Beyond Stability Constraints: A Biophysical Model of Enzyme Evolution with Selection on Stability and Activity

Julian Echave*,[1]

[1]Escuela de Ciencia y Tecnología, Universidad Nacional de San Martín (UNSAM), Buenos Aires, Argentina

***Corresponding author:** E-mail: jechave@unsam.edu.ar.

Associate editor: Jeffrey Thorne

Abstract

The rate of evolution varies among sites within proteins. In enzymes, two rate gradients are observed: rate decreases with increasing local packing and it increases with increasing distance from catalytic residues. The rate-packing gradient would be mainly due to stability constraints and is well reproduced by biophysical models with selection for protein stability. However, stability constraints are unlikely to account for the rate-distance gradient. Here, to explore the mechanistic underpinnings of the rate gradients observed in enzymes, I propose a stability–activity model of enzyme evolution, M_{SA}. This model is based on a two-dimensional fitness function that depends on stability, quantified by ΔG, the enzyme's folding free energy, and activity, quantified by ΔG^*, the activation energy barrier of the enzymatic reaction. I test M_{SA} on a diverse data set of enzymes, comparing it with two simpler models: M_S, which depends only on ΔG, and M_A, which depends only on ΔG^*. I found that M_{SA} clearly outperforms both M_S and M_A and it accounts for both the rate-packing and rate-distance gradients. Thus, M_{SA} captures the distribution of stability and activity constraints within enzymes, explaining the resulting patterns of rate variation among sites.

Key words: **protein evolution model, stability constraints, activity constraints, site-specific rates.**

Introduction

It has long been known that the rate of evolution varies among sites within proteins. Ever since the early days of Molecular Evolution, this rate variation has been typically interpreted in terms of purifying selection to maintain function and structure: slowly evolving sites are those where mutations are more likely to be discarded by natural selection because they perturb the protein's structure or function too much; rate would be determined by so called structural and functional constraints (Perutz et al. 1965; Kimura and Ohta 1974).

Further insight into structural constraints came from work that studied the dependence of evolutionary rate on various properties of the local environment of protein sites. Rate increases with decreasing local packing density (Franzosa and Xia 2009; Yeh, Huang, et al. 2014; Yeh, Liu, et al. 2014; Marcos and Echave 2015; Shahmoradi and Wilke 2016; Sharir-Ivry and Xia 2017); rate increases with increasing solvent accessibility (Dean et al. 2002; Franzosa and Xia 2009; Ramsey et al. 2011; Scherrer et al. 2012; Franzosa and Xia 2012); and rate increases with increasing local flexibility (Liao et al. 2005; Liu and Bahar 2012; Nevin Gerek et al. 2013; Marsh and Teichmann 2014; Huang et al. 2014; Shahmoradi et al. 2014). These findings led to the view that the rate of evolution is mainly determined by protein structure, increasing from a slowly evolving, buried, tightly packed, and rigid protein core, toward a rapidly evolving, solvent-exposed, loosely packed, flexible surface (Echave et al. 2016).

In the previous view, functional constraints play only a minor role, affecting the conservation of just a few sites, such as enzyme catalytic residues and some of their immediate neighbors (Bartlett et al. 2002; Torrance et al. 2005; Echave et al. 2016). However, in enzymes, site-specific rates depend not only on local structural properties, but also on distance from active residues (Dean et al. 2002). Site-specific substitution rates increase rather slowly with increasing distance, so that enzyme active sites seem to influence evolutionary rates at long distances, affecting most protein sites (Jack et al. 2016).

Summarizing, within enzymes there are two clear evolutionary rate gradients. First, rate increases with decreasing local packing density, which I will call the rate-packing gradient. (The rate-packing gradient also represents the dependence of rate on solvent accessibility and local flexibility, because these properties strongly correlate with packing.) Second, rate increases with distance from the active site, which I will call the rate-distance gradient. To understand the biophysical origin of these gradients beyond the useful but vague notions of structural and functional constraints, necessitates *mechanistic* models grounded as much as possible on first principles of Molecular Evolution and Protein Biophysics.

Most biophysical models developed so far assume that protein fitness depends on stability (Echave and Wilke 2017). Such stability-based models account for much of the observed variation of evolutionary rates among sites (Huang et al. 2014; Echave et al. 2015; Marcos and Echave 2015; Goldstein and Pollock 2017; Jimenez et al. 2018). More importantly, stability-based models account for the rate-packing

gradient, which would, therefore, reflect stability constraints (Huang et al. 2014; Marcos and Echave 2015). However, stability constraints are unlikely to account for the rate-distance gradient, which would more likely reflect activity constraints (Jack et al. 2016).

To the best of my knowledge, none of the biophysical models developed so far can be used to explain the rate-distance gradient found in enzymes. There are a few biophysical models with selection on activity, but they are based on ligand binding rather than catalysis (Echave and Wilke 2017). While these models do predict rates depending on distance from the binding site (Nelson and Grishin 2016), they are not suitable for enzymes because the key to enzymatic activity is not improving ligand binding but lowering the activation energy barrier of the catalyzed reaction. Therefore, advancing our understanding of enzyme evolution demands the development of new biophysical models based on realistic principles of enzymatic catalysis.

Here, I propose a mechanistic biophysical *stability–activity model* of enzyme evolution, M_{SA}, that includes explicitly selection on stability and activity. I will show that M_{SA} captures the distribution of stability and activity constraints within enzymes, explaining both the rate-packing and the rate-distance gradients.

New Approach

In this section, I briefly describe the most important assumptions and formulas of the M_{SA} model of enzyme evolution. A sketch of the key model elements is shown in figure 1. For a detailed derivation of M_{SA} see Supplementary Material online.

Fitness, Fixation Probability, and Substitution Rate

I model fitness as a two-dimensional step function (fig. 1A):

$$F(\Delta G, \Delta G^*) = \begin{cases} 1 & \text{if } \Delta G < \Delta G_{thr} \text{ and } \Delta G^* < \Delta G^*_{thr}, \\ 0 & \text{if } \Delta G >= \Delta G_{thr} \text{ or } \Delta G^* >= \Delta G^*_{thr}, \end{cases} \quad (1)$$

where ΔG is the folding free energy, quantifying stability, and ΔG^* is the activation free energy, quantifying activity. ΔG_{thr} and ΔG^*_{thr} are stability and activity thresholds.

In the limit of rare mutations, evolution can be modeled as an origination–fixation process (McCandlish and Stoltzfus 2014). Consider a monoclonal population of N individuals with common wild type genotype with stability ΔG and activation energy ΔG^* somewhere within the viable $F = 1$ region of the fitness landscape. A mutant arises with stability $\Delta G_{mut} = \Delta G + \Delta\Delta G$ and activation energy $\Delta G^*_{mut} = \Delta G^* + \Delta\Delta G^*$. The mutation will be either neutral, if the mutant remains in the viable $F = 1$ region of the fitness landscape, or lethal, if the mutant falls into the $F = 0$ region. A neutral mutation may become fixed with fixation probability $p^{fix} = 1/N$; a lethal mutation becomes lost ($p^{fix} = 0$).

For computational tractability, I further simplify the rate prediction problem using a mean field approximation (Bloom and Glassman 2009; Echave et al. 2015). Briefly, this approximation consists of 1) assuming that $\Delta\Delta G$ and $\Delta\Delta G^*$ are independent of the sequence background and 2) assuming that at evolutionary equilibrium stability and activity are distributed according to $\rho(\Delta G, \Delta G^*) = a_S e^{a_S(\Delta G - \Delta G_{thr})} \times a_A e^{a_A(\Delta G^* - \Delta G^*_{thr})}$ (fig. 1B). Then, averaging p^{fix} over ΔG and ΔG^* leads to (fig. 1C):

$$p^{fix}(\Delta\Delta G, \Delta\Delta G^*) = \frac{1}{N} \min\left(1, e^{-a_S \Delta\Delta G}\right) \times \min\left(1, e^{-a_A \Delta\Delta G^*}\right). \quad (2)$$

This mean field fixation probability depends on $\Delta\Delta G$ and $\Delta\Delta G^*$, and on two positive parameters a_S and a_A. Since increasing either of these parameters decreases p^{fix}, they can be interpreted as quantifying the degree of selection pressure.

Knowing the fixation probability, we can calculate the rate of evolution. The rate of $i \to j$ substitutions (i.e., fixed mutations) at a protein site r is given by

$$Q^r_{ji} = N M_{ji} p^{fix}(\Delta\Delta G_{ji,r}, \Delta\Delta G^*_{ji,r}), \quad (3)$$

where N is population size, M_{ji} is the rate of $i \to j$ mutations, and p^{fix}, given by equation (2), depends on $\Delta\Delta G_{ji,r}$ and $\Delta\Delta G^*_{ji,r}$, the stability and activation energy changes due to the $i \to j$ mutation at site r. The average substitution rate of site r is given by

$$K^r = \sum_i \sum_{j \neq i} Q^r_{ji} \pi^r_i, \quad (4)$$

where Q^r_{ji} is given by equation (3) and π^r_i is the equilibrium probability of finding amino acid i at site r.

Mutational Change of Stability and Activation Energy

To calculate $\Delta\Delta G$ and $\Delta\Delta G^*$, I used the Linearly Forced Elastic Network Model (LFENM; Echave 2008; Echave and Fernández 2010; Huang et al. 2014; Marcos and Echave 2015). The LFENM model represents a given protein as an elastic network of nodes (amino acids) connected by harmonic springs (interactions) and models mutations as random perturbations of the lengths of the springs that connect the mutated site to other sites. The elastic network has a quadratic energy function with a single minimum at the equilibrium conformation. (A protein conformation is represented by $\mathbf{r}$, the vector of Cartesian coordinates of all network nodes.) Given a wild type protein with energy $V_{wt}(\mathbf{r})$ with minimum at $\mathbf{r}^0_{wt}$ and a mutant with energy $V_{mut}(\mathbf{r})$ with minimum at $\mathbf{r}^0_{mut}$, it is possible to derive:

$$\Delta\Delta G = V_{mut}(\mathbf{r}^0_{mut}) - V_{wt}(\mathbf{r}^0_{wt}) = \frac{1}{2}\sum_{ij} k_{ij} \delta l^2_{ij} - \frac{1}{2}(\mathbf{r}^0_{mut} - \mathbf{r}^0_{wt})^T \mathbf{K} (\mathbf{r}^0_{mut} - \mathbf{r}^0_{wt}), \quad (5)$$

where k_{ij} is the force constant of the spring connecting sites i and j, δl_{ij} is the change of the length of spring ij due to the mutation, and $\mathbf{K}$ is the Hessian matrix common to $V_{mut}(\mathbf{r})$ and $V_{wt}(\mathbf{r})$. $\Delta\Delta G$ given by equation (5) is the difference between the mutant's and wild-type's minimum energies (see fig. 1D).

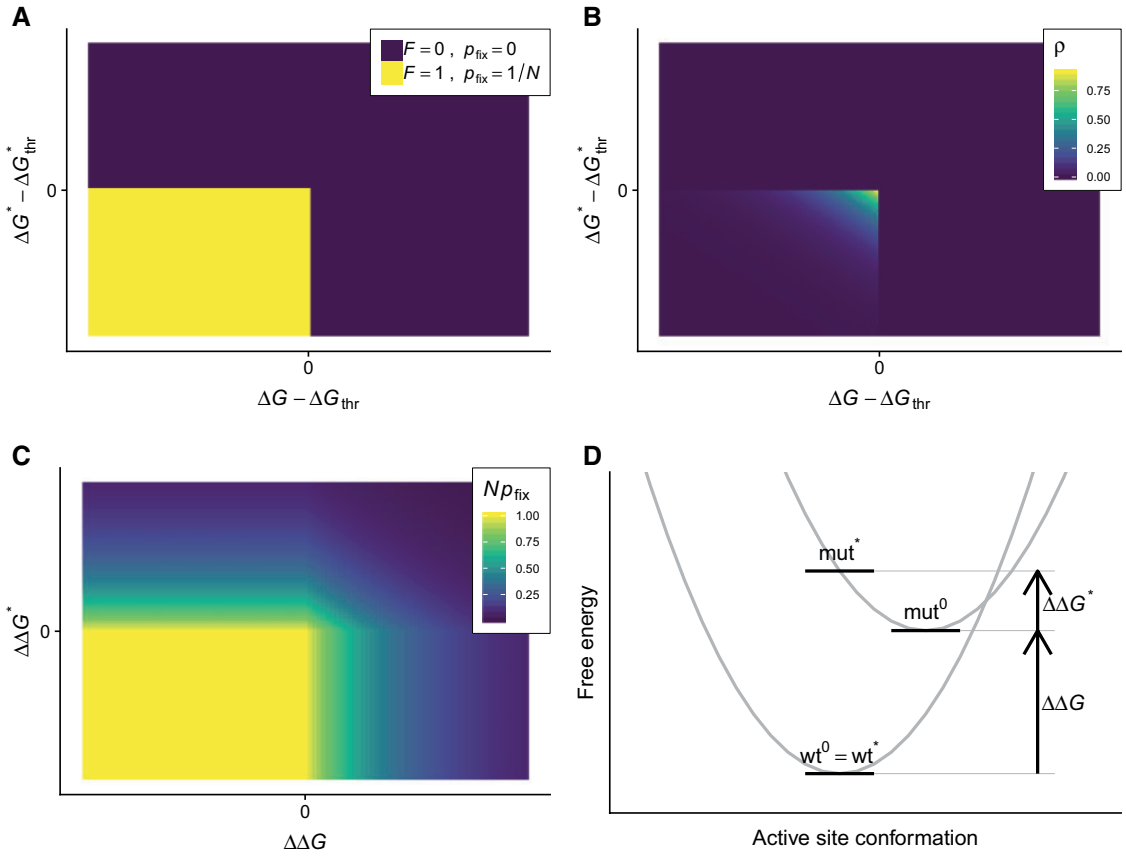

Fig. 1. The stability–activity model M_{SA}. (*A*) Fitness is a 2D step function: enzymes are viable ($F=1$) and may become fixed ($p^{fix}=1/N$, N being population size) only if their folding free energy ΔG is more negative than a threshold energy ΔG_{thr} and if their activation energy barrier ΔG^* is lower than a threshold ΔG^*_{thr}. (*B*) Distribution of ΔG and ΔG^* at evolutionary equilibrium. (*C*) A mean field p^{fix} that depends only on $\Delta\Delta G$ (mutational change of stability) and $\Delta\Delta G^*$ (mutational change of activation energy) is obtained by integrating the p^{fix} of *A* over the distribution of *B*. (*D*) $\Delta\Delta G$ is the difference between energy minima; $\Delta\Delta G^*$ is the energy needed to distort the mutant from its equilibrium conformation mut^0 to the active conformation mut^* (the wild type's active site is assumed preorganized in the active conformation: $wt^0 = wt^*$).

To calculate $\Delta\Delta G^*$ I made the following considerations. First, at low substrate concentrations, the activation energy barrier is the free energy difference between the transition state ES^* and enzyme E and substrate S free in solution: $\Delta G^* = G(ES^*) - G(E) - G(S)$ (Schowen 1978). Second, ΔG^* can be written as the sum of two contributions, a *distortion energy*, which is the energy necessary for the enzyme and substrate to adopt their transition state conformations, E^* and S^*, plus a *vertical binding energy*, which is the energy released when E^* and S^* bind to form ES^* (Schowen 1978; Stein 2011). Third, I assume that mutations have no effect on the vertical binding energy or the substrate distortion, but affect only the enzyme's distortion free energy: $\Delta\Delta G^* = \Delta G(E \rightarrow E^*)$. Finally, I assume that the wild-type has a preorganized active site with a conformation which is identical to the transition state conformation. Using these assumptions and the LFENM model, I derived that $\Delta\Delta G^*$ is the energy needed to distort the mutant's active site from the mutant's conformation $\mathbf{r}^0_{a,mut}$ to the wild type conformation $\mathbf{r}^0_{a,wt}$ (fig. 1*D*):

$$\Delta\Delta G^* = \frac{1}{2}(\mathbf{r}^0_{a,wt} - \mathbf{r}^0_{a,mut})^T \mathbf{K}^{eff}_{aa}(\mathbf{r}^0_{a,wt} - \mathbf{r}^0_{a,mut}), \quad (6)$$

where $\mathbf{K}^{eff}_{aa}$ is a matrix that allows the calculation of the effective energy of distortions within the conformational subspace spanned by active residue coordinates $\mathbf{r}_a$.

By construction, LFENM assumes that the wild type protein used to build the elastic network is the most stable protein. Thus, the energy change from wild type state $k=0$ to any other state $k \neq 0$ obtained using equation (5) will have a positive $\Delta\Delta G_{k0}$. Similarly, due to the assumption that the wild type has an ideally preorganized active site, the $\Delta\Delta G^*_{k0}$ calculated using equation (6) are all positive. However, $\Delta\Delta G_{ji} = \Delta\Delta G_{j0} - \Delta\Delta G_{i0}$ and $\Delta\Delta G^*_{ji} = \Delta\Delta G^*_{j0} - \Delta\Delta G^*_{i0}$ may be positive or negative.

Note, finally, that LFENM energy functions depend only on network topology and not on the specific amino acids represented by the network nodes. Therefore, LFENM "mutations" do not correspond to actual amino acid mutations. In other words, possible LFENM node states are not actual amino acids. However, rates averaged over LFENM states do vary among sites in the same way as rates averaged

over amino acid substitutions, as I will show and discuss below.

Results

I tested the M_{SA} model on a data set of 163 monomeric enzymes with diverse sizes, functions, and structures. In addition to M_{SA}, I considered a stability model M_S, dependent only on $\Delta\Delta G$, and an activity model M_A, dependent on $\Delta\Delta G^*$. I obtained site-specific rates predicted by the models, $K_{M_{SA}}$, K_{M_S}, and K_{M_A}, by fitting the models' parameters to observed rates K_{obs}, which are rate estimates obtained from protein sequence alignments (see Materials and Methods).

M_{SA} Outperforms M_S and M_A

First, I considered which model provides the best rate predictions. Results are shown in figure 2. According to the Akaike Information Criterion *AIC* (see Materials and Methods), M_{SA} outperforms M_S for 151 out of the 163 enzymes studied (fig. 2*A*) and it outperforms M_A for 162/163 enzymes (fig. 2*B*). Thus, rate variation among protein sites is influenced by constraints on stability *and* activity.

M_{SA} Accounts for the Rate-packing and Rate-distance Gradients

Beyond the overall goodness of fit, I am particularly interested in whether the models account for the rate gradients observed in enzymes. To this end, I studied the dependence of observed and model rates on *WCN*, the Weighted Contact Number, a metric of local packing, and on d_{active}, the distance from the closest active residue. For clarity, in figure 3, I explicitly show one example (phosphomannose isomerase of Candida albicans, which has pdb code 1PMI). For this case, M_S fits the $K_{obs} \sim WCN$ gradient reasonably well, but it fails to account for the $K_{obs} \sim d_{active}$ gradient (fig. 3*A*). Conversely, M_A fails to fit $K_{obs} \sim WCN$, but fits well the $K_{obs} \sim d_{active}$ gradient (fig. 3*B*). M_{SA} reproduces almost perfectly the $K_{obs} \sim WCN$ dependence and it fits very well the $K_{obs} \sim d_{active}$ dependence. Thus, the one-dimensional models M_S and M_A reproduce well either one gradient or the other, but not both. In contrast, the two-dimensional model M_{SA} accounts for both the rate-packing and the rate-distance gradients.

To generalize the previous analysis, I repeated it for each of the enzymes of the data set. Results are shown in figure 4. From the M_{SA} versus M_S comparison (fig. 4*A*), M_{SA} outperforms M_S in reproducing both gradients: the rate-packing gradient for 123/163 cases and the rate-distance gradient for *all* 163/163 cases. From the M_{SA} versus M_A comparison (fig. 4*B*), M_{SA} also outperforms M_A in reproducing both gradients: the rate-packing gradient for 160/163 cases and the rate-distance gradient for 139/163 cases. Thus, M_{SA} accounts for both the rate-packing gradient, for which M_A fails, and the rate-distance gradient, for which M_S fails.

Discussion

I proposed the M_{SA} model of enzyme evolution with selection on stability and activity. I derived all the formulas needed to calculate site-specific substitution rates from mutational changes of stability, $\Delta\Delta G$, and activation energy, $\Delta\Delta G^*$. Further, I derived the equations needed to calculate $\Delta\Delta G$ and $\Delta\Delta G^*$ using a LFENM. For comparison, I also considered a stability-based model M_S and an activity-based model M_A.

I tested M_{SA} in comparison with M_S and M_A on a diverse data set of monomeric enzymes. I found that M_{SA} fits observed rates better than both M_S and M_A for most proteins studied. More importantly, M_{SA} is the only model that accounts for both gradients observed in enzymes: the rate-packing $K \sim WCN$ gradient and the rate-distance $K \sim d_{active}$ gradient. In contrast, M_S fails to account for the rate-distance gradient and M_A fails to account for the rate-packing gradient. Taken together, these findings suggest that these rate gradients reflect the distribution of stability and activity constraints within enzymes.

To further support the model, I run some extra tests (see Supplementary Material online). Briefly, I found that the use of LFENM to calculate $\Delta\Delta G$ is validated by the fact that using an all-atom energy function to calculate $\Delta\Delta G$ gives similar

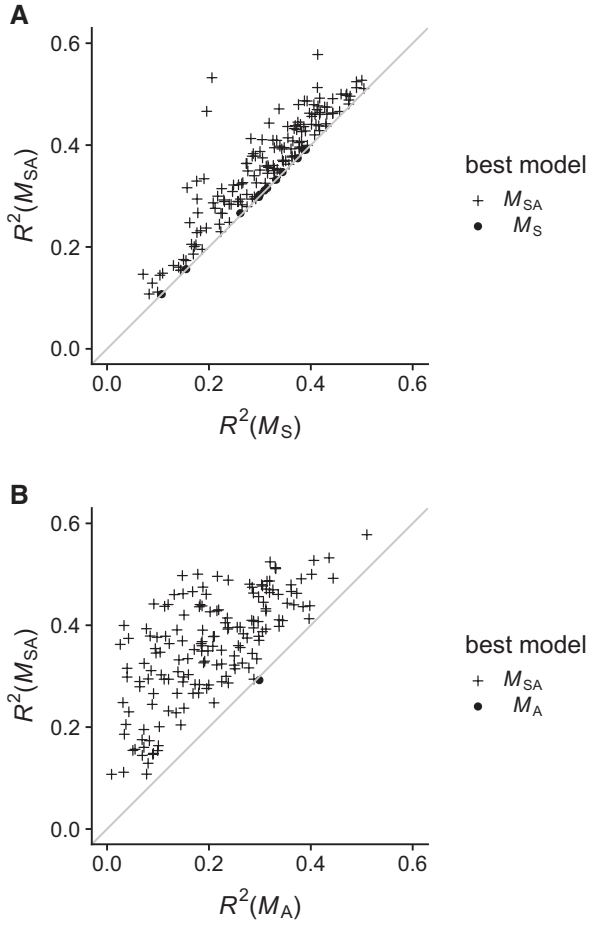

Fig. 2. M_{SA} fits observed rates better than M_S and M_A do for almost all enzymes. $R^2(M) = R^2(K_M, K_{obs})$ is the square correlation coefficient between model rates and observed rates. Each point corresponds to one protein. The $y = x$ line is shown for reference. Point types indicate which of the two compared models is the best model according to the Akaike Information Criterion *AIC*. (*A*) M_{SA} versus M_S. (*B*) M_{SA} versus M_A.

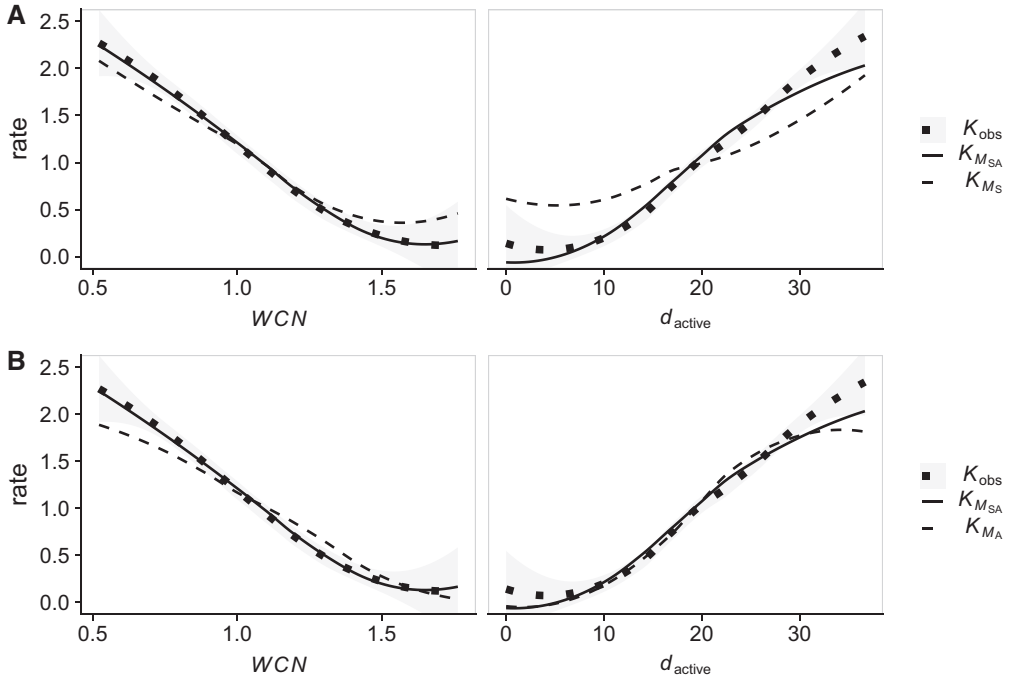

Fig. 3. M_{SA} accounts for the rate-packing and rate-distance gradients for 1PMI. Observed and predicted rate-packing gradient ($K \sim WCN$, left panels) and rate-distance gradient ($K \sim d_{active}$, right panels) for example case 1PMI. The rate-property lines shown were obtained fitting rate-property scatter plots using local polynomial regression (see Materials and Methods). The shaded area represents the error of the smooth fit for the case of K_{obs}. (*A*) M_{SA} versus M_S. (*B*) M_{SA} versus M_A. M_S fails to reproduce the $K_{obs} \sim d_{active}$ dependence; M_A fails to reproduce the $K_{obs} \sim WCN$ dependence; M_{SA} accounts for both gradients.

rate predictions. Second, the preorganized-active-site assumption used to calculate $\Delta\Delta G^*$ is supported by the finding that model rates are insensitive to whether the enzyme form used to build the LFENM is free or complexed. Finally, I found that a major part of the $K_{M_{SA}}-K_{obs}$ lack of fit is due to poor accuracy of K_{obs} estimates, so that better fit will more likely come from better data rather than model improvements.

Having shown that M_{SA} is a good model to explain rate variation among sites, I discuss some possible caveats related to the model's assumptions. One caveat is the assumption that $\Delta\Delta G$ and $\Delta\Delta G^*$ are independent of the precise sequence background in which mutations occur. This assumption may have to be removed to study some epistatic phenomena such as, for instance, the co-evolution of sites in contact. However, the background-independence assumption is reasonable for the calculation of single-site properties such as estimating site-specific free energy changes from protein alignments (Bloom and Glassman 2009) and predicting site-specific rates(Echave et al. 2015). The present results further support to this assumption.

A second caveat is the assumption that mutations affect only the distortion energy contribution to the activation energy barrier. This assumption will be invalid for *some* sites: mutating catalytic or ligand-binding residues will also affect the binding energy contribution to $\Delta\Delta G^*$. However, that $\Delta\Delta G^*$ is due to distortion is valid for *most* sites, which are not directly involved in binding. This is probably why M_{SA} successfully reproduces the overall pattern of rate variation among sites.

A third issue is that LFENM models mutations as perturbations that cannot be mapped to actual mutations. Accordingly, LFENM matrix elements Q^r_{ji} (eq. 3) represent rates of interchange between node states rather than rates of substitution between specific amino acids. Therefore, LFENM cannot be used to predict substitution rates between actual amino acids. This limitation, however, does not prevent the calculation of average site-specific rates K^r. Importantly, the good fit between these rates and observed rates means that they depend more on the on the topology of the network of amino acids than on their specific identity.

A final caveat is that since LFENM node states are not actual amino acids, it is not clear how to choose mutational rates M_{ji}. For this reason, here I assumed a single mutational rate $M_{ji} = \mu$. A priori, this assumption seems unrealistic because it would not account for the possible effect on site-specific rates of mutational biases (e.g., due to the structure of the genetic code). Yet, M_{SA} rates do fit observed rates and reproduce the gradients found in enzymes. This suggests that relative rates are insensitive to the mutational pattern, which is consistent with recent work that found that relative rate estimates at codon level or amino-acid level are very similar (Sydykova and Wilke 2017) and that rate estimates are insensitive to the substitution model (Spielman and Pond 2018).

To finish, I point out some possible future applications of this work. I believe the M_{SA} model will be helpful to explore several fundamental issues and to develop applications. For example, M_{SA} might be helpful to explain why proteins evolve to be moderately efficient (Bar-Even et al. 2011, 2015), just as stability-based models have explained why proteins are

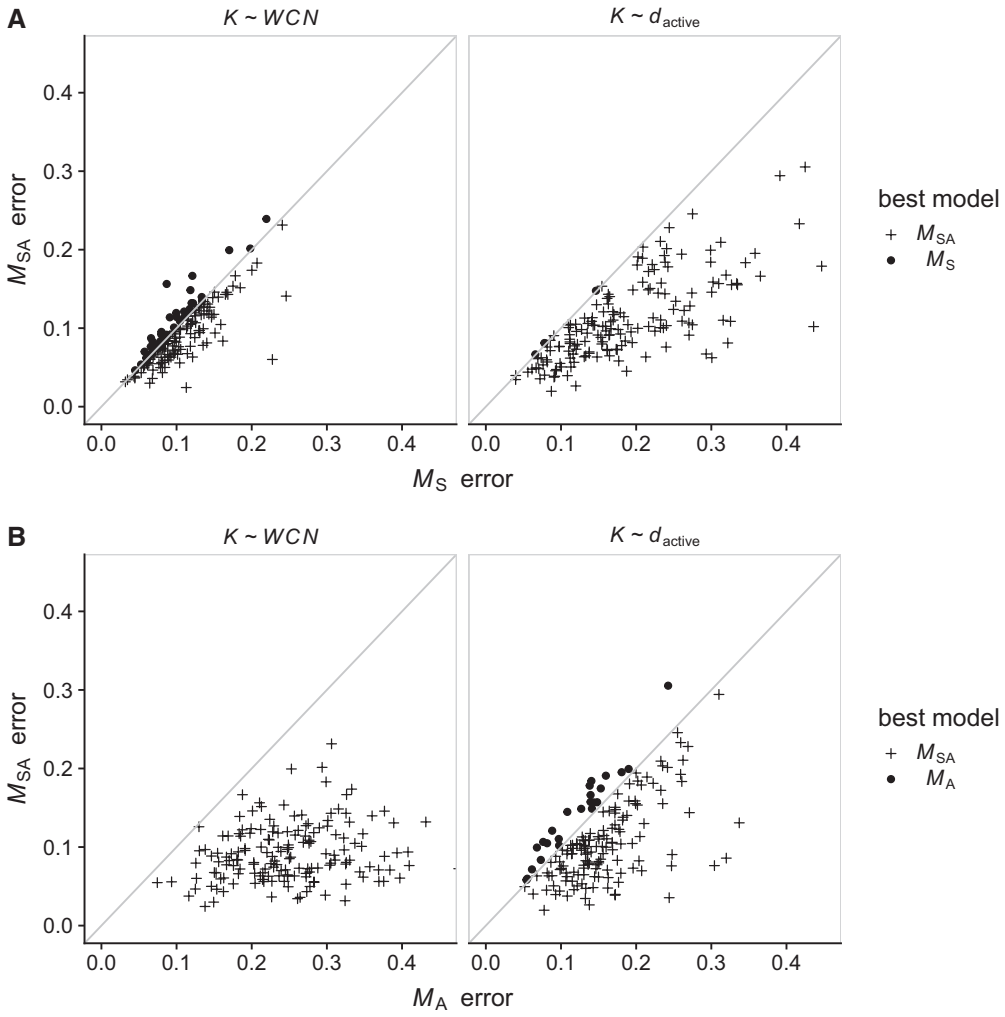

Fig. 4. M_{SA} accounts for the rate-packing and the rate-distance gradients. Dissimilarity (error) between model and observed $K \sim WCN$ gradients (left panels) and $K \sim d_{active}$ gradients (right panels). (For the error metric see Materials and Methods.) Each point represents one protein. The $y = x$ line is shown as reference. Point types represent which of the models compared fits the given gradient best (has lower error). (*A*) M_{SA} versus M_S; (*B*) M_{SA} versus M_A. M_S fails to predict the rate-distance gradient (*A*, right panel) and M_A fails to predict the rate-packing gradient (*B*, left panel). In general, M_{SA} outperforms M_S and M_A in all panels, reproducing well both the rate-packing and rate-distance gradients.

marginally stable (Taverna and Goldstein 2002; GoldStein 2011). Another classical issue that may be explored using M_{SA} is the evolutionary implications of a trade-off between stability and activity (Miller 2017). On the applied side, M_{SA} could be used to improve active site prediction (Jack et al. 2016). Another application would be to add activity constraints, as modeled by M_{SA}, to improve probabilistic evolution models used for phylogenetic inference purposes (Rodrigue and Lartillot 2017). In general, I expect that developing biophysical models of protein evolution that consider selection on stability and activity, such as M_{SA}, is a promising research endeavor that will advance our understanding of protein evolution and impact many areas of evolutionary biology.

Materials and Methods

Data Set and Empirical Rates

I tested the models on a data set of 163 enzymes, a subset of the 524 enzymes used in Jack et al. (2016). Specifically, I kept only monomeric enzymes and removed those that had missing amino acids or broken chains, which could result in wrong elastic network models. Catalytic residue information was obtained from the Catalytic Site Atlas (Furnham et al. 2014). The structures of these proteins were obtained from the RCSB protein database (Berman et al. 2000). The data set is diverse: no two enzymes have >25% sequence identity and there are representatives of the main SCOP structural classes (Murzin et al. 1995) and of the six main EC functional classes (Webb 1992). For some details of the data set, see Supplementary Material online.

Site-specific "observed" rates K_{obs} for these enzymes have been calculated before (Shih et al. 2012; Echave et al. 2015; Jack et al. 2016). Briefly, estimating K_{obs} involves finding homologous sequences, aligning them, inferring a phylogeny, and using the sequence alignment and phylogeny as input for the program `rate4site` to estimate the substitution rate of each site (Mayrose et al. 2004). Here, I have mostly used the rates of Echave et al. (2015), but to assess possible

estimation errors I compared them also with the rates of Jack et al. (2016).

Model Parameters

Mutation Matrix

I assume that all mutations have the same probability, thus $M_{ji} = \mu$ in equation (3).

Elastic Network Model

There is a large variety of ENMs (Fugebakk et al. 2013; Bastolla 2014; López-Blanco and Chacón 2016). Here, I used the ENM of Ming and Wall (2005): amino acids are represented by single nodes; nodes are connected if they are within $R_0 = 10.5$Å; l_{ij} is the distance between the nodes in the pdb structure; the force constants are $k_{ij} = 189\ kcal/mol$ for sequence neighbors and $k_{ij} = 4.5\ kcal/mol$ otherwise. I placed nodes at the side-chain geometric centers (except for glycine, which has no side chain, for which I used C_α coordinates).

LFENM Perturbations

A mutation is emulated by perturbing independently each of the springs that connect to the mutated site by adding perturbations δl_{ij}. Here, I used $\delta l_{ij} \sim N(0, \sigma = 0.3\text{Å})$.

Protein-dependent Parameters

To calculate model rates for a given protein, I started by performing a full mutational scan. For each enzyme, I used its pdb structure to build the ENM. Then, at each site I introduced $N = 19$ LFENM mutations and, for each mutation, I calculated $\Delta\Delta G$ (eq. 5) and $\Delta\Delta G^*$ (eq. 6). Then, for each model $M = M_{SA}, M_S, M_A$, I calculated model rates $K_M(\mathbf{a})$ using equations (2)–(4). I further normalized K_M so that $\langle K_M \rangle = 1$, so that rates are relative to the protein average, which is the normalization of K_{obs}. Finally, I found the model parameters $\mathbf{a}$ by minimizing the Residual Sum of Squares $RSS = \sum_r (K^r_{obs} - K^r_M(\mathbf{a}))^2$ using the general purpose optimization function optim of R package stats. Since models are fit independently, $a_S(M_S) \neq a_S(M_{SA})$ and $a_A(M_A) \neq a_A(M_{SA})$.

Assessing Whether Models Fit Observed Rates

To quantify model-data fit I used R^2 and AIC. $R^2(K_M, K_{obs})$, the square Pearson correlation coefficient, is the most frequently used goodness-of-fit measure. I have used it quantify the degree of improvement provided by M_{SA} over M_S or M_A. However, R^2 is not adequate for selecting among alternative nonlinear models (Spiess and Neumeyer 2010). To this end, I used the Akaike Information Criterion $AIC = 2k - 2\ln L$, where k is the number of parameters and L is the maximum likelihood. AIC is smaller for larger likelihoods and less parameters; the best model is that with the smallest AIC. Assuming normally distributed residuals $\ln L = -n/2[\ln(2\pi RSS/n) + 1]$, n being the number of sites and RSS the minimum residual sum of squares.

Assessing Whether Models Account for the Rate-packing and Rate-distance Gradients

To describe the dependence of site-specific rate K on a metric X, I obtained smooth fits to the K versus X scatter plots, using local polynomial regression using the function loess of R package stats. Specifically, for each protein, I obtained the loess functions $\hat{K}_{observed}(X)$ and $\hat{K}_M(X)$ for $M = M_{SA}, M_S, M_A$ and $X = WCN, d_{active}$. Then, I calculated Root Mean Square Errors: $RMSE(M, X) = 1/n \sum_i^n \left[\hat{K}_{observed}(X_i) - \hat{K}_M(X_i)\right]^2$, where X_i are n evenly spaced points that cover the range of X. The lower $RMSE(M, X)$, the better M accounts for the $K_{obs} \sim X$ dependence.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

Most of this work was done at the Max Planck Institute for the Physics of Complex Systems (Dresden, Germany), where I was on sabbatical. This work was supported by Consejo Nacional de Investigaciones Científicas y Técnicas (grant number PIP 11220150100385CO) and by Agencia Nacional de Promoción Científica y Tecnológica (grant number PICT-2016-4209).

References

Bar-Even A, Milo R, Noor E, Tawfik DS. 2015. The moderately efficient enzyme: futile encounters and enzyme floppiness. *Biochemistry* 54(32):4969–4977.

Bar-Even A, Noor E, Savir Y, Liebermeister W, Davidi D, Tawfik DS, Milo R. 2011. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry* 50(21):4402–4410.

Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. 2002. Analysis of catalytic residues in enzyme active sites. *J Mol Biol.* 324(1):105–121.

Bastolla U. 2014. Computing protein dynamics from protein structure with elastic network models. *WIREs Comput Mol Sci.* 4(5):488–503.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. *Nucleic Acids Res.* 28(1):235–242.

Bloom JD, Glassman MJ. 2009. Inferring stabilizing mutations from protein phylogenies: application to influenza hemagglutinin. *PLoS Comput Biol.* 5(4):e1000349.

Dean AM, Neuhauser C, Grenier E, Golding GB. 2002. The pattern of amino acid replacements in alpha/beta-barrels. *Mol Biol Evol.* 19(11):1846–1864.

Echave J. 2008. Evolutionary divergence of protein structure: the linearly forced elastic network model. *Chem Phys Lett.* 457(4–6):413–416.

Echave J, Fernández FM. 2010. A perturbative view of protein structural variation. *Proteins Struct Funct Bioinformatics.* 78(1):173–180.

Echave J, Jackson ELE, Wilke COC. 2015. Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. *Phys Biol.* 12(2):025002.

Echave J, Spielman SJ, Wilke CO. 2016. Causes of evolutionary rate variation among protein sites. *Nat Rev Genet.* 17(2):109–121.

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

Franzosa EA, Xia Y. 2009. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol Biol Evol.* 26(10):2387–2395.

Franzosa EA, Xia Y. 2012. Independent effects of protein core size and expression on residue-level structure–evolution relationships. *PLoS One* 7(10):e46602.
Fugebakk E, Reuter N, Hinsen K. 2013. Evaluation of protein elastic network models based on an analysis of collective motions. *J Chem Theory Comput.* 9:5618–5628.
Furnham N, Holliday GL, de Beer TAP, Jacobsen JOB, Pearson WR, Thornton JM. 2014. The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.* 42(D1):D485–D489.
Goldstein R. 2011. The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins Struct Funct Bioinformatics.* 79(5):1396–1407.
Goldstein RA, Pollock DD. 2017. Sequence entropy of folding and the absolute rate of amino acid substitutions. *Nat Ecol Evol.* 1(12):1923–1930.
Huang TT, Del Valle Marcos ML, Hwang JK, Echave J. 2014. A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. *BMC Evol Biol.* 14:78.
Jack BR, Meyer AG, Echave J, Wilke CO. 2016. Functional sites induce long-range evolutionary constraints in enzymes. *PLoS Biol.* 14(5):e1002452.
Jimenez MJ, Arenas M, Bastolla U. 2018. Substitution rates predicted by stability-constrained models of protein evolution are not consistent with empirical data. *Mol Biol Evol.* 35(3):743–755.
Kimura M, Ohta T. 1974. On some principles governing molecular evolution. *Proc Natl Acad Sci U S A.* 71(7):2848–2852.
Liao H, Yeh W, Chiang D, Jernigan RL, Lustig B. 2005. Protein sequence entropy is closely related to packing density and hydrophobicity. *Protein Eng Des Sel.* 18(2):59–64.
Liu Y, Bahar I. 2012. Sequence evolution correlates with structural dynamics. *Mol Biol Evol.* 29(9):2253–2263.
López-Blanco JR, Chacón P. 2016. New generation of elastic network models. *Curr Opin Struct Biol.* 37:46–53.
Marcos ML, Echave J. 2015. Too packed to change: side-chain packing and site-specific substitution rates in protein evolution. *PeerJ* 3:e911.
Marsh JA, Teichmann SA. 2014. Parallel dynamics and evolution: protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *Bioessays* 36(2):209–218.
Mayrose I, Graur D, Ben-Tal N, Pupko T. 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol Biol Evol.* 21(9):1781–1791.
McCandlish DM, Stoltzfus A. 2014. Modeling evolution using the probability of fixation: history and implications. *Q Rev Biol.* 89(3):225–252.
Miller SR. 2017. An appraisal of the enzyme stability–activity trade-off. *Evolution* 71(7):1876–1887.
Ming D, Wall ME. 2005. Allostery in a coarse-grained model of protein dynamics. *Phys Rev Lett.* 95(19):198103.
Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol.* 247(4):536–540.
Nelson ED, Grishin NV. 2016. Long-range epistasis mediated by structural change in a model of ligand binding proteins. *PLoS One* 11(11): e0166739.
Nevin Gerek Z, Kumar S, Banu Ozkan S. 2013. Structural dynamics flexibility informs function and evolution at a proteome scale. *Evol Appl.* 6(3):423–433.
Perutz M, Kendrew J, Watson H. 1965. Structure and function of haemoglobin: II. Some relations between polypeptide chain configuration and amino acid sequence. *J Mol Biol.* 13(3):669–678.
Ramsey DC, Scherrer MP, Zhou T, Wilke CO. 2011. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics* 188(2):479–488.
Rodrigue N, Lartillot N. 2017. Detecting adaptation in protein-coding genes using a Bayesian site-heterogeneous mutation–selection codon substitution model. *Mol Biol Evol.* 34(1):204–214.
Scherrer MP, Scherrer MP, Meyer AG, Meyer AG, Wilke CO, Wilke CO. 2012. Modeling coding-sequence evolution within the context of residue solvent accessibility. *BMC Evol Biol.* 12:179.
Schowen RL. 1978. Catalytic power and transition-state stabilization. In Gandour RD, Schowen RL, editors. Transition states of biochemical processes. Boston: Springer, chapter 20. p. 77–144.
Shahmoradi A, Sydykova DK, Spielman SJ, Jackson EL, Dawson ET, Meyer AG, Wilke CO. 2014. Predicting evolutionary site variability from structure in viral proteins: buriedness, packing, flexibility, and design. *J Mol Evol.* 79(3–4):130–142.
Shahmoradi A, Wilke CO. 2016. Dissecting the roles of local packing density and longer-range effects in protein sequence evolution. *Proteins Struct Funct Bioinformatics.* 84(6):841–854.
Sharir-Ivry A, Xia Y. 2017. The impact of native state switching on protein sequence evolution. *Mol Biol Evol.* 34(6):1378–1390.
Shih C-H, Chang C-M, Lin Y-S, Lo W-C, Hwang J-K. 2012. Evolutionary information hidden in a single protein structure. *Proteins* 80(6):1647–1657.
Spielman SJ, Pond SLK. 2018. Relative evolutionary rates in proteins are largely insensitive to the substitution model. *Mol Biol Evol.*35:2307–2317.
Spiess AN, Neumeyer N. 2010. An evaluation of R^2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. *BMC Pharmacol.* 10:6.
Stein RL. 2011. Kinetics of enzyme action: essential principles for drug hunters. Hoboken (NJ): Wiley.
Sydykova DK, Wilke CO. 2017. Calculating site-specific evolutionary rates at the amino-acid or codon level yields similar rate estimates. *PeerJ* 5:e3391.
Taverna DM, Goldstein RA. 2002. Why are proteins marginally stable? *Proteins Struct Funct Genet.* 46(1):105–109.
Torrance JW, Bartlett GJ, Porter CT, Thornton JM. 2005. Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. *J Mol Biol.* 347(3):565–581.
Webb EC. 1992. Enzyme nomenclature. San Diego (CA): Wiley.
Yeh SW, Huang TT, Liu JW, Yu SH, Shih CH, Hwang JK, Echave J. 2014. Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. *Biomed Res Int.* 2014:572409.
Yeh SW, Liu JW, Yu SH, Shih CH, Hwang JK, Echave J. 2014. Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure. *Mol Biol Evol.* 31(1):135–139.