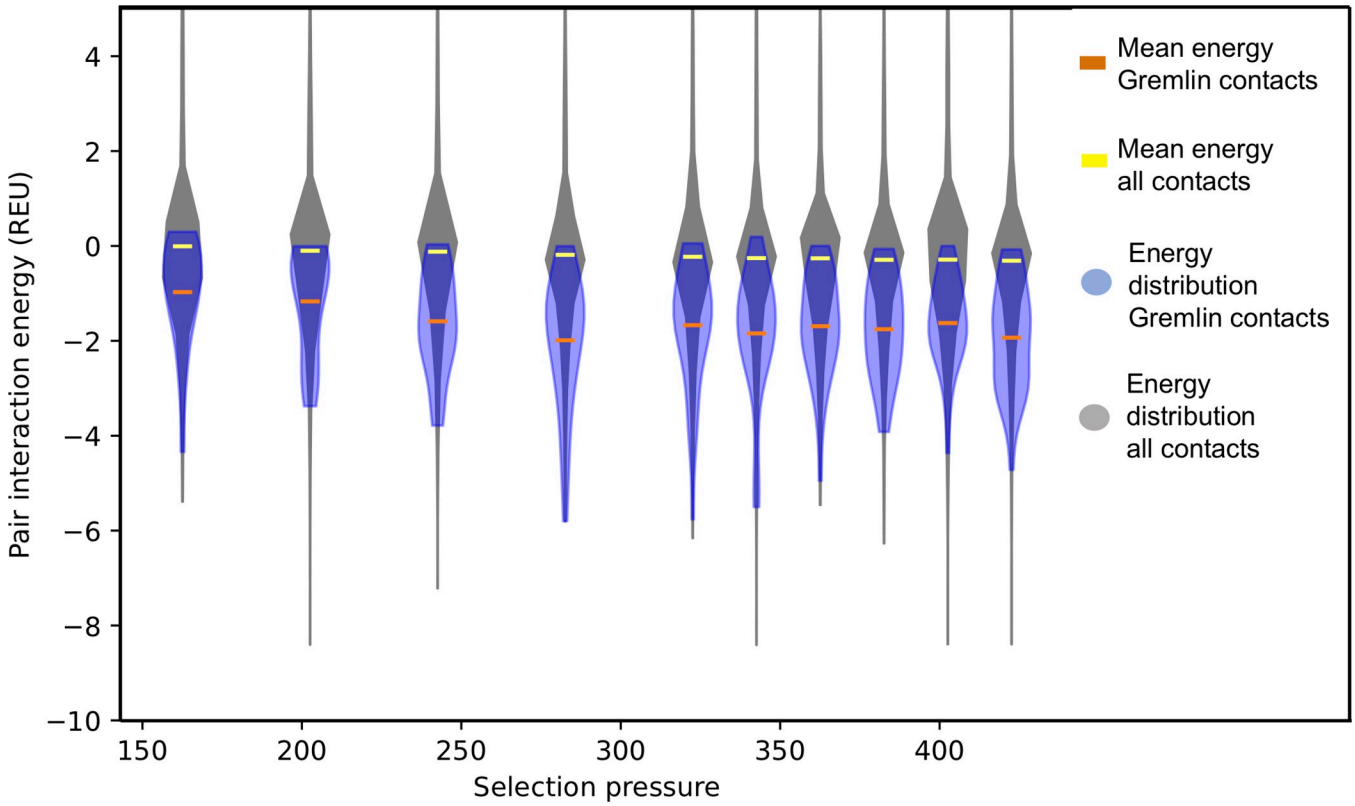

Fig 6. Pair interaction energy for contacts predicted from alignments. Distribution of pair interaction energies for contacts predicted by Gremlin (blue, mean represented by the orange line) and all contacts within the structure (gray, mean represented by the yellow line) as a function of selection pressure. The width of the violin is related to the frequency of a given pair interaction value. Energies in Rosetta Energy Units (REU).

https://doi.org/10.1371/journal.pcbi.1010262.g006

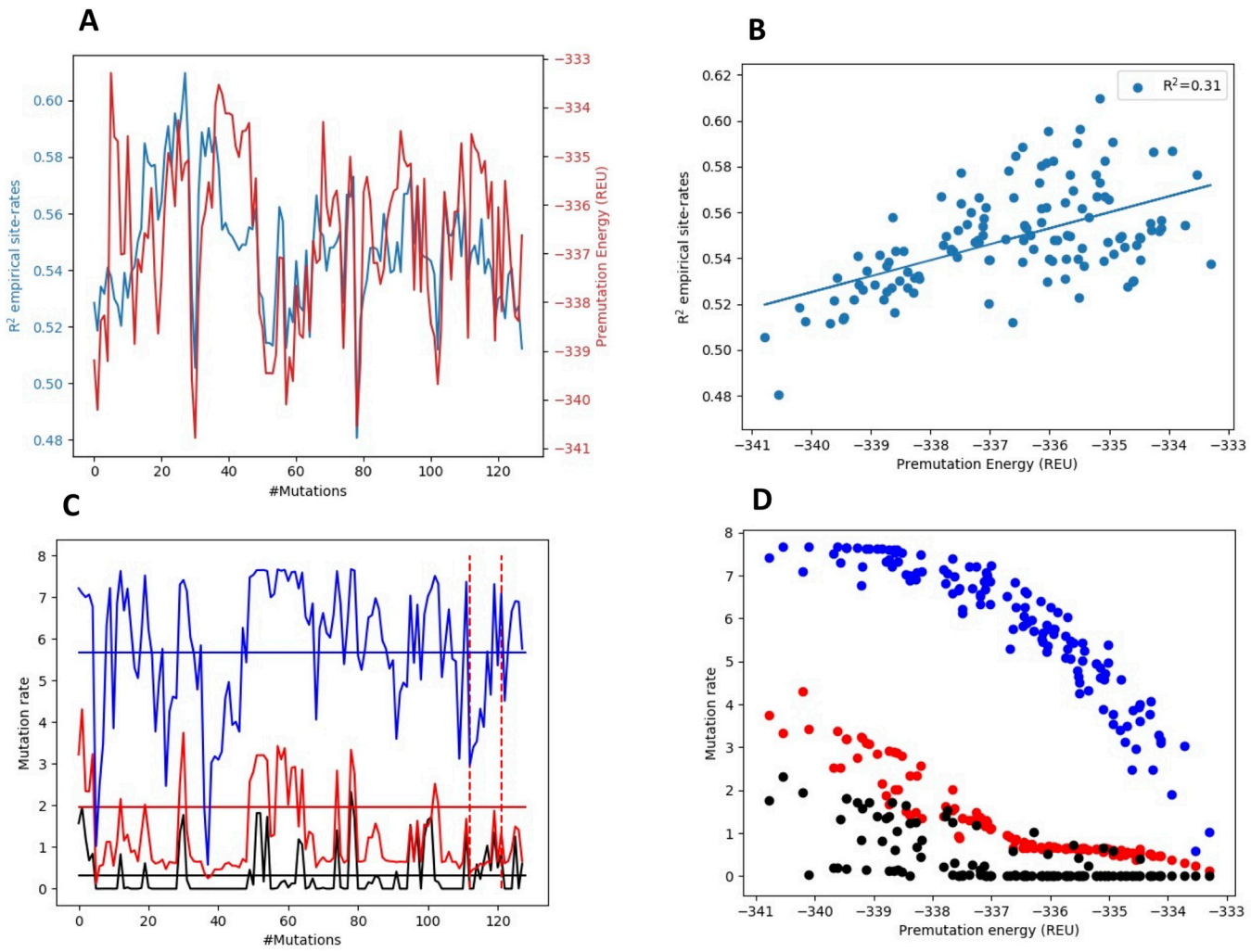

Fig 7. Fluctuation in substitution rates during a trajectory and correlations with stability. A) Correlation between calculated and empirical site-specific rates (calculated as R^2-values) for azurin (blue) as a function of the number of introduced mutations. Fluctuation in energy as a function of introduced mutations (blue). B) Correlation between R^2-values and energies are shown in A). Selection pressure in trajectory is set to -322. C) Site-specific rates for the sites in azurin. Residue 41 (blue), residue 49 (green) and residue 66 (red). Empirical site rates from rate4site as solid lines and dashed lines show mutational events at the site. D) Dependence of site-rates in C) with premutation stability. Energies in Rosetta Energy Units (REU).

https://doi.org/10.1371/journal.pcbi.1010262.g007

mutation rates. About 31% of the variation in the correlation with empirical site rates can be explained by the fluctuation in stability of the protein during the trajectory ($R^2 = 0.31$).

The rate at individual sites in the protein will also fluctuate considerably. In Fig 7C, the rates for three different individual sites in the protein are plotted as a function of the number of mutations in the trajectory. The mutation rate can drop an order of magnitude during the trajectory, even though there are no mutations occurring at that site. Fig 7D demonstrates that rates at individual sites can be highly coupled to the stability of the proteins.

Discussion

Protein stability results from the net balance of forces involving thousands of interacting atoms. The result is typically a protein with only marginal stability where small changes in atomic interactions can shift the protein from a folded to unfolded state. The marginal stability of proteins can be understood as the result of the balance between the introduction of

predominantly destabilizing mutations and selection [7,8,10]. This marginal stability also emerges in evolutionary simulations employing a simple contact-based potential of protein energetics [7] and in simulations where ΔΔG values are sampled from a probability distribution [8]. Nonetheless, many mechanistically important aspects of protein evolution may be lost without consideration of the detailed atomic interactions in proteins.

A few investigations have been presented where evolutionary trajectories have been simulated with atomistic energy functions. Typically, these studies have employed the FoldX energy function [22] to evaluate ΔΔG values [18,36], but also the ERIS [24] ΔΔG predictor have been used [15,20,25]. In this manuscript, we simulate evolution with the Rosetta macromolecular modeling package, which provides a powerful framework for modeling the structure and energetics of proteins [28] and where side-chain and backbone flexibility can be modeled with a wide range of structure-prediction protocols. RosettaEvolve can be readily extended to use additional fitness models and additional methods to model conformational changes upon mutations, such as the flexible backbone approach developed by Bartlow et al. to model mutations in protein interfaces [37].

A unique aspect of the approach in RosettaEvolve is the fact that mutations are introduced on the fly, enabling studies of issues like entrenchment and residue-residue coevolution. This follows work from Jiang et al. [18] that used Rosetta to generate evolutionary trajectories along a single branch. There are several differences between their approach and ours. We simulate evolution based on nucleotide mutations, rather than at the protein level, we use ΔΔG calculations that include backbone flexibility rather than the raw Rosetta energy, a new way of acceleration mutation fixation events, and our models differ in the selection pressure and the effective population sizes. Jiang et al. [18] used protein design calculations to set the selection pressure value and used a smaller effective population size (100).

We have previously developed a Rosetta-based method to predict amino acid substitution rates [33] from protein structure using the combination of structure-based stability calculations and mutation-selection model, which we refer to as the TMS (Thermodynamic Mutation-Selection) model. Amino acid substitution rates at a site calculated can readily be summed up to evaluate the site substitution rates [14]. A benefit of the TMS method is that it enables us to evaluate the site-specific rates for all sequences continuously along an evolutionary trajectory. Our results show that the site rates fluctuate considerably during the trajectory, even for sites that are not mutated. Natural proteins have also experienced significant variation in backbone structure during their evolutionary trajectories as reflected by the structural variability found in sequence homologs. Such relatively large-scale structural fluctuations are not modeled with the limited backbone flexibility ΔΔG method used in this study. Our simulation results highlight that relatively small changes in structure and energetics in proteins can have considerable consequences for substitution rates at individual sites in proteins and that accurate prediction of site-rates hinges on modeling the detailed structural and energetic consequences of amino acid substitutions. Nonetheless, a significant amount of the fluctuation in substitution rates can be explained simply by fluctuations in the overall stability of proteins during the evolutionary trajectory (Fig 7B and 7D). The relationship between protein stability and evolutionary rates has been studied previously at the level of genes, where it has been demonstrated that proteins that are less stable evolve faster [38].

An underlying approximation in this study is that protein folding can largely be described by two states, a folded and active native state, and an unfolded/misfolded/inactive state. Two-state folding is a good model for the folding of many small single-domain proteins [39]. Based on this underlying assumption, several evolutionary phenomena can be rationalized: the relationship between gene expression and evolutionary rates [5], distributions of evolutionary rates [40], mutational robustness [41], distribution of thermodynamic stabilities in proteins

[42] and stability-mediated epistasis in virus evolution [43]. Nonetheless, two-state behavior is an oversimplification and even single mutations can trigger alternatively folded states [44]. General treatment of how the presence of alternative folding states shapes sequence evolution with an atomistic model is probably computationally intractable, but the effect of a single additional state could be studied by an expanded stability fitness function.

We show that phylogenetic trees populated with sequences using an evolutionary all-atom structural and energetic model result in sequences with a significant covariation signal. Sites with high statistical coupling have considerably more favorable pair interaction energies than average contacts in proteins. This suggests that the basic premise behind statistical coupling analysis for contact prediction—that strong residue-residue interactions lead to covariation signal—is correct. Nonetheless, although some covariation signal is also observed at lower selection pressures, only at very high fitness pressures does the covariation signal reach the levels seen for natural sequences. Furthermore, the limited backbone flexibility in the simulation likely overestimates the relative strength of specific residue-residue interactions, resulting in enhanced covariation signals. We, therefore, expect that more realistic modeling of structural variability would reduce the covariation signal. Taken together, this may suggest that additional mechanisms can be behind the strong covariation signal found in natural protein sequences. Further investigations of the correlation between the substitution history of RosettaEvolve trajectories, statistical coupling score, and protein energetics should enable a more detailed understanding of how covariation emerges among homologous proteins.

Materials and methods

DNA substitution model

The DNA mutational model has two parameters, the transition/transversion rate ratio κ and the whole codon mutation rate ρ. The relative rate of single base pair changes to multi-codon mutation depends on ρ but also on the number of states that are accessible for the multi-nucleotide route: We calculate the probability of multi-nucleotide changes as

$$p_{multi-nucleotide} = \frac{63*\rho}{63*\rho + 3*\kappa + 6}$$

κ and ρ are parameters in the simulation. In this study, we have set the values found to optimize the correlations with empirical amino acid substitution rates presented in Norn et. al. [33], κ = 2.7 and ρ = 0.1.

Fixation probabilities are scaled to improve computational efficiency

Computed fixation probabilities are generally too low to enable efficient simulation of evolutionary trajectories. To accelerate sampling, we used adaptive importance sampling [45] In the simplest case, importance sampling relates the target distribution, f(x), to the sampling distribution, q(x), by a scale-factor, w:

$$w = f(x)/q(x)$$

For each substitution in a given sequence, *i*, we compute a scaling factor, $w = f_i^{MAX}$, such that a proposed sequence, j, with maximum possible fitness, $\omega^{MAX} = 1$ (achieved for $\Delta G \rightarrow -\infty$) is fixed with a $q_{i \rightarrow j} = 1$:

$$f_i^{MAX} = \frac{1 - exp(-2s_i^{MAX})}{1 - exp(-4Ns_i^{MAX})}$$

where the maximum selection coefficient is

$$s_i^{MAX} = \omega^{MAX}/\omega_i - 1$$

During an evolutionary trajectory and given sequence, i, our target distribution is thus $f(x) = f_{i\to j}$ and the sampling distribution $q(x) = f_{i\to j}/f_i^{MAX}$.

Structural modeling

The crystal structure of azurin (PDB ID: 5AZU [35]) was used as the basis for all modeling. All structural modeling was done with the Rosetta macromolecular modeling suite [28] using the beta_nov16 energy function. A monomer from 5AZU was energy refined before running the evolutionary trajectories using the method described by Niven et al. [46] to make the crystal structure compatible with the energy function. The copper ion was not maintained in the simulation. Prediction of ΔΔG values for mutations was done using a modified version of the approach presented by Park et al. [27], with a 6.0 instead of 9.0 Å distance cutoff in the Lennard-Jones energy term. Backbone flexibility is allowed at the mutated and neighboring residues, and side-chains are repacked for all residues that have at least an interaction energy more than 0.1 REU. To put the REU units for ΔΔG values on the kcal/mol scale a conversion factor for beta_nov16 was used and calculated by correlation to experimental ΔΔG values. A single ΔΔG prediction was used to evaluate each mutation. Rosetta version unknown: aafaa0d91c9e6b83998e8592570267d5fca2a501 was used for the simulations. Simulations times varies greatly with applied selection pressure. For reference, the equilibration runs shown in Fig 2 took between 134 to 303 CPU hours.

Site-specific rate calculations

Site rates were calculated with the TMS method presented in Norn et al. [7] as described in [40]. In reference [40] a single selection pressure was fitted for a benchmark of 66 proteins based on maximizing the similarity with empirical site-specific rates. In this study, the selection pressure used in the rate calculation corresponds to the value used in the evolutionary trajectory that generated the structure. Empirical rates for azurin were calculated with rate4site [35] using the empirical Bayes method with the LG instantaneous rate matrix and an alignment consisting of 500 sequences.

Simulation of evolution along predefined phylogenetic trees

A phylogenetic tree was generated based on a sequence alignment generated by Gremlin [47] using RAxML [48] with the LG as the instantaneous rate matrix. The phylogenetic tree was filled with a node at the center of the tree to optimize computational speed. For each branch, RosettaEvolve was run with the number of mutations expected from the branch length in the empirical tree. At each internal node, a structure is stored and used as a basis for the next set of branches originating from each leaf. The starting structures/sequences for the simulations are the models generated at the end of the equilibration trajectories at each studied offset (selection pressure).

Statistical coupling analysis

Leaf sequences from the phylogenetic tree simulation (1050 sequences) were analyzed with Gremlin [47] web server (gremlin.bakerlab.org). Gremlin was run without MSA enrichment so that only simulated sequences was used in the analysis. Classification of contact prediction was done using the standard distance threshold of 8.0 Å between Cb (Ca for glycine) using the

coordinates in 5AZU. A default threshold value of a scaled score above 1.0 was used to select contacts predicted by Gremlin.

Pair energies were determined using the residue_energy_breakdown.linuxgccrelease application using the beta_nov16 energy function and the energy-refined 5AZU structure.

Command lines and code

RosettaEvolve is available through Rosetta [28], which can be downloaded at rosettacommons.org. Additional scripts and running information can be found at https://github.com/Andre-lab/RosettaEvolve and in S1 Text. Command lines used in this study are found in S1 Text.

Supporting information

S1 Text. Command lines and run examples.
(PDF)

Acknowledgments

We thank Douglas L. Theobald for helpful discussions on the implementation of RosettaEvolve.

Author Contributions

Conceptualization: Christoffer Norn, Ingemar André.

Data curation: Ingemar André.

Formal analysis: Ingemar André.

Funding acquisition: Ingemar André.

Investigation: Ingemar André.

Methodology: Christoffer Norn, Ingemar André.

Project administration: Ingemar André.

Software: Christoffer Norn.

Supervision: Ingemar André.

Writing – original draft: Ingemar André.

Writing – review & editing: Christoffer Norn, Ingemar André.

References

1. Ghosh K, Dill K. Cellular Proteomes Have Broad Distributions of Protein Stability. Biophys J. 2010; 99 (12):3996–4002. https://doi.org/10.1016/j.bpj.2010.10.036 WOS:000285438900023. PMID: 21156142
2. Christensen S, Ramisch S, Andre I. DnaK response to expression of protein mutants is dependent on translation rate and stability. Commun Biol. 2022; 5(1). ARTN 597 WOS:000812308700002. https://doi.org/10.1038/s42003-022-03542-2 PMID: 35710941
3. Geiler-Samerotte KA, Dion MF, Budnik BA, Wang SM, Hartl DL, Drummond DA. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. Proc Natl Acad Sci U S A. 2011; 108(2):680–5. Epub 2010/12/29. https://doi.org/10.1073/pnas.1017570108 PMID: 21187411; PubMed Central PMCID: PMC3021021.
4. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins evolve slowly. P Natl Acad Sci USA. 2005; 102(40):14338–43. https://doi.org/10.1073/pnas.0504070102 WOS:000232392900040. PMID: 16176987

5. Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell. 2008; 134(2):341–52. Epub 2008/07/30. https://doi.org/10.1016/j.cell.2008.05.042 PMID: 18662548; PubMed Central PMCID: PMC2696314.

6. Williams PD, Pollock DD, Blackburne BP, Goldstein RA. Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput Biol. 2006; 2(6):e69. Epub 2006/06/23. https://doi.org/10.1371/journal.pcbi.0020069 PMID: 16789817; PubMed Central PMCID: PMC1480538.

7. Goldstein RA. The evolution and evolutionary consequences of marginal thermostability in proteins. Proteins Struct Funct Bioinform. 2011; 79(5):1396–407. https://doi.org/10.1002/prot.22964 PMID: 21337623.

8. Serohijos AW, Rimas Z, Shakhnovich EI. Protein biophysics explains why highly abundant proteins evolve slowly. Cell Rep. 2012; 2(2):249–56. Epub 2012/09/04. https://doi.org/10.1016/j.celrep.2012.06.022 PMID: 22938865; PubMed Central PMCID: PMC3533372.

9. Serohijos AWR, Shakhnovich EI. Contribution of Selection for Protein Folding Stability in Shaping the Patterns of Polymorphisms in Coding Regions. Molecular Biology and Evolution. 2014; 31(1):165–76. https://doi.org/10.1093/molbev/mst189 PMID: 24124208.

10. Taverna DM, Goldstein RA. Why are proteins marginally stable? Proteins Struct Funct Bioinform. 2002; 46(1):105–9. https://doi.org/10.1002/prot.10016 PMID: 11746707.

11. Serohijos AW, Lee SY, Shakhnovich EI. Highly abundant proteins favor more stable 3D structures in yeast. Biophys J. 2013; 104(3):L1–3. Epub 2013/02/28. https://doi.org/10.1016/j.bpj.2012.11.3838 PMID: 23442924; PubMed Central PMCID: PMC3566449.

12. Echave J, Jackson EL, Wilke CO. Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. Phys Biol. 2015; 12(2):025002. Epub 2015/03/20. https://doi.org/10.1088/1478-3975/12/2/025002 PMID: 25787027; PubMed Central PMCID: PMC4391963.

13. Echave J, Wilke CO. Biophysical Models of Protein Evolution: Understanding the Patterns of Evolutionary Sequence Divergence. Annu Rev Biophys. 2017; 46:85–103. Epub 2017/03/17. https://doi.org/10.1146/annurev-biophys-070816-033819 PMID: 28301766; PubMed Central PMCID: PMC5800964.

14. Norn HC. An evolutionary basis for protein desig and structure prediction: Lund University; 2019.

15. Dasmeh P, Serohijos AWR, Kepp KP, Shakhnovich EI. The Influence of Selection for Protein Stability on dN/dS Estimations. Genome Biol Evol. 2014; 6(10):2956–67. https://doi.org/10.1093/gbe/evu223 PMID: 25355808.

16. Jiang Q, Teufel AI, Jackson EL, Wilke CO. Beyond Thermodynamic Constraints: Evolutionary Sampling Generates Realistic Protein Sequence Variation. Genetics. 2018; 208(4):1387–95. Epub 2018/02/01. https://doi.org/10.1534/genetics.118.300699 PMID: 29382650; PubMed Central PMCID: PMC5887137.

17. Pollock DD, Thiltgen G, Goldstein RA. Amino acid coevolution induces an evolutionary Stokes shift. Proc Natl Acad Sci U S A. 2012; 109(21):E1352–9. Epub 2012/05/02. https://doi.org/10.1073/pnas.1120084109 PMID: 22547823; PubMed Central PMCID: PMC3361410.

18. Shah P, McCandlish DM, Plotkin JB. Contingency and entrenchment in protein evolution under purifying selection. Proc Natl Acad Sci U S A. 2015; 112(25):E3226–35. Epub 2015/06/10. https://doi.org/10.1073/pnas.1412933112 PMID: 26056312; PubMed Central PMCID: PMC4485141.

19. Wylie CS, Shakhnovich EI. A biophysical protein folding model accounts for most mutational fitness effects in viruses. Proc Natl Acad Sci U S A. 2011; 108(24):9916–21. Epub 2011/05/26. https://doi.org/10.1073/pnas.1017572108 PMID: 21610162; PubMed Central PMCID: PMC3116435.

20. Gauthier L, Di Franco R, Serohijos AWR. SodaPop: a forward simulation suite for the evolutionary dynamics of asexual populations on protein fitness landscapes. Bioinformatics. 2019; 35(20):4053–62. Epub 2019/03/16. https://doi.org/10.1093/bioinformatics/btz175 PMID: 30873519.

21. Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS. The stability effects of protein mutations appear to be universally distributed. J Mol Biol. 2007; 369(5):1318–32. Epub 2007/05/08. https://doi.org/10.1016/j.jmb.2007.03.069 PMID: 17482644.

22. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol. 2002; 320(2):369–87. Epub 2002/06/25. https://doi.org/10.1016/S0022-2836(02)00442-4 PMID: 12079393.

23. Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins. 2011; 79(3):830–8. Epub 2011/02/03. https://doi.org/10.1002/prot.22921 PMID: 21287615; PubMed Central PMCID: PMC3760476.

24. Yin S, Ding F, Dokholyan NV. Eris: an automated estimator of protein stability. Nat Methods. 2007; 4(6):466–7. Epub 2007/06/01. https://doi.org/10.1038/nmeth0607-466 PMID: 17538626.

25. Serohijos Adrian WR, Lee SYR, Shakhnovich Eugene I. Highly Abundant Proteins Favor More Stable 3D Structures in Yeast. Biophysical Journal. 2013; 104(3):L1–L3. https://doi.org/10.1016/j.bpj.2012.11.3838 PMID: 23442924.

26. Starr TN, Thornton JW. Epistasis in protein evolution. Protein Sci. 2016; 25(7):1204–18. Epub 2016/02/03. https://doi.org/10.1002/pro.2897 PMID: 26833806; PubMed Central PMCID: PMC4918427.

27. Park H, Bradley P, Greisen P Jr, Liu Y, Mulligan VK, Kim DE, et al. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. J Chem Theory Comput. 2016; 12(12):6201–12. Epub 2016/10/22. https://doi.org/10.1021/acs.jctc.6b00819 PMID: 27766851; PubMed Central PMCID: PMC5515585.

28. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011; 487:545–74. Epub 2010/12/29. https://doi.org/10.1016/B978-0-12-381270-4.00019-6 PMID: 21187238; PubMed Central PMCID: PMC4083816.

29. Topal MD, Fresco JR. Base pairing and fidelity in codon-anticodon interaction. Nature. 1976; 263 (5575):289–93. Epub 1976/09/23. https://doi.org/10.1038/263289a0 PMID: 958483

30. Topal MD, Fresco JR. Complementary base pairing and the origin of substitution mutations. Nature. 1976; 263(5575):285–9. Epub 1976/09/23. https://doi.org/10.1038/263285a0 PMID: 958482

31. Harris K, Nielsen R. Error-prone polymerase activity causes multinucleotide mutations in humans. Genome Res. 2014; 24(9):1445–54. Epub 2014/08/01. https://doi.org/10.1101/gr.170696.113 PMID: 25079859; PubMed Central PMCID: PMC4158752.

32. Reid TM, Loeb LA. Tandem double CC—>TT mutations are produced by reactive oxygen species. Proc Natl Acad Sci U S A. 1993; 90(9):3904–7. Epub 1993/05/01. https://doi.org/10.1073/pnas.90.9.3904 PMID: 8483909; PubMed Central PMCID: PMC46414.

33. Norn C, Andre I, Theobald DL. A thermodynamic model of protein structure evolution explains empirical amino acid substitution matrices. Protein Sci. 2021; 30(10):2057–68. Epub 2021/07/05. https://doi.org/10.1002/pro.4155 PMID: 34218472; PubMed Central PMCID: PMC8442976.

34. McCandlish DM, Stoltzfus A. Modeling evolution using the probability of fixation: history and implications. Q Rev Biol. 2014; 89(3):225–52. Epub 2014/09/10. https://doi.org/10.1086/677571 PMID: 25195318.

35. Nar H, Messerschmidt A, Huber R, Vandekamp M, Canters GW. Crystal-Structure Analysis of Oxidized Pseudomonas-Aeruginosa Azurin at Ph 5.5 and Ph 9.0—a Ph-Induced Conformational Transition Involves a Peptide-Bond Flip. Journal of Molecular Biology. 1991; 221(3):765–72. https://doi.org/10.1016/0022-2836(91)80173-R WOS:A1991GL15400008. PMID: 1942029

36. Ashenberg O, Gong LI, Bloom JD. Mutational effects on stability are largely conserved during protein evolution. Proc Natl Acad Sci U S A. 2013; 110(52):21071–6. Epub 2013/12/11. https://doi.org/10.1073/pnas.1314781111 PMID: 24324165; PubMed Central PMCID: PMC3876214.

37. Barlow KA, S OC, Thompson S, P Suresh, Lucas JE, Heinonen M, et al. Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity upon Mutation. J Phys Chem B. 2018; 122(21):5389–99. Epub 2018/02/06. https://doi.org/10.1021/acs.jpcb.7b11367 PMID: 29401388; PubMed Central PMCID: PMC5980710.

38. Serohijos Adrian WR, Rimas Z, Shakhnovich Eugene I. Protein Biophysics Explains Why Highly Abundant Proteins Evolve Slowly. Cell Reports. 2012; 2(2):249–56. https://doi.org/10.1016/j.celrep.2012.06.022 PMID: 22938865.

39. Jackson SE. How do small single-domain proteins fold? Fold Des. 1998; 3(4):R81–91. Epub 1998/08/26. https://doi.org/10.1016/S1359-0278(98)00033-9 PMID: 9710577.

40. Lobkovsky AE, Wolf YI, Koonin EV. Universal distribution of protein evolution rates as a consequence of protein folding physics. P Natl Acad Sci USA. 2010; 107(7):2983–8. https://doi.org/10.1073/pnas.0910445107 WOS:000274599500051. PMID: 20133769

41. Bloom JD, Labthavikul ST, Otey CR, Arnold FH. Protein stability promotes evolvability. Proc Natl Acad Sci U S A. 2006; 103(15):5869–74. Epub 2006/04/04. https://doi.org/10.1073/pnas.0510098103 PMID: 16581913; PubMed Central PMCID: PMC1458665.

42. Zeldovich KB, Chen P, Shakhnovich EI. Protein stability imposes limits on organism complexity and speed of molecular evolution. Proc Natl Acad Sci U S A. 2007; 104(41):16152–7. Epub 2007/10/05. https://doi.org/10.1073/pnas.0705366104 PMID: 17913881; PubMed Central PMCID: PMC2042177.

43. Gong LI, Suchard MA, Bloom JD. Stability-mediated epistasis constrains the evolution of an influenza protein. Elife. 2013; 2. ARTN e00631 WOS:000328614800004. https://doi.org/10.7554/eLife.00631 PMID: 23682315

44. Connell KB, Horner GA, Marqusee S. A Single Mutation at Residue 25 Populates the Folding Intermediate of E. coli RNase H and Reveals a Highly Dynamic Partially Folded Ensemble. Journal of Molecular

Biology. 2009; 391(2):461–70. https://doi.org/10.1016/j.jmb.2009.05.084 WOS:000269227300017. PMID: 19505477

45. Huber M. Handbook of Markov Chain Monte Carlo. Chapman Hall Crc Handbooks Mod Statistical Methods. 2011. https://doi.org/10.1201/b10905-10

46. Nivon LG, Moretti R, Baker D. A Pareto-optimal refinement method for protein design scaffolds. PLoS One. 2013; 8(4):e59004. Epub 2013/04/09. https://doi.org/10.1371/journal.pone.0059004 PMID: 23565140; PubMed Central PMCID: PMC3614904.

47. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci U S A. 2013; 110(39):15674–9. Epub 2013/09/07. https://doi.org/10.1073/pnas.1314045110 PMID: 24009338; PubMed Central PMCID:PMC3785744.

48. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30(9):1312–3. Epub 2014/01/24. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623; PubMed Central PMCID:PMC3998144.