Europe PMC Funders Group
Author Manuscript
Nat Ecol Evol. Author manuscript; available in PMC 2018 April 23.

Published in final edited form as:
Nat Ecol Evol. 2017 December ; 1(12): 1923–1930. doi:10.1038/s41559-017-0338-9.

Sequence entropy of folding and the absolute rate of amino acid substitutions

Richard A. Goldstein[1] and **David D. Pollock**[2,*]

[1]Division of Infection & Immunity, University College London, London, WC1E 6BT, UK

[2]Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045 USA

Summary

Adequate representations of protein evolution should consider how the acceptance of mutations depends on the sequence context in which they arise. However, epistatic interactions among sites in a protein result in time and spatial substitution rate heterogeneity beyond the capabilities of current models. Here, we exploit parallels between amino acid substitutions and chemical reaction kinetics to develop an improved theory of protein evolution. We constructed a mechanistic framework for modelling amino acid substitution rates that employs the formalisms of statistical mechanics, with population genetics principles underlying the analysis. Theoretical analyses and computer simulations of proteins under purifying selection for thermodynamic stability show that substitution rates and the stabilisation of resident amino acids (the 'evolutionary Stokes shift') can be predicted from biophysics and the effect of sequence entropy alone. Furthermore, we demonstrate that substitutions predominantly occur when epistatic interactions result in near neutrality; substitution rates are determined by how often epistasis results in such nearly neutral conditions. This theory provides a general framework for modelling protein sequence change under purifying selection, potentially explains patterns of convergence and mutation rates in real proteins that are incompatible with previous models, and provides a better null model for the detection of adaptive changes.

Keywords

protein evolution; epistasis; coevolution; transition state theory; statistical mechanics; substitution rate; evolutionary Stokes shift; entrenchment; contingency

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:http://www.nature.com/authors/editorial_policies/license.html#terms

*Correspondence to: David.Pollock@ucdenver.edu.

Author contributions
R.A.G. and D.D.P. jointly designed the study, analysed the results, and wrote the paper. R.A.G wrote the simulation software and performed all mathematical derivations.

Competing interests
The authors declare no competing financial interests.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Introduction

Proteins continuously change as mutations are fixed or eliminated depending on their effect on the protein's structure, stability, functionality, and intermolecular interactions. These holistic properties result from interactions between amino acids throughout the protein, inducing epistatic (non-additive) fitness interactions among mutations, and leading to long-term coevolution such that a substitution at one site alters the relative probability of substitutions at other sites1–9. Because of the complexity of epistatic interactions, it has been difficult to identify what determines substitution rates at a site, characterise how these rates depend on the rest of the sequence, and understand how they vary with time and location in the protein.

Standard empirical substitution rate models neglect epistatic interactions and the resultant rate heterogeneity beyond simple scaling factors10–12. Although these models have had a major impact throughout the life sciences, they cannot estimate the effect of epistatic interactions on structural stability, function or fitness, predict the role of compensatory substitutions in protein evolution13,14, predict which of the 10% of deleterious mutations in humans are harmless in other species15, or accurately represent the rate and time dependence of convergence and homoplasy14. More complicated empirical models have been developed16–21, but their utility is limited by the large number of required parameters. More mechanistic substitution models that represent the underlying process of molecular evolution hold the promise of increased accuracies using fewer biologically meaningful parameters, but require a deeper understanding of the process of sequence change, especially the characteristics and effects of epistasis.

We previously demonstrated that computational simulations of protein evolution, with fitness determined by thermodynamic stability, can reproduce many of the puzzling aspects of protein evolution including the rate- and time-dependence of convergence14 and the site- and time-dependence of substitution rates22. These models exhibit a phenomenon called the 'evolutionary Stokes shift'6, the tendency for newly resident amino acids at a site to be stabilised, or 'entrenched'8 over evolutionary time following a substitution. We also observed a tendency for the new amino acid to be pre-stabilised prior to the substitution by chance or contingency8,9. The pre- and post-adjustments of the protein to the new amino acid occur without corresponding changes in fitness, distinguishing this process from paired compensatory substitutions where the constituent substitutions have non-zero fitness effects23. Despite debate about the relationship between structural stability and fitness and the frequency of large reversals in amino acid propensities, the role of epistasis in protein evolution has become increasingly recognised1–9.

To understand how selection for structural stability determines amino acid substitution rates, and what drives the pre- and post-substitution structural stabilisations in the absence of changes in fitness, we developed a mechanistic framework employing the formalisms of statistical mechanics. We find that average substitution rates in proteins selected for structural stability can be explained by the evolutionary equivalent of transition state theory, with fluctuations in amino acid preferences due to epistatic interactions representing an essential aspect of the substitution process. Just as entropy plays a preeminent role in

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

statistical mechanics, the sequence entropy of folding, defined as the log of the number of possible sequences that fold with the evolutionarily determined degree of structural stability, is central to evolutionary mechanics. We test our mathematical approximations and predictions using computational simulations of protein evolution. We demonstrate that average substitution rates at a site can be predicted from site-specific structural stability distributions estimated in the absence of selection on that site and the dependence of the sequence entropy of folding on overall protein stability. The effect of other global factors such as effective population size, protein structure designability, and selective strength are combined in the entropy term and do not need to be considered or estimated separately.

Results

Site-specific stabilities and relative substitution rates

To develop a mechanical theory of protein evolution, we considered how purifying selection for stability determines site-specific substitution rates. To make the material more accessible, detailed equations underlying our results are in the Methods section; however, we endeavor here to provide an outline of the main concepts. Although real proteins are under selection for a range of properties, this specific form of selective pressure is well defined, theoretically tractable, and a common constraint. Analysing selection on stability also provides a 'null model' to examine other forms of selection. The stability $\Phi(\mathbf{X})$ of protein sequence $\mathbf{X}$ was defined as the negative of the free energy of folding, with more positive values indicating greater stability. The fitness $m(\mathbf{X}) = m(\Phi(\mathbf{X}))$ was set equal to the probability that the encoded protein would be folded at thermodynamic equilibrium (Equation (1))6,24,25. Thus, increases in stability correspond to increases in fitness.

Consider site k, occupied by amino acid α. The protein stability can be partitioned into two contributions $\Phi(\mathbf{X}) = \phi_{k,\alpha}(\mathbf{X}_{\not\ni k}) + \Phi_{k,\mathrm{Bath}}(\mathbf{X}_{\not\ni k})$. The first term includes interactions between the amino acid α resident at site k and those resident at other sites; in statistical mechanics, this term represents the system of interest. The second sequence-dependent term includes the much larger set of interactions amongst amino acids at all sites not including k, and corresponds to the thermodynamic bath in statistical mechanics. Both terms include interactions in the folded and unfolded states. For simplicity, we use Φ, ϕ_α and Φ_{Bath} to represent $\Phi(\mathbf{X})$, $\phi_{k,\alpha}(\mathbf{X}_{\not\ni k})$ and $\Phi_{k,\mathrm{Bath}}(\mathbf{X}_{\not\ni k})$, respectively.

In the low mutation rate regime, the instantaneous substitution rate from resident amino acid α to new amino acid β is equal to the corresponding mutation rate times the fixation probability, which in our model depends on the impact of the two amino acids on the protein's stability. We note that real and simulated proteins usually evolve within a narrow range of stabilities around an average value $\bar{\Phi}$24,26–29 where selection and the genetic drift of large numbers of slightly destabilising mutations balance30,31, dependent on factors such as temperature24 and effective population size. Under this condition, the change in fitness $\Delta m_{\alpha\to\beta}$, and therefore the probability of fixation, is determined by the difference between ϕ_α and ϕ_β.

Values of ϕ_α and ϕ_β will vary as substitutions occur in the rest of the protein, which will be affected by the amino acid occupying position k6. We assume other sites are sufficiently

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

numerous and change sufficiently rapidly that the protein is always equilibrated with the amino acid occupying site k. The joint probability distribution of ϕ_α and ϕ_β when site k is occupied by α can be described by the stationary distribution $\rho(\phi_\alpha, \phi_\beta|\alpha)$. The average substitution rate from α to β is completely determined by the mutation rate and this resident-dependent joint probability distribution. Because of its importance, we focus on characterising this joint distribution.

With our approximations, all sequences with a specific $\rho(\phi_\alpha, \phi_\beta|\alpha)$ have identical fitnesses. $\rho(\phi_\alpha, \phi_\beta|\alpha)$ is then proportional to the *number* of sequences with specific values of ϕ_α and ϕ_β; in analogy to Boltzmann's description of entropy as the log of the number of microscopic configurations corresponding to a specified macroscopic state, we characterise the log of the number of sequences corresponding to specific values of ϕ_α, ϕ_β, $x_k = \alpha$ and $\bar{\Phi}$ as the 'sequence entropy of folding' $S(\phi_\alpha, \phi_\beta|\alpha)$. This quantity is notably different from the 'sequence entropy' used to represent site-specific variability in a set of aligned sequences32.

To explore and evaluate our theoretical analysis, we simulated the evolution of a 300-residue protein with fitness equal to the probability of the protein being folded at thermodynamic equilibrium, matching our theoretical model. These simulations are not meant to make specific quantitative predictions, but rather to predict general characteristics of evolutionary behaviour for proteins that require a native confirmation to carry out their biological function, and have demonstrated their ability to reproduce fundamental aspects of protein evolution6,24,25. By using a simple pair-contact model of protein thermodynamics, we were able to perform replicate simulations corresponding to about 5 billion years given typical eukaryotic substitution rates.

We grouped sites with similar substitution patterns into four different site classes, with class 1 the most exposed and 4 the most buried. Figs. 1a-d illustrate $\rho_{\mathbb{C}}(\phi_{\mathrm{Glu}}, \phi_{\mathrm{Lys}}|\mathrm{Glu})$ and $\rho_{\mathbb{C}}(\phi_{\mathrm{Glu}}, \phi_{\mathrm{Lys}}|\mathrm{Lys})$, the joint probability distribution of site-specific contributions of glutamic acid and lysine when one or the other is resident, for sites belonging to site class $\mathbb{C} = \{1,2,3,4\}$. Corresponding distributions of population-scaled selective coefficients are shown in Supplementary Fig. S2. The distributions are broad, consistent with earlier results demonstrating that selective pressures vary over a wide range as substitutions occur elsewhere in the protein6. Exposed, rapidly evolving sites with few selective constraints (site class 1, Fig. 1a) have more compact distributions with smaller variances compared to buried, slowly evolving sites (site class 4, Fig. 1d). The potential contribution of an amino acid to protein stability is usually greater when that amino acid is resident at a site, a reflection of the 'evolutionary Stokes shift'6. The amount of this increase appears correlated with the variance in ϕ_α.

The bivariate distributions are surprisingly independent of population size (Figs. 1e-h, S2), but highly dependent on amino acid pair (Figs. 1i-l, S2). Distributions for physicochemically similar amino acids (e.g., glutamic acid versus aspartic acid, Fig. 1I) are highly correlated, while those for dissimilar amino acids (e.g., glutamic acid versus alanine, Fig. 1J) are anti-correlated. A non-resident amino acid is generally stabilised if the distributions are correlated (e.g. ϕ_{Glu} is positive when aspartic acid is present), but destabilised if the distributions are anti-correlated (e.g. ϕ_{Glu} is negative when alanine is present).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Predicting relative substitution rates

Substitution rates should be predictable from knowledge of $\boldsymbol{\rho(\phi_\alpha, \phi_\beta|\alpha)}$. To test this, we modelled $\boldsymbol{\rho(\phi_\alpha, \phi_\beta|\alpha)}$ as a bivariate normal distributions and numerically integrated over these distributions to calculate substitution rates33–35 (Equation (2)). There is extremely good agreement between expected substitution rates and those obtained by counting substitutions that occurred during simulations, as shown in Figs. 2a-c. The population size independence of the predicted (Figs. 2a-c) and observed substitution rates (Fig. S1b) matches previous observations36, and arises from our use of a concave-down fitness function (Equation 1)37.

We next investigated whether substitution rate calculations could be simplified by considering the dynamics of the substitution process. As described above, the values of ϕ_α and ϕ_β vary as the rest of the protein sequence changes. A part of the evolutionary trajectory before and after a glutamic acid to lysine substitution is shown in Fig. 3. Prior to substitution when glutamic acid is resident, glutamic acid is generally stabilised by the evolutionary Stokes shift, while lysine is slightly destabilised, reflecting the physicochemical differences between these two amino acids. The pattern is reversed after the substitution. Strikingly, substitutions occur in a narrow region along the diagonal $\phi_{Glu} = \phi_{Lys}$, where substitutions are nearly neutral (Figs. 1 and S2). These observations suggest the applicability of transition state theory (TST), a method for predicting the rate of chemical reactions38. TST focuses on how the energies of the reactants and products vary as the reactants undergo conformational fluctuations. The reaction occurs when the reactants are in a 'transition state' in which the energies of reactant and product are equal. The predicted reaction rate is equal to the probability that the reactants are in the transition state, times the conversion rate from reactants in the transition state to products.

Adapting this theory, the substitution rate from α to β was estimated from the probability that the protein is in the nearly neutral region (the 'transition state') times the rate of neutral substitution. The width of the neutral zone is approximately $\frac{2}{\gamma'}$ where γ describes γ the dependence of the number of sequences on protein stability ($\rho(\Phi) \sim e^{\gamma \Phi}$), which can be estimated by the relative numbers of destabilising and stabilising mutations. We obtained a closed-form expression for substitution rates (Equation (6)) that produces strikingly accurate substitution rate predictions (Fig. 2d-f). Notably, because this calculation considers only neutral substitutions, Kimura's fixation probability formula is no longer needed, greatly simplifying the calculations.

The equilibrium distributions of site-specific stabilities and the evolutionary 'Stokes Shift'

As described above, the rate of amino acid substitutions is determined by $\boldsymbol{\rho(\phi_\alpha, \phi_\beta|\alpha)}$ in the region where $\boldsymbol{\phi_\alpha \approx \phi_\beta}$. A full mechanistic description requires that we understand how these distributions, and therefore the substitution rates, are determined; we approach this goal using the principles of statistical mechanics and sequence entropy of folding.

$\rho(\phi_\alpha, \phi_\beta|\alpha)$ reflects the number of sequences corresponding to specified values of ϕ_α, ϕ_β, $x_k = \alpha$ and $\bar{\Phi}$. We approximate $\rho(\phi_\alpha, \phi_\beta|\alpha)$ by the product of two terms, $\rho_{Loc}(\phi_\alpha, \phi_\beta) \times$

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

$\rho_{Bath}(\Phi_{Bath} = \bar{\Phi} - \phi_\alpha)$. The local term $\rho_{Loc}(\phi_\alpha, \phi_\beta)$ represents the fraction of sequences with site-specific ϕ_α and ϕ_β, while the second term $\rho_{Bath}(\bar{\Phi} - \phi_\alpha)$ represents the fraction of sequences where the bath interactions provide sufficient contributions to the stability so that $\phi_\alpha + \Phi_{Bath} = \bar{\Phi}$. We approximated the first 'absolute' reference term by performing simulations in which a non-interacting amino acid ∅ was fixed at that site and all other sites were allowed to evolve under selection, as before. We then calculated the values of ϕ_α and ϕ_β that would result if amino acids α and β were substituted for ∅ in the resulting sequences. Interactions involving the focal amino acid represent a small fraction of total stability contributions, so the second term $\rho(\Phi_{Bath})$ was approximated by the distribution of protein sequences with total stability Φ, $\rho(\Phi) \sim e^{\gamma \Phi}$.

The product of these distributions suggests that the evolutionary Stokes shift alters the average value of ϕ_α by an amount $\zeta_{\alpha|\alpha} = \gamma \sigma^2_{\alpha|\varnothing}$, where $\sigma^2_{\alpha|\varnothing}$ is the variance in the distribution of $\rho(\phi_\alpha|\varnothing)$, while the variance itself is unaltered (Equation (7)). We can understand this shift by comparing the relative contributions of ϕ_α and Φ_{Bath} to $\bar{\Phi}$. Increasing values of ϕ_α decrease the value of Φ_{Bath} necessary to fulfill $\phi_\alpha + \Phi_{Bath} = \bar{\Phi}$. As the number of possible sequences increases rapidly with decreasing Φ_{Bath}, there is a strong bias towards increased values of ϕ_α. This stabilisation resulting from the large increase in sequence entropy of folding with decreasing Φ_{Bath} is precisely the evolutionary Stokes shift.

The predicted distributions of $\rho(\phi_\alpha, \phi_\beta|\alpha)$ versus those observed in thermodynamic simulations are shown in Figs. 1m-p. Estimated $\zeta_{\alpha|\alpha}$ values match observations surprisingly well given the approximations made (Fig. 4a-c). As predicted, the entropic stabilisation is approximately linear with $\sigma^2_{\alpha|\varnothing}$, with slope close to the estimated value of $\gamma = 1.26$ (kcal $mol^{-1})^{-1}$ (Fig. 4d-f), confirming the trends evident in Fig. 1. The observed entropic stabilisation is smaller than predicted for the largest shifts in the two slowest rate classes, involving the charged lysine, arginine, aspartic acid and glutamic acid. Earlier work demonstrated that equilibration for the most buried states can be extremely slow6, so deviations may result from insufficient time to adjust to the presence of the new amino acid.

In earlier work, we described how the evolutionary Stokes shift results in stabilisation of amino acids that are similar to the current resident, and destabilisation of amino acids that have large physicochemical differences. The basis of this effect, according to our theory, is that the presence of α at the site shifts values of ϕ_β by $\zeta_{\beta|\alpha} = \gamma \varphi_{\alpha\beta|\varnothing} \sigma_{\alpha|\varnothing} \sigma_{\beta|\varnothing}$, where $\varphi_{\alpha\beta|\varnothing}$ is the correlation between ϕ_α and ϕ_β in $\rho(\phi_\alpha, \phi_\beta|\varnothing)$; these shifts can be to higher or lower values depending on whether the physicochemical properties of the amino acids are similar or different (positive or negative $\varphi_{\alpha\beta|\varnothing}$, respectively), increasing or decreasing the density of the distribution in the region $\phi_\alpha \approx \phi_\beta$ and the corresponding substitution rate. Substitution rates estimated with the TST approximation (Equation (6)) and site-specific stabilities calculated using Equation (7) are remarkably accurate for all four site classes over four orders of magnitude (Fig. 2g-i).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Discussion

The probability of fixation of an amino acid-altering mutation in proteins selected for stability depends on the relative contributions to stability made by the resident (ϕ_α) and mutant amino acid (ϕ_β), and the effect of the resulting stability change ($\phi_\beta - \phi_\alpha$) on organismal fitness33–35. The situation is complicated by epistatic interactions connecting the mutating site to other sites throughout the protein, resulting in fluctuations in these contributions and corresponding fixation probabilities as substitutions occur throughout the protein. Surprisingly, this apparent complication leads to a simplification; substitutions that occur when stability contributions are similar dominate the evolutionary process, and the nearly neutral zone can be modeled using transition state theory. This represents a fundamental change in understanding of substitution rates; the focus shifts from estimating the fitness change resulting from a substitution to calculating the fraction of the time that substitution is nearly neutral. Fluctuations in stability contributions cannot be ignored because they create the conditions under which substitutions occur.

The frequency of nearly neutral conditions depends on the joint distribution of ϕ_α and ϕ_β. Amino acids with similar physicochemical properties make correlated contributions to stability, increasing the probability of near-neutrality resulting in higher rates of conservative change, a phenomenon first described by Fisher39. The multiplicity of interactions at buried sites increase the variances of ϕ_α and ϕ_β, reducing the probability of near neutrality and thus the substitution rate (Figs. 1a-1d), consistent with observed slower substitution rates observed at internal sites.

The joint stability distribution is affected by the tendency for the resident amino acid (and similar amino acids) to be stabilised by substitutions at other sites, the ‘evolutionary Stokes shift’6. This shift can be understood purely in terms of biophysics and the sequence entropy of folding; increases in the stabilising contributions of resident amino acids reduce the amount of stabilisation required from interactions amongst the remaining amino acids (the ‘bath’). Because more sequences are able to fulfill this reduced stabilisation requirement, the contributions of the bath to the sequence entropy of folding is larger, and higher affinities for the resident amino acid are entropically favoured. Although describable as an adjustment, this evolutionary mechanism can be fully reversible, as are the simulations described here, with similar processes of moving into and away from the neutral zone6. These processes, called ‘contingency’ and ‘entrenchment’ by Plotkin and colleagues8, are mirrors of each other, so that a tape of the dissipation (entrenchment) process, played backwards, would have the same statistical properties as the pre-adaptation (contingency) process played forwards.

The predicted average substitution rates for sites under purifying selection for stability can be estimated solely based on the mutation rate and the joint distribution of ϕ_α and ϕ_β, with *no adjustable parameters*. Thus, when we show that, as long as the assumptions and approximations of the analysis remain valid, $\rho(\phi_\alpha, \phi_\beta)$ depends only on details of the protein structure and function (which affect $\rho(\phi_\alpha|\varnothing)$ and γ), we can infer that Kimura’s formula is not required to predict and explain substitution rates amongst amino acids. Although evolutionary mechanics theory fully incorporates population genetics theory, substitution

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

rates and the evolutionary Stokes shift do not depend on population size (Figure 2, Figure 4), despite its centrality to Kimura's formulation.

Here, we addressed only the theoretical predictions and simulations near equilibrium. Some discrepancies between the predicted and observed Stokes shifts for charged residues in buried sites, however, may be explained by inadequate time for equilibration. Individual sites at specific time points may be constrained by conserved neighbouring sites as well as conserved structural features. Such effects may influence the time-dependent probability of back and subsequent substitutions, an important topic for investigation. The interaction of fluctuating selection and fluctuating population size also requires further investigation.

We have considered purifying selection with fixed population size and fitness based purely on folding stability. We and others have previously shown that evolutionary simulations based on protein thermodynamics produce patterns of epistasis, convergence, and entrenchment that are qualitatively similar to patterns in real proteins4–7,14,40; we now have a clear explanation why these patterns may have been produced, from statistical mechanics considerations. Selection for other properties involving contributions from multiple amino acid sites would define their own nearly neutral landscape. We argue that other forms of constant selection such as interactions with other proteins, ligand binding, catalysis, and avoiding proteolysis and aggregation should restrict the number of acceptable sequences but not otherwise affect the theory or calculations. Other selective pressures may force occasionally adaptive, non-neutral substitutions when external pressures change. An evolutionary Stokes shift would still occur following such a forced substitution, but the process would no longer be reversible6. The simple theory described here provides an improved conceptual basis to understand what would happen in the absence of further complications, and should allow more confident prediction of non-structural functional constraints, adaptation, and fluctuating population sizes when they do occur. Analogous models can be applied to other forms of selection acting at higher levels such as development41.

The work described here establishes a theory of evolutionary mechanics, and simulations demonstrate that this theory can be used to predict substitution rates from the basic properties of how amino acids interact. Assuming that most proteins under constant selective pressures are dominated by thermodynamic (or similar) processes, we provide a mechanistic explanation for the known predominance of nearly neutral evolution and a better understanding of what happens during purifying selection. As with statistical mechanics and thermodynamics, the theory of evolutionary mechanics allows us to connect the microscopic events of evolutionary mechanics (mutation rates, fitness differences, and fixation probabilities) with the macroscopic events of molecular evolution (relative rates of substitution, and distributions of fluctuating rates across sequences and over time).

Methods

Simulations of protein evolution

The methods used to simulate protein evolution have been described previously6,24,25. Our simulations modelled proteins evolving under selection for a common requirement for

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

globular proteins, stability of the native conformation. The free energy $\boldsymbol{G}(\mathbf{X}, \mathbf{r})$ of a protein sequence $\mathbf{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3} \ldots \boldsymbol{x_n}\}$ in conformation $\mathbf{r}$ was calculated by summing the pairwise energies of amino acids in contact in that conformation, using the contact potentials derived by Miyazawa and Jernigan42. The free energy of folding $\boldsymbol{\Delta G}_{\text{Folding}}(\mathbf{X})$ was computed by first determining the free energy of the sequence in a pre-chosen native state, the conformation of the 300-residue purple acid phosphatase, PDB 1QHW43. The energies of unfolded states were assumed to follow a Gaussian distribution; the parameters characterising this distribution were estimated by calculating the free energies of the sequence in a widely diverse set of 55 different protein structures. The energy of the unfolded state was then calculated by assuming a large set (10^{160}) of possible unfolded structures with free energies drawn from this distribution. The free energy of folding $\boldsymbol{\Delta G_{\text{Folding}}}(\mathbf{X})$ was calculated as the expected difference between the two, and stability was $\boldsymbol{\Phi}(\mathbf{X}) = -\boldsymbol{\Delta G_{\text{Folding}}}(\mathbf{X})$. The Malthusian fitness of a sequence $\boldsymbol{m}(\mathbf{X})$ was defined as the fraction of that sequence that would be folded to the native state at equilibrium

$$m\left(\Phi\left(\mathbf{X}\right)\right) = \frac{\exp\left(\frac{\Phi(\mathbf{X})}{T}\right)}{1+\exp\left(\frac{\Phi(\mathbf{X})}{T}\right)} \quad (1)$$

where T is temperature in units of energy, 0.6 kcal mol^{-1}.

The simulations implemented a Gillespie algorithm44 representing the evolution of a protein in the low mutation rate limit where the monomorphic population is represented by a single sequence. Starting from a single randomly chosen nucleotide sequence encoding a 300 amino-acid protein, we simulated evolution by considering in each step all possible nucleotide mutations with rates given by the K80 nucleotide model ($\boldsymbol{\kappa} = 2$)45. The fixation probability of each mutation was calculated based on the Kimura formula for diploid organisms33–35,

$$P_{\text{Fix}}\left(\mathbf{X}, \mathbf{X}'\right) = \frac{1 - \mathrm{e}^{-2\left(m\left(\Phi\left(\mathbf{X}'\right)\right) - m(\Phi(\mathbf{X}))\right)}}{1 - \mathrm{e}^{-4N_e\left(m\left(\Phi\left(\mathbf{X}'\right)\right) - m(\Phi(\mathbf{X}))\right)}} \quad (2)$$

where $\mathbf{X}$ and $\mathbf{X}'$ are the sequences before and after the mutation. The total substitution rate was set equal to the product of the mutation rate times the fixation probability, summed over all possible mutations. At each step, the evolutionary time was advanced by an amount chosen from an exponential distribution based on the total substitution rate, and one substitution was chosen to be fixed at random with relative probabilities determined by the product of the mutation rates times the acceptance probabilities.

Sequence evolution was simulated for a sufficient number of generations such that protein stability was roughly constant, representing mutation-selection-drift selection balance. 100 equilibrated proteins were chosen, and two longer simulations were performed using each these equilibrated proteins as initial starting sequences, for a total of 200 simulations. The evolution of each lineage was simulated for an evolutionary distance of approximately seven

amino acid replacements per amino acid position. The sequence and energy were sampled at regular time intervals.

Grouping of sites

For ease of analysis, we divided protein sites into four classes with similar substitution rates. Substitution matrices were calculated individually for each site; due to the length of simulations, we had on average over 1400 substitutions at each site. Sites were then clustered based on the off-diagonal elements of the substitution matrices using K-means clustering46,47. The resulting clusters were approximately equal in size, and class membership strongly depended on how buried or exposed sites were in the native state (as indicated by number of contacts). We ranked clusters by surface exposure, where class 1 is the most exposed and 4 is the most buried.

Calculating the site-specific contribution to protein stability

The site-specific contribution $\phi_{k,\alpha}(\mathbf{X}_{\not\ni k})$ of amino acid α at focal site k as a function of the amino acids $\mathbf{X}_{\not\ni k}$ at all sites excluding k is equal to $\Phi\{x_1, x_2, x_3 \ldots x_{k-1}, \alpha, x_{k+1} \ldots x_n\}$, the stability when the focal site is occupied by α, minus $\Phi\{x_1, x_2, x_3 \ldots x_{k-1}, \varnothing, x_{k+1} \ldots x_n\}$, the stability of a reference state when α is replaced by a non-interacting amino acid ∅, with the rest of the sequence and thus all other interactions unchanged. The part of the stability unaffected by this replacement is represented by the 'bath' interactions $\Phi_{k,\mathrm{Bath}}(\mathbf{X}_{\not\ni k}) = \Phi(\mathbf{X}) - \phi_{k,\alpha}(\mathbf{X}_{\not\ni k})$ so that $\Phi(\mathbf{X}) = \phi_{k,\alpha}(\mathbf{X}_{\not\ni k}) + \Phi_{k,\mathrm{Bath}}(\mathbf{X}_{\not\ni k})$.

Determining the change in fitness

Prior to a mutation, when amino acid α is resident, the protein stability is equal to $\Phi = \phi_\alpha + \Phi_{\mathrm{Bath}}$ with corresponding fitness $m(\Phi)$, given by Equation 1 (Methods). After a mutation to amino acid β, the stability is equal to $\Phi' = \phi_\beta + \Phi_{\mathrm{Bath}} = \Phi + \phi_\beta - \phi_\alpha$, corresponding to fitness $m(\Phi') = m(\Phi + \phi_\beta - \phi_\alpha)$, where we have used the fact that Φ_{Bath} is unchanged by the mutation. The situation is complicated by the non-linear relationship between fitness and stability (Equation 1, Methods), but can be greatly simplified by noting that real proteins, as well as proteins from this and other evolutionary simulations under purifying selection for thermostability, evolve within a narrow range of stability values around an average value $\bar{\Phi}$24,26–29; see Supplementary Fig. S1. This narrow stability range occurs where the effectiveness of selection for greater stability is balanced by large numbers of slightly destabilising mutations fixed by genetic drift30,31. We therefore approximate the protein's stability prior to the mutation as equal to $\Phi = \bar{\Phi}$; the resulting change in fitness is then equal to $\Delta m_{\alpha\to\beta} = m(\bar{\Phi} + \phi_\beta - \phi_\alpha) - m(\bar{\Phi})$. The value of $\bar{\Phi}$ depends on factors such as temperature24, effective population size (as shown in24 and Fig. S1), and protein structure and function, but will be constant as long as these factors are approximately constant. With these assumptions, the change in fitness and thereby the probability of fixation of the mutation is therefore determined by the difference between the current values of ϕ_α and ϕ_β.

While the total stability value of $\bar{\Phi}$ is a constant, the manner in which this stability is distributed amongst the various interactions, and therefore the values of ϕ_α and ϕ_β as well as the corresponding substitution rate, will vary as substitutions occur along the rest of the protein sequence. The nature of this variation depends on which amino acid occupies

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

position k because that amino acid affects the evolution in the rest of the protein6. In order to compute the estimated average substitution rate, we assume that the other sites are sufficiently numerous and change sufficiently rapidly that the protein is always fully adjusted to the current amino acid at site k. (This assumption is most likely to break down following non-conservative substitutions, as discussed below.) The joint probability distribution of ϕ_α and ϕ_β given total stability $\bar{\Phi}$ and the occupation of site k by amino acid α can then be described by the stationary distribution $\rho(\phi_\alpha, \phi_\beta|\Phi(\mathbf{X}) = \bar{\Phi}, x_k = \alpha)$, which we simplify to $\rho(\phi_\alpha, \phi_\beta|\alpha)$.

Calculating the substitution rate integrating over distributions of local contributions

The average rate for substitution $a \rightarrow \beta$ at site k, $Q_{k,a \rightarrow \beta}$, is equal to the neutral substitution rate $\upsilon_{a \rightarrow \beta}$ times the average probability of fixation, which is a function of the stability of the protein before and after the substitution. The standard deviation of observed values of protein stabilities Φ, for example 0.71 kcal mol^{-1} for population size N_e set to 10^6, was small compared with the range of values of $\boldsymbol{\phi}_{k,\alpha}(\mathbf{X}_{\not\ni k})$, allowing us to represent the distribution $\boldsymbol{\Phi}$ by its average, $\boldsymbol{\Phi} \simeq \bar{\boldsymbol{\Phi}} = 9.26$ kcal mol^{-1}. We assumed that the stability before the substitution was $\bar{\boldsymbol{\Phi}}$ and afterwards was $\bar{\boldsymbol{\Phi}} + (\boldsymbol{\phi}_{k,\beta}(\mathbf{X}_{\not\ni k}) - \boldsymbol{\phi}_{k,\alpha}(\mathbf{X}_{\not\ni k}))$. The average substitution rate was then estimated as

$$Q_{k,\alpha\rightarrow\beta}=2N_e\nu_{\alpha\rightarrow\beta}\int\int\frac{1-e^{-2\Delta m\left(\phi_{k,\alpha},\phi_{k,\beta}\right)}}{1-e^{-4N_e\Delta m\left(\phi_{k,\alpha},\phi_{k,\beta}\right)}}\rho\left(\phi_{k,\alpha},\phi_{k,\beta}|x_k=\alpha\right)\mathrm{d}\phi_{k,\alpha}\,\mathrm{d}\phi_{k,\beta}. \quad (3)$$

where $\Delta m(\phi_{k,\alpha}, \phi_{k,\beta}) = m(\bar{\Phi} + (\phi_{k,\beta}, -\phi_{k,\alpha})) - m(\bar{\Phi})$ and $\rho(\phi_{k,\alpha}, \phi_{k,\beta}|x_k = \alpha)$ is the joint distribution of $\phi_{k,\alpha}(\mathbf{X}_{\not\ni k})$ and $\phi_{k,\beta}(\mathbf{X}_{\not\ni k})$ for the equilibrium distribution of sequences $\mathbf{X}_{\not\ni k}$ when α occupies site k, and we use the fixation probabilities assuming small values of $2N\Delta m$.

Based on observations in Fig. 1, $\rho(\phi_{k,\alpha}, \phi_{k,\beta}|x_k = \alpha)$ was modeled as a bivariate normal distribution of the form $\rho\left(\phi_{k,\alpha},\phi_{k,\beta}|x_k=\alpha\right)=\mathcal{N}\left(\overline{\phi}_{k,\alpha|\alpha},\overline{\phi}_{k,\beta|\alpha},\sigma^2_{k,\alpha|\alpha},\sigma^2_{k,\beta|\alpha},\varphi_{k,\alpha\beta|\alpha}\right)$. Parameters were calculated directly from evolutionary simulation, and Equation (3) integrated numerically. The neutral substitution rate was calculated using the same K80 nucleotide model ($\kappa = 2$)45 as used in the simulation, with all non-nonsense codons considered equally likely.

Calculating the substitution rate integrating assuming only neutral substitutions

As observed in Fig. 1, substitutions generally occur in a neutral region in which $\boldsymbol{\Delta\Phi}_{k,\alpha\rightarrow\beta} = \boldsymbol{\phi}_{k,\beta}(\mathbf{X}_{\not\ni k}) - \boldsymbol{\phi}_{k,\alpha}(\mathbf{X}_{\not\ni k}) \approx \mathbf{0}$, so that

$$\frac{1-e^{-2\Delta m\left(\phi_{k,\alpha},\phi_{k,\beta}\right)}}{1-e^{-4N_e\Delta m\left(\phi_{k,\alpha},\phi_{k,\beta}\right)}}\approx\frac{1}{2N_e}. \quad (4)$$

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

This condition is satisfied in a band of width 2ε centred on $\phi_{k,\beta}(\mathbf{X}_{\not\ni k}) - \phi_{k,\alpha}(\mathbf{X}_{\not\ni k})$, where ε represents a deviation from strict neutrality that is sufficiently close for Equation (4) to be sufficiently accurate.

A natural scale for ε was obtained by considering the 'free fitness' Γ(Φ) of the protein equal to $\Gamma(\Phi) = m(\Phi) + \frac{S(\Phi)}{4N_e}$ where $S(\Phi)$ is the sequence entropy of folding, equal to the log of the number of sequences corresponding to a given total stability Φ48,49. Free fitness is analogous to thermodynamic free energy but with temperature T replaced by $4N_e$, and encompasses contributions from both fitness and sequence entropy to determine the distribution of states; evolutionary dynamics moves towards maximising this quantity. As the stability represents the sum of many small interactions, we would expect the distribution of stabilities to obey the central limit theorem and to resemble a Gaussian distribution. We are, however, on the tail of the distribution where the Gaussian is indistinguishable from an exponential, with one additional unidentifiable parameter. We instead assume $S(\Phi) = \ln(\Omega_0 e^{-\gamma\Phi})$ where Ω_0 is constant. Noting that the system is at equilibrium with $\frac{\partial\Gamma(\Phi)}{\partial\Phi} = 0$ when Φ = $\bar{\Phi}$, it can be demonstrated that a good approximation at the mode is

$$\frac{\partial\, 4N_e m(\Phi)}{\partial\Phi}\Big|_{\Phi=\bar{\Phi}} = \gamma \qquad (5)$$

Thus, γ defines the rate of change of the population-weighted fitness $4N_e m(\Phi)$ with stability. Alternatively, a change in stability of $\frac{1}{\gamma}$ corresponds to a unit change in population-weighted fitness. In our calculations, we equated $\varepsilon = \frac{1}{\gamma}$; the estimation of γ is described below. Note that this calculation demonstrates that ε is, surprisingly, independent of effective population size N_e. This is a result of the balance between selection and mutational drift at equilibrium; for fixed effect of mutational drift, the degree of selection $(\frac{\partial m(\Phi)}{\partial\Phi})$ adjusts to changes in effective population size so that their product is constant36,37.

If we assume that $\rho(\phi_{k,\alpha}, \phi_{k,\beta}|x_k = \alpha)$ is broader than ε, and that Equation (4) is satisfied, Recalling that the marginal probability density for a multivariate normal distribution is still normal, Equation (3) becomes

$$Q^{\mathrm{TST}}_{k,\alpha\to\beta} = 2\varepsilon\, \upsilon_{\alpha\to\beta} \iint \rho(\phi_{k,\alpha}, \phi_{k,\beta}|x_k=\alpha)\, \delta(\phi_{k,\alpha} - \phi_{k,\beta})\, d\phi_{k,\alpha} d\phi_{k,\beta}$$
$$= \frac{\upsilon_{\alpha\to\beta}}{\gamma} \frac{\exp\left(-\frac{(\bar{\phi}_{k,\alpha|\alpha} - \bar{\phi}_{k,\beta|\alpha})^2}{2(\sigma^2_{k,\alpha|\alpha} + \sigma^2_{k,\beta|\alpha} - 2\varphi_{k,\alpha\beta|\alpha}\sigma_{k,\alpha|\alpha}\sigma_{k,\beta|\alpha})}\right)}{\sqrt{2\pi(\sigma^2_{k,\alpha|\alpha} + \sigma^2_{k,\beta|\alpha} - 2\varphi_{k,\alpha\beta|\alpha}\sigma_{k,\alpha|\alpha}\sigma_{k,\beta|\alpha})}} \qquad (6)$$

where $\delta(\phi_{k,\alpha} - \phi_{k,\beta})$ is the Dirac delta function.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

For highly similar amino acids the entire distribution of $\rho(\phi_{k,\alpha}, \phi_{k,\beta}|x_k = \alpha)$ may be contained in a region significantly narrower than the neutral zone, resulting in an overestimation of $Q_{k,\alpha\to\beta} > \upsilon_{\alpha\to\beta}$. For this reason, the estimated rate was capped at the neutral rate $\upsilon_{\alpha\to\beta}$.

Estimating $\rho(\phi_{k,\alpha}, \phi_{k,\beta})$

As described in the Results section, we approximate $\rho(\phi_{k,\alpha}, \phi_{k,\beta}|x_k = \alpha)$ as the product of two terms, $\rho_{\text{Loc}}(\phi_{k,\alpha}, \phi_{k,\beta}) \times \rho_{\text{Bath}}(\Phi_{k,\text{Bath}} = \bar{\Phi} - \phi_{k,\alpha})$, where $\rho_{\text{Loc}}(\phi_{k,\alpha}, \phi_{k,\beta})$ represents the fraction of sequences with given values of $\phi_{k,\alpha}$ and $\phi_{k,\beta}$ independently of how the rest of the protein adjusts to the current amino acid resident at site k, while $\rho_{\text{Bath}}(\Phi_{k,\text{Bath}} = \bar{\Phi} - \phi_{k,\alpha})$, represents the fraction of sequences where the bath interactions contribute sufficiently to the stability so that $\phi_{k,\alpha} + \Phi_{k,\text{Bath}} = \bar{\Phi}$.

$\rho_{\text{Loc}}(\phi_{k,\alpha}, \phi_{k,\beta})$ was approximated by $\rho(\phi_{k,\alpha}, \phi_{k,\beta}|x_k = \varnothing)$, the observed distribution observed when site k was occupied by a non-interacting amino acid $\varnothing$. We assumed that the contribution to the stability was small and approximated the distribution of Bath contributions with the distribution of total protein stabilities, $\rho_{\text{Bath}}(\Phi_{k,\text{Bath}} = \bar{\Phi} - \phi_{k,\alpha}) \simeq \rho_{\Phi}(\Phi_{k,\text{Bath}} = \bar{\Phi} - \phi_{k,\alpha}) \propto \exp(-\gamma(\bar{\Phi} - \phi_{k,\alpha}))$.

Because the number of possible sequences is immense, and because $\phi_{k,\alpha}$ and $\phi_{k,\beta}$ are the result of many interactions, the central limit theorem suggests that $\rho(\phi_{k,\alpha}, \phi_{k,\beta}|x_k = \varnothing)$ can be approximated by a bivariate normal distribution $\rho(\phi_{k,\alpha},\phi_{k,\beta}|x_k=\varnothing) \propto \mathcal{N}\left\{\overline{\phi}_{k,\alpha}|_{\varnothing},\overline{\phi}_{k,\beta}|_{\varnothing},\sigma^2_{k,\alpha}|_{\varnothing},\sigma^2_{k,\beta}|_{\varnothing},\varphi_{k,\alpha\beta}|_{\varnothing}\right\}$. The normalised product of $\rho(\phi_{k,\alpha}, \phi_{k,\beta}|x_k = \varnothing)$ and $\rho_{\text{Bath}}(\Phi_{k,\text{Bath}} = \bar{\Phi} - \phi_{k,\alpha}) \propto \exp(-\gamma(\bar{\Phi} - \phi_{k,\alpha}))$ results in an estimated shifted bivariate normal distribution $\tilde{\rho}_{k,\alpha}(\phi_{k,\alpha},\phi_{k,\beta}) = \mathcal{N}\left\{\tilde{\phi}_{k,\alpha}|_{\alpha},\tilde{\phi}_{k,\beta}|_{\alpha},\tilde{\sigma}^2_{k,\alpha}|_{\alpha},\tilde{\sigma}^2_{k,\beta}|_{\alpha},\tilde{\varphi}_{k,\alpha\beta}\right\}$ with

$$\begin{aligned}
\tilde{\phi}_{k,\alpha}|_{\alpha} &= \tilde{\phi}_{\text{k}\alpha}|_{\varnothing,} + \gamma\sigma^2_{k,\alpha}|_{\varnothing} \\
\tilde{\sigma}^2_{\text{k},\alpha}|_{\alpha} &= \sigma^2_{k,\alpha}|_{\varnothing} \\
\tilde{\phi}_{\text{k},\beta}|_{\alpha} &= \overline{\phi}_{\text{k},\beta}|_{\varnothing} + \gamma\varphi_{k,\alpha\beta}|_{\varnothing}\sigma_{k,\alpha}|_{\varnothing}\sigma_{k,\beta}|_{\varnothing} \\
\tilde{\sigma}^2_{k,\beta}|_{\alpha} &= \sigma^2_{k,\beta}|_{\varnothing} \\
\tilde{\varphi}_{k,\alpha\beta}|_{\alpha} &= \varphi_{k,\alpha\beta}|_{\varnothing}
\end{aligned} \tag{7}$$

Substituting these results into Equation (6) yields

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

$$Q_{k,\alpha\to\beta}^{\mathrm{TST},\varnothing}=\frac{\upsilon_{\alpha\to\beta}}{\gamma}\frac{exp\left(-\frac{\left(\bar{\phi}_{k,\alpha|\varnothing}-\bar{\phi}_{k,\beta|\varnothing}+\gamma\sigma^2_{k,\alpha|\varnothing}\left(1-\varphi_{k,\alpha\beta|\varnothing}\frac{\sigma_{k,\beta|\varnothing}}{\sigma_{k,\alpha|\varnothing}}\right)\right)^2}{2(\sigma^2_{k,\alpha|\varnothing}+\sigma^2_{k,\beta|\varnothing}-2\varphi_{k,\alpha\beta|\varnothing}\sigma_{k,\alpha|\varnothing}\sigma_{k,\beta|\varnothing})}\right)}{\sqrt{2\pi(\sigma^2_{k,\alpha|\varnothing}+\sigma^2_{k,\beta|\varnothing}-2\varphi_{k,\alpha\beta|\varnothing}\sigma_{k,\alpha|\varnothing}\sigma_{k,\beta|\varnothing})}} \quad (8)$$

Characterising the bath state distribution

As described above, we assume that the number of protein sequences with a given value of Φ in the range of interest around $\Phi = \bar{\Phi}$ is approximately exponential $\Omega(\Phi) \sim e^{-\gamma\Phi}$. We estimated γ from the average change in stability resulting from random mutations, $\langle\rho_{mut}(\Delta\Phi)\rangle$, which is negative due to the greater number of sequences coding for proteins with lower stability. This suggests that by correcting for the dependence of Ω on Φ by multiplying $\rho_{mut}(\Delta\Phi)$ and $e^{\gamma\Delta\Phi}$, this bias would disappear. We adjusted γ so that $\langle\Delta\Phi e^{\gamma\Delta\Phi}\rangle = 0$ where the average was over all possible mutations during the simulations with N_e equal to 10^6, yielding $\gamma = 1.26$ (kcal mol^{-1})$^{-1}$.

The bath state distribution determines the equilibrium stabilities through Equation (5).

Substituting Equation (1) into Equation (5) yields $\bar{\Phi} \approx T \ln\left(\frac{4N_e}{\gamma T}\right)$. This expression results in estimations for $\bar{\Phi}$ of 6.53, 9.26, and 12.05 for N_e equal to 10^4, 10^6, and 10^8, respectively. These agree well with the average of the distributions shown in Supplementary Figure S1: 6.40, 9.15, and 11.90.

We note that under this model, the population scaled fixed load $2N_e(1 - m(\bar{\Phi}))$ is equal to

$$2N_e\left(1-m(\bar{\Phi})\right)=2N_e\left(\frac{1}{1+\left(\frac{4N_e}{\gamma T}\right)}\right)\approx\frac{\gamma T}{2} \quad (9)$$

that is, it only depends on the dependence of the sequence entropy on the stability and the temperature. For our system, $2N_e(1 - m(\bar{\Phi})) \approx 0.38$.

Data and Code Availability

Data, including structures used, contact potentials, tables of outcomes, and raw program data output, is available on Dryad (doi:10.5061/dryad.7b8vb). All simulations and analysis software is available on GitHub (https://github.com/EvolutionaryMechanics/Goldstein-Pollock-2017).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Acknowledgments

We thank Bhavin Khatri for helpful discussions. We acknowledge the support of the Medical Research Council (UK) (MC_U117573805) and the Biotechnology and Biological Sciences Research Council (UK) (BB/P007562/1) to RAG and the National Institutes of Health (NIH; GM083127 and GM097251) to DDP.

References

1. Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. Epistasis as the primary factor in molecular evolution. Nature. 2012; 490:535–538. DOI: 10.1038/nature11510 [PubMed: 23064225]
2. Usmanova DR, Ferretti L, Povolotskaya IS, Vlasov PK, Kondrashov FA. A model of substitution trajectories in sequence space and long-term protein evolution. Mol Biol Evol. 2015; 32:542–554. DOI: 10.1093/molbev/msu318 [PubMed: 25415964]
3. Sarkisyan KS, et al. Local fitness landscape of the green fluorescent protein. Nature. 2016; 533:397–401. DOI: 10.1038/nature17995 [PubMed: 27193686]
4. Ashenberg O, Gong LI, Bloom JD. Mutational effects on stability are largely conserved during protein evolution. Proc Natl Acad Sci USA. 2013; 110:21071–21076. DOI: 10.1073/pnas.1314781111 [PubMed: 24324165]
5. Gong LI, Bloom JD. Epistatically interacting substitutions are enriched during adaptive protein evolution. PLoS Genet. 2014; 10:e1004328.doi: 10.1371/journal.pgen.1004328 [PubMed: 24811236]
6. Pollock DD, Thiltgen G, Goldstein RA. Amino acid coevolution induces an evolutionary Stokes shift. Proc Natl Acad Sci USA. 2012; 109:E1352–1359. DOI: 10.1073/pnas.1120084109 [PubMed: 22547823]
7. Pollock DD, Goldstein RA. Strong evidence for protein epistasis, weak evidence against it. Proc Natl Acad Sci USA. 2014; 111:E1450.doi: 10.1073/pnas.1401112111 [PubMed: 24706894]
8. Shah P, McCandlish DM, Plotkin JB. Contingency and entrenchment in protein evolution under purifying selection. Proc Natl Acad Sci USA. 2015; 112:E3226–3235. DOI: 10.1073/pnas.1412933112 [PubMed: 26056312]
9. Pollock DD, Taylor WR, Goldman N. Coevolving protein residues: maximum likelihood and relationship to structue. J Mol Biol. 1999; 287:187–198. [PubMed: 10074416]
10. Muse SV, Gaut BS. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol. 1994; 11:715–724. [PubMed: 7968485]
11. Nielsen R, Yang Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 1998; 148:929–936. [PubMed: 9539414]
12. Tamuri AU, dos Reis M, Hay AJ, Goldstein RA. Identifying Changes in Selective Constraints: Host Shifts in Influenza. Plos Comput Biol. 2009; 5:e1000564. doi:ArtnE1000564. doi: 10.1371/Journal.Pcbi.1000564 [PubMed: 19911053]
13. Castoe TA, et al. Evidence for an ancient adaptive episode of convergent molecular evolution. Proc Natl Acad Sci USA. 2009; 106:8986–8991. [PubMed: 19416880]
14. Goldstein RA, Pollard ST, Shah SD, Pollock DD. Nonadaptive Amino Acid Convergence Rates Decrease over Time. Mol Biol Evol. 2015; 32:1373–1381. DOI: 10.1093/molbev/msv041 [PubMed: 25737491]
15. Kondrashov AS, Sunyaev S, Kondrashov FA. Dobzhansky-Muller incompatibilities in protein evolution. Proc Natl Acad Sci USA. 2002; 99:14878–14883. DOI: 10.1073/pnas.232565499 [PubMed: 12403824]
16. Halpern AL, Bruno WJ. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol. 1998; 15:910–917. [PubMed: 9656490]
17. Lartillot N, Philippe H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 2004; 21:1095–1109. [PubMed: 15014145]
18. Tamuri AU, dos Reis M, Goldstein RA. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. Genetics. 2012; 190:1101–1115. DOI: 10.1534/genetics.111.136432 [PubMed: 22209901]

Europe PMC Funders Author Manuscripts

19. Tamuri AU, Goldman N, dos Reis M. A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. Genetics. 2014; 197:257–271. DOI: 10.1534/genetics.114.162263 [PubMed: 24532780]
20. Rodrigue N. On the statistical interpretation of site-specific variables in phylogeny-based substitution models. Genetics. 2013; 193:557–564. DOI: 10.1534/genetics.112.145722 [PubMed: 23222651]
21. Spielman SJ, Wilke CO. Extensively Parameterized Mutation-Selection Models Reliably Capture Site-Specific Selective Constraint. Mol Biol Evol. 2016; 33:2990–3002. DOI: 10.1093/molbev/msw171 [PubMed: 27512115]
22. Goldstein RA, Pollock DD. The tangled bank of amino acids. Protein Sci. 2016; doi: 10.1002/pro.2930
23. Kimura M. The role of compensatory neutral mutations in molecular evolution. Journal of Genetics. 1985; 64doi: 10.1007/BF02923549
24. Goldstein RA. The evolution and evolutionary consequences of marginal thermostability in proteins. Proteins. 2011; 79:1396–1407. DOI: 10.1002/prot.22964 [PubMed: 21337623]
25. Williams PD, Pollock DD, Blackburne BP, Goldstein RA. Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput Biol. 2006; 2:e69.doi: 10.1371/journal.pcbi.0020069 [PubMed: 16789817]
26. Privalov PL. Stability of proteins: small globular proteins. Adv Protein Chem. 1979; 33:167–241. [PubMed: 44431]
27. Privalov PL, Gill SJ. Stability of protein-structure and hydrophoboc interaction. Advances in Protein Chemistry. 1988; 39:191–234. [PubMed: 3072868]
28. Taverna DM, Goldstein RA. Why are proteins marginally stable? Proteins. 2002; 46:105–109. [PubMed: 11746707]
29. Zeldovich KB, Shakhnovich EI. Understanding protein evolution: from protein physics to Darwinian selection. Annu Rev Phys Chem. 2008; 59:105–127. DOI: 10.1146/annurev.physchem.58.032806.104449 [PubMed: 17937598]
30. Iwasa Y. Free fitness that always increases in evolution. J Theor Biol. 1988; 135:265–281. [PubMed: 3256719]
31. Sella G, Hirsh AE. The application of statistical physics to evolutionary biology. Proc Natl Acad Sci USA. 2005; 102:9541–9546. [PubMed: 15980155]
32. Shenkin PS, Erman B, Mastrandrea LD. Information-theoretical entropy as a measure of sequence variability. Proteins. 1991; 11:297–313. DOI: 10.1002/prot.340110408 [PubMed: 1758884]
33. Crow, JF., Kimura, M. An introduction to population genetics theory. Harper & Row; 1970.
34. Kimura M. Some problems of stochastic processes in genetics. Ann Math Stat. 1957; 28:882–901.
35. Kimura M. On the probability of fixation of mutant genes in a population. Genetics. 1962; 47:713–719. [PubMed: 14456043]
36. Goldstein RA. Population size dependence of fitness effect distribution and substitution rate probed by biophysical model of protein thermostability. Genome Biol Evol. 2013; 5:1584–1593. DOI: 10.1093/gbe/evt110 [PubMed: 23884461]
37. Cherry JL. Should we expect substitution rate to depend on population size? Genetics. 1998; 150:911–919. [PubMed: 9755219]
38. Eyring H. The Activated Complex in Chemical Reactions. J Chem Phys. 1935; 3:107–115.
39. Fisher, R. The Genetic Theory of Natural Selection. Oxford University Press; 1930.
40. Wylie CS, Shakhnovich EI. A biophysical protein folding model accounts for most mutational fitness effects in viruses. Proc Natl Acad Sci USA. 2011; 108:9916–9921. DOI: 10.1073/pnas.1017572108 [PubMed: 21610162]
41. Izaguirre JA, et al. CompuCell, a multi-model framework for simulation of morphogenesis. Bioinformatics. 2004; 20:1129–1137. DOI: 10.1093/bioinformatics/bth050 [PubMed: 14764549]
42. Miyazawa S, Jernigan R. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules. 1985; 18:534–552.

Europe PMC Funders Author Manuscripts

43. Lindqvist Y, Johansson E, Kaija H, Vihko P, Schneider G. Three-dimensional structure of a mammalian purple acid phosphatase at 2.2 A resolution with a mu-(hydr)oxo bridged di-iron center. Journal of Molecular Biology. 1999; 291:135–147. [PubMed: 10438611]
44. Gillespie DT. Exact Stochastic Simulation of Coupled Chemical Reactions. J Phys Chem. 1977; 81:2340–2361.
45. Kimura M. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J Mol Evol. 1980; 16:111–120. [PubMed: 7463489]
46. Forgy EW. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics. 1965; 21:768–769.
47. Lloyd S. Least squares quantization in PCM. IEEE Transactions on Information Theory. 1982; 28:129–137. DOI: 10.1109/TIT.1982.1056489
48. Khatri BS, Goldstein RA. A coarse-grained biophysical model of sequence evolution and the population size dependence of the speciation rate. J Theor Biol. 2015; 378:56–64. DOI: 10.1016/j.jtbi.2015.04.027 [PubMed: 25936759]
49. Khatri BS, McLeish TC, Sear RP. Statistical mechanics of convergent evolution in spatial patterning. Proc Natl Acad Sci USA. 2009; 106:9564–9569. DOI: 10.1073/pnas.0812260106 [PubMed: 19497876]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

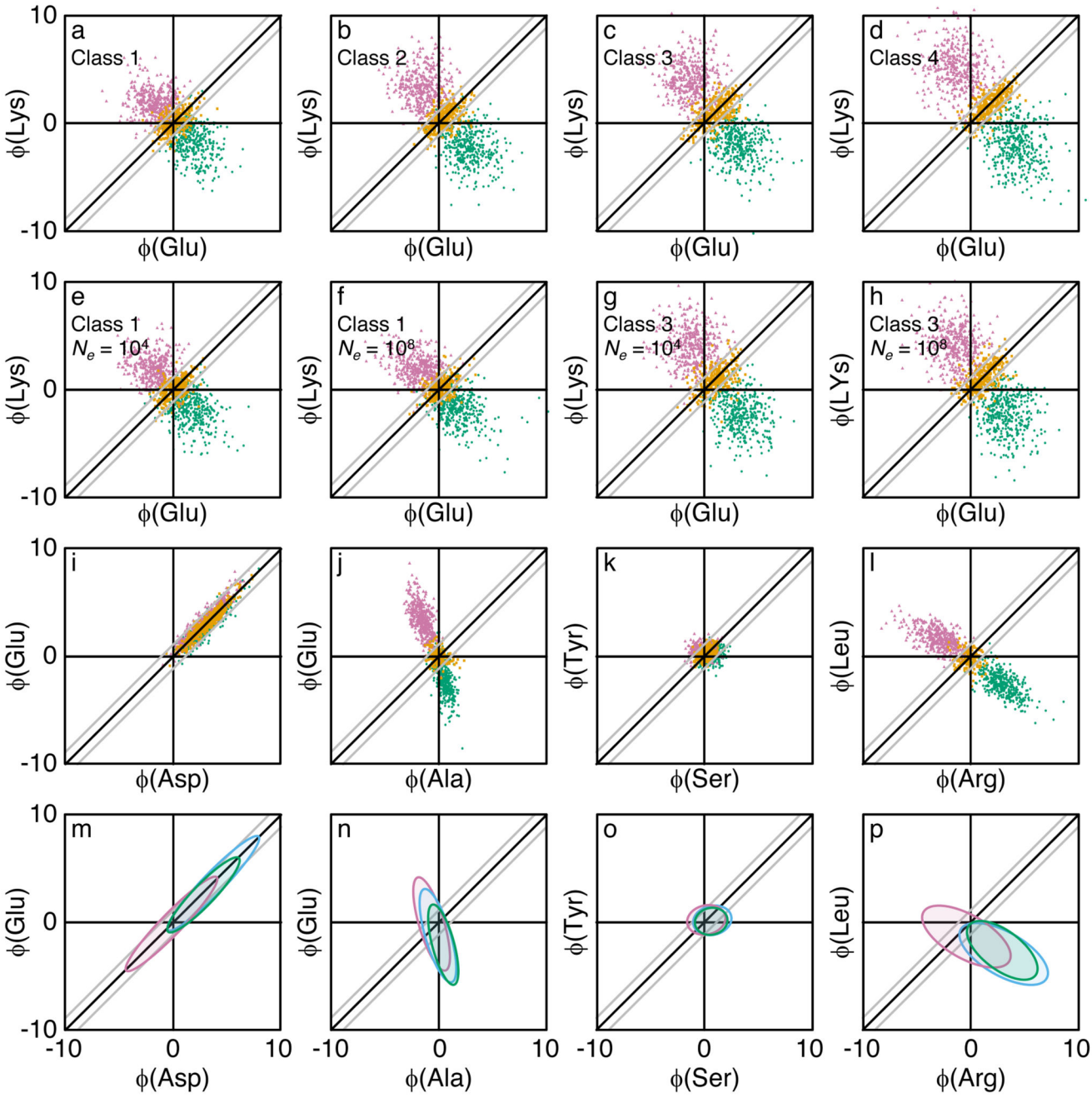

Figure 1.

Relative stabilities of amino acid pairs. The stability values of two amino acid residues (Res, labelled with standard three-letter abbreviations) are shown on the two axes (ϕ(Res)). In a-l, relative local contributions to stability are shown for glutamic acid and lysine in different site classes (a-d), or for different population sizes for site class 1 and 3 (e-h), or various amino acid pairs in site class 3 (i-l). Points were sampled when the amino acid on the x-axis was resident (green), when the amino acid on the y-axis was resident (pink), or during transitions between the two (yellow). In m-p, distributions of local contributions to stability

in reference state are shown when the non-interacting null amino acid was present ($\rho(\phi_\alpha, \phi_\beta|\varnothing)$, pink), when the amino acid on the x-axis was present as predicted using Equation (7) ($\tilde{\rho}(\phi_\alpha, \phi_\beta|\alpha)$, cyan), or as observed ($\rho(\phi_\alpha, \phi_\beta|\alpha)$, green). Grey diagonal lines mark the boundaries of regions of near-neutral substitutions. These and all other stability values are in kcal mol^{-1}.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

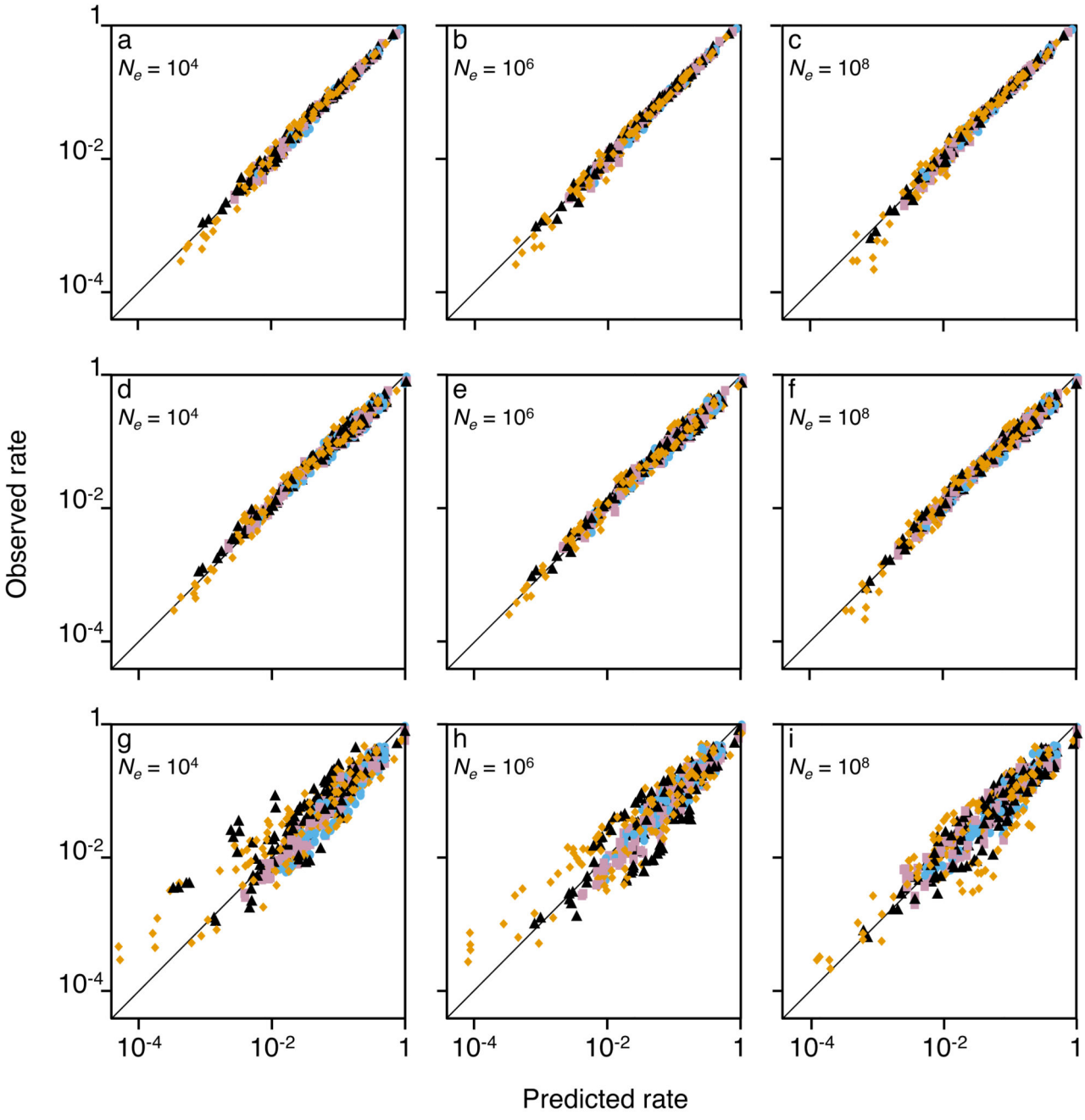

Figure 2.

Comparison of predicted and observed substitution rates. Predicted rates (x-axis) and observed rates (y-axis) are shown for all pairs of amino acids separated by a single base change for all sites in the different site classes (Class 1, blue circles; Class 2, pink squares; Class 3, black triangles; Class 4, orange diamonds). In a-c: predicted substitution rates calculated by integrating over $\rho(\phi_\alpha, \phi_\beta|\alpha)$ for three different population sizes; d-f: Predicted substitution rates calculated using transition state theory (Equation (6)), which assumes only

near-neutral substitutions occur; g-i: Predicted substitution rates calculated using transition state theory with parameters estimated using Equation (7).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

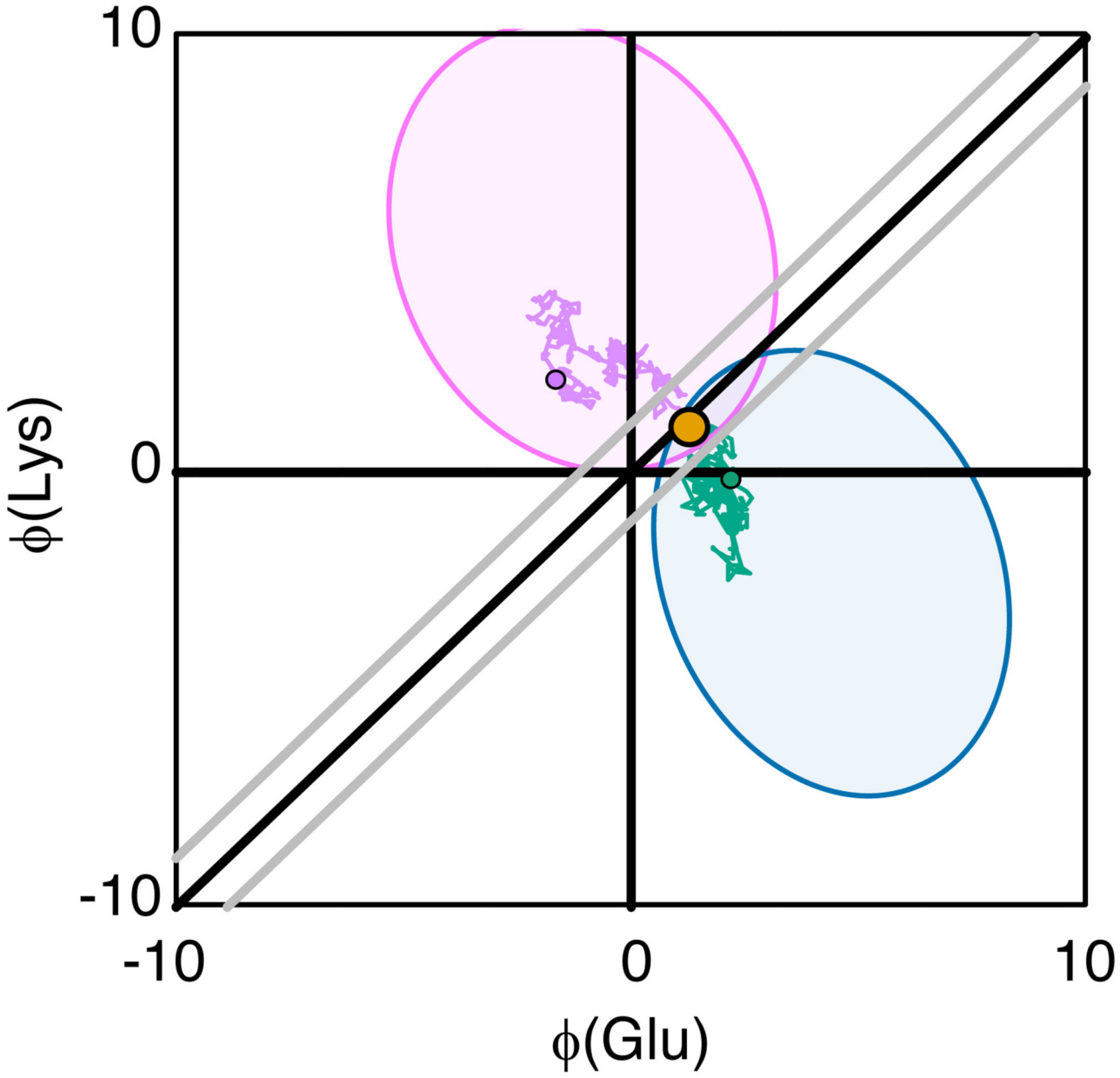

Figure 3.

Example of a trajectory before and after a substitution from glutamic acid to lysine. Stabilities on the x-axis and y-axis are shown as in Figure 1. Local contribution to stability when either is resident is shown for before (green) and after the substitution (pink) (green). Values during the substitution shown in yellow; beginning and end points are shown as black circles. The observed distributions over the simulations when glutamic acid or lysine is resident shown as shaded region. Grey diagonal lines mark regions of near-neutral substitutions.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

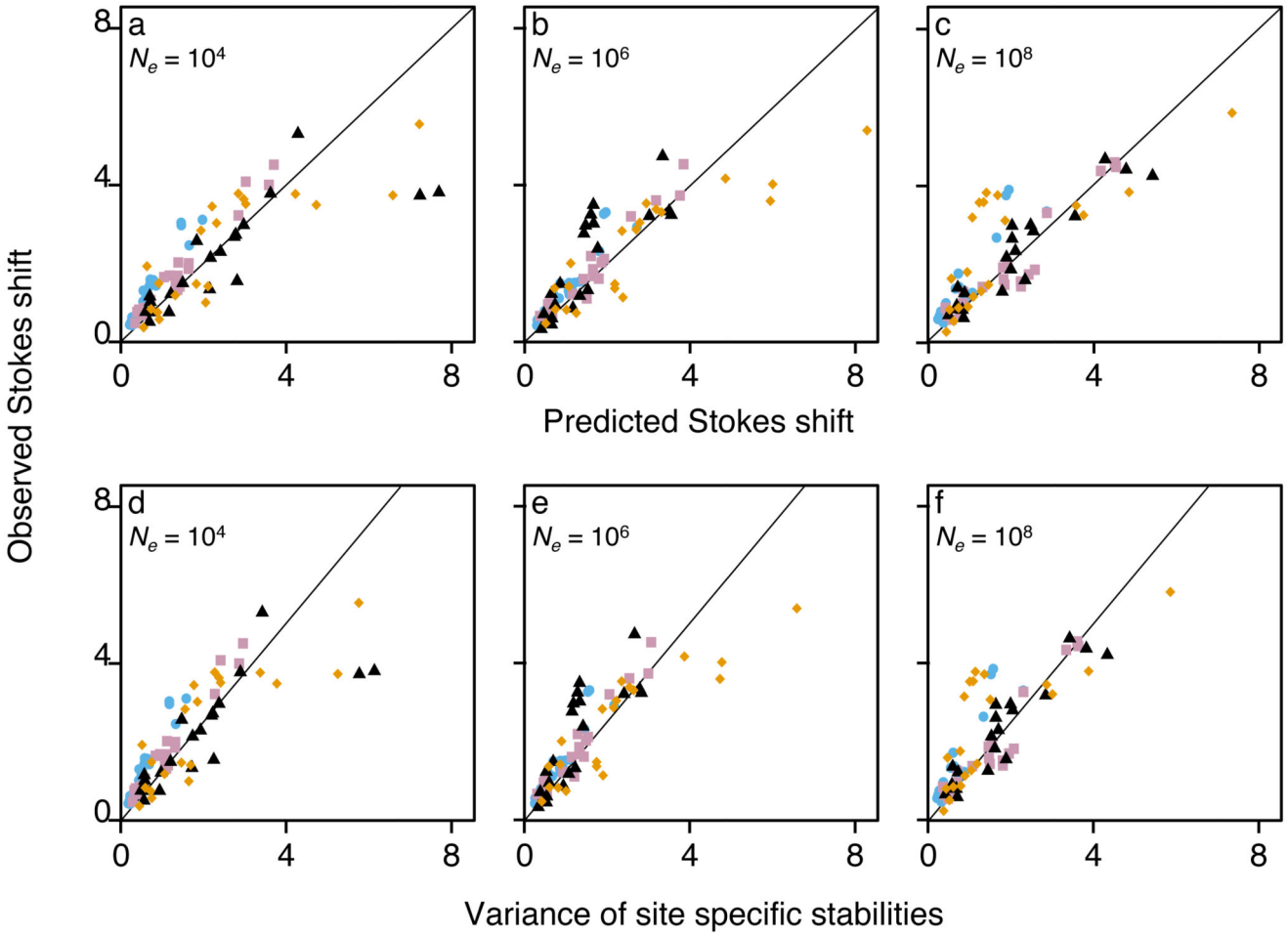

Figure 4.

Accuracy of site-specific stability and evolutionary Stokes shift predictions. Observed (y-axis) versus estimated values of the evolutionary Stokes shift ($\zeta_{\alpha|\alpha}$, x – axis) are shown in a-c for all four site rate classes (Class 1, blue circles; Class 2, pink squares; Class 3, black triangles; Class 4, orange diamonds), for three different population sizes. The linear relationship between the observed evolutionary Stokes shift (y-axis) and the variance in amino acid-specific stability contributions in the absence of selection on the site $(\sigma^2_{\alpha|\varnothing}, \mathrm{x-axis})$ are shown in d-f. The lines shown are theoretical predictions with $\gamma = 1.26$. Outliers are identified.