Too packed to change: site-specific substitution rates and side-chain packing in protein evolution

María Laura Marcos and Julian Echave

Escuela de Ciencia y Tecnología, Universidad Nacional de San Martín, Martín de Irigoyen 3100, 1650 San Martín, Buenos Aires, Argentina

ABSTRACT

In protein evolution, due to functional and biophysical constraints, the rates of amino acid substitution differ from site to site. Among the best predictors of site-specific rates is packing density. The packing density measure that best correlates with rates is the weighted contact number (WCN), the sum of inverse square distances between the site's C_α and the other C_α. According to a mechanistic stress model proposed recently, rates are determined by packing because mutating packed sites stresses and destabilizes the protein's active conformation. While WCN is a measure of C_α packing, mutations replace side chains, which prompted us to consider whether a site's evolutionary divergence is constrained by main-chain packing or side-chain packing. To address this issue, we extended the stress theory to model side chains explicitly. The theory predicts that rates should depend solely on side-chain packing. We tested these predictions on a data set of structurally and functionally diverse monomeric enzymes. We found that, on average, side-chain contact density (WCN_ρ) explains 39.1% of among-sites rate variation, larger than main-chain contact density (WCN_α) which explains 32.1%. More importantly, the independent contribution of WCN_α is only 0.7%. Thus, as predicted by the stress theory, site-specific evolutionary rates are determined by side-chain packing.

Keywords: protein evolution, rate variation among sites, structural constraints, packing, contact density, side chain

INTRODUCTION

Why do some protein sites evolve more slowly than others? Protein evolution is driven by random mutations and shaped by natural selection (Liberles et al., 2012; Sikosek and Chan, 2014). Mutations are selected depending on their impact on functional properties, such as the chemical nature of catalytic residues, active site conformation, and the protein's ability to fold rapidly and stably. Since changes of these properties depend on the mutated site, amino acid substitution rates vary from site to site.

We can reformulate the question opening the previous paragraph: What *specific properties* account for site-dependent rates of evolution? Substitution rates have been found to correlate with several properties. Amongst the best predictors are solvent accessibility (Bustamante et al., 2000; Conant and Stadler, 2009; Franzosa and Xia, 2009; Ramsey et al., 2011; Shahmoradi et al., 2014) and stability changes (Echave et al., 2014). For a large data set of enzymes, it was found that the main structural determinant is the *weighted contact number* WCN (Shih and Hwang, 2012; Yeh et al., 2014a,b). A site's WCN is the sum of inverse square distances from its C_α to the C_αs of other sites, therefore, it is a measure of C_α packing density.

The relationship between WCN and substitution rates can be understood in terms of a mechanistic stress model of protein evolution (Huang et al., 2014). Given an ancestral wild-type protein, the model assumes that its native conformation is the active conformation. Mutating a site perturbs (stresses) its interactions with other sites, destabilizing the active conformation. Such a destabilization determines the probability of the mutation being accepted or rejected, and therefore the rate of amino acid substitutions. Using the parameter-free Anisotropic Network Model (Yang et al., 2009), the expected destabilization was found to be proportional to WCN, and site-specific substitution rates were predicted to decrease linearly with increasing WCN, in agreement with observations.

So far, substitution rate vs. WCN studies were based on main chain (C_α) packing (Shih and Hwang, 2012; Yeh et al., 2014a; Huang et al., 2014). However, mutations replace side chains. Consider a protein residue, e.g. Thr93 of Human Carbonic Anhidrase II (pdb code 1CA2) (Fig. 1). The environment of the main chain (panel A) differs from that of the side chain (panel B). When Thr93 is mutated, what environment would determine whether the mutation is accepted or rejected? More specifically: Do site-specific substitution rates depend on main-chain packing or on side-chain packing? To address this issue, we extended the stress model to consider main and side chains explicitly, we derived substitution rates as a function of packing, and tested the theory on a data set of monomeric enzymes.

METHODS

The stress model

The stress model provides a mechanism for the observed correlation between rates and packing density (Huang et al., 2014). The model is based on the idea that a mutant is viable to the extent that it spends time in the active conformation. When a site is mutated, the interactions with its neighbors are perturbed (stressed), which destabilizes the active conformation by an amount δV^*, the *local mutational stress*. Mutational stress is related to site-specific evolutionary rates:

$$K^i \propto -\langle \delta V^* \rangle^i , \tag{1}$$

i.e. the substitution rate of site i, K^i, decreases linearly with the *mean local mutational stress*, $\langle \delta V^* \rangle^i$ (δV^* averaged over mutations at i). (1) is the main equation of the stress theory.

To calculate the mutational stress, we need an energy function. Huang et al. (2014) used the *parameter-free Anisotropic Network Model* (pfANM) of (Yang et al., 2009), which models the protein using an elastic network where each residue is represented by a node placed at its C_α. Pairs of nodes are connected by springs with force constants $k_{ij} = 1/{d_{ij}^0}^2$, where d_{ij}^0 is the distance between C_{α_i} and C_{α_j} in the active conformation. Following (Echave, 2008; Echave and Fernández, 2010), mutations are modeled as random perturbations of the lengths of the springs connected to the mutated site, which leads to:

$$K^i \propto -\mathrm{WCN}^i, \tag{2}$$

where

$$\mathrm{WCN}^i = \sum_{j \neq i} \frac{1}{{d_{ij}^0}^2} \tag{3}$$

is the weighted contact number introduced by Lin et al. (2008) and found to be among the best structural predictors of site-dependent evolutionary rates (Yeh et al., 2014a,b). Thus, according to the stress theory combined with the C_α-based pfANM, substitution rates should decrease linearly with WCN.

Since point mutations replace *side chains*, including them explicitly might improve the predictions of the stress theory. To explore this possibility, we model the protein as an elastic network where each residue is represented by two nodes, one for the main chain, α, placed at the residue's C_α, and another for the side chain, ρ, placed at the side-chain geometric center (only α nodes for Glycines). Mutations affect only the side chain of the mutated site. We model them adding random perturbations to the lengths of the springs connected to the mutated site. Assuming, as before, that the force constant of the spring connecting nodes n_i and n_j (n is α or ρ) is $k_{n_i n_j} = 1/{d_{n_i n_j}^0}^2$, it follows that:

$$K^i \propto -\mathrm{WCN}_\rho^i, \tag{4}$$

where

$$\mathrm{WCN}_\rho^i = \sum_{j \neq i} \left(\frac{1}{{d_{\rho_i \alpha_j}^0}^2} + \frac{1}{{d_{\rho_i \rho_j}^0}^2} \right). \tag{5}$$

WCN_ρ, defined here, is the weighted contact number of the side chain. Thus, when using the pfANM based on main chain nodes α and side-chain nodes ρ, the stress model predicts that site-specific rates will

depend only on the contact density of the side chain WCN_ρ. To check this prediction, we also consider the main chain weighted contact number:

$$\mathrm{WCN}_\alpha^i = \sum_{j\neq i}\left(\frac{1}{{d_{\alpha_i\alpha_j}^0}^2}+\frac{1}{{d_{\alpha_i\rho_j}^0}^2}\right), \tag{6}$$

According to the stress model, main-chain packing should not contribute independently to substitution rates.

In this section, we briefly presented the main results of the stress theory. A more detailed derivation can be found in the Appendix and in (Huang et al., 2014).

Dataset and comparison of empirical and predicted rates

To test our theory, we used the data set of (Echave et al., 2014). The set consists of 209 monomeric enzymes of known structure covering diverse structural and functional classes. Each structure is accompanied by up to 300 homologous sequences.

We used the empirical site-specific rates of evolution of (Echave et al., 2014). They were calculated as follows. First, the homologous sequences for each structure were aligned using MAFFT (Multiple Alignment using Fast Fourier Transform) (Katoh et al., 2005; Katoh and Standley, 2013). Second, using the resulting alignments as input, Maximum Likelihood phylogenetic trees were inferred with RAxML (Randomized Axelerated Maximum Likelyhood), using the LG substitution matrix (named after Le and Gacuel) and the CAT model of rate heterogeneity (Stamatakis, 2014). Third, the alignment and phylogenetic tree for each structure was used as input of Rate4Site to obtain the site-specific rates of substitution using the empirical Bayesian method and the amino-acid Jukes-Cantor mutational model (aaJC) (Mayrose et al., 2004). Finally, site-specific *relative* rates were obtained by dividing site-specific rates by their average over all sites of the protein. We denote the empirical rates by K_{R4S}.

For each protein, we calculated three packing density measures and predicted rates using linear fits. For brevity, we will use the shorthand $y\sim x$ for one-variable linear fits and $y\sim x_1+x_2$ for two-variable fits. Using the protein's pdb structure, we calculated WCN, WCN_α, and WCN_ρ, using (3), (6), and (5), respectively. Then, we calculated predicted rates by fitting $K\sim\mathrm{WCN}$, $K\sim\mathrm{WCN}_\alpha$, and $K\sim\mathrm{WCN}_\rho$ to the set of empirical rates. We also considered the two-variable fit $K\sim\mathrm{WCN}_\alpha+\mathrm{WCN}_\rho$. The goodness of fit of each model was assessed using R^2, the square correlation coefficients between predicted and empirical rates.

For statistical analysis we used R (R Core Team, 2014). For linear fits we used the built-in function `lm()`. Correlation coefficients were calculated using `cor()`. Binomial tests were performed `binom.test()`.

RESULTS AND DISCUSSION

According to the stress model, site-specific substitution rates depend only on side-chain packing. Main chain packing should not be directly related to substitution rates. To test this theory, we compared rate predictions based on main-chain packing and side-chain packing for a data set of 209 diverse monomeric enzymes.

Consider, for example, Human Carbonic Anhidrase II (pdb code 1CA2). Empirical rates K_{R4S} were obtained from the multiple sequence alignment as described in Methods. Using the pdb structure, we calculated the packing measures WCN, WCN_α, and WCN_ρ, using (3), (6), and (5), respectively. We used these packing measures to predict rates using linear fits to empirical rates, as described in Methods. As we mentioned in the Introduction, main chain environments and side-chain environments are different (Fig. 1). Accordingly, WCN_α and WCN_ρ result in different predicted rates (Fig. 2). The two site-dependent profiles of predicted rates are similar to the empirical K_{R4S} profile. WCN_ρ-based predictions look better (Fig. 2) and are better (Fig. 3): the R^2 values are 0.41 for WCN_α and 0.56 for WCN_ρ. R^2 increases only by 0.02 for the two-variable fit $K\sim\mathrm{WCN}_\alpha+\mathrm{WCN}_\rho$ ($R^2=0.58$). Since WCN (Eq. (3)) was, so far, the best structural predictor of site-specific rates for enzymes (Yeh et al., 2014a,b), we also calculated $R^2(K_{\mathrm{R4S}},\mathrm{WCN})$: it is 0.40. To summarize, for 1CA2, $\mathrm{WCN}_\rho>\mathrm{WCN}_\alpha\gtrsim\mathrm{WCN}$; the best predictor of site-specific rates is WCN_ρ. Moreover, WCN_α has only a small independent effect on substitution rates.

We repeated the previous assessment for each protein of the data set. For each of the 209 enzymes, we calculated the densities WCN, WCN_α, and WCN_ρ and calculated predicted rates from linear fits to

empirical rates K_{R4S}. R^2 values averaged over all proteins are 0.316, 0.321, and 0.391 for WCN, WCN_α, and WCN_ρ, respectively (Fig. 4). Thus, as for 1CA2, the predictive power of single-variable fits follows $\mathrm{WCN}_\rho > \mathrm{WCN}_\alpha \gtrsim \mathrm{WCN}$. When going from $K \sim \mathrm{WCN}_\rho$ to $K \sim \mathrm{WCN}_\alpha + \mathrm{WCN}_\rho$, R^2 increases from 0.391 to 0.398, only a 0.7% increase in explained variance (Fig. 4). Therefore, on average, WCN_ρ is the best predictor of site-specific substitution rates.

Beyond average R^2, we performed a protein-by-protein comparison (Fig. 5). We found that WCN_ρ is a better predictor than WCN_α for 204 of the 209 proteins studied ($p \ll 10^{-3}$, binomial test). Similarly, WCN_ρ outperforms WCN for 206/209 proteins ($p \ll 10^{-3}$, binomial test). Thus, WCN_ρ is the best rate predictor for almost all proteins of the data set.

To summarize, side-chain contact density (WCN_ρ) is the best predictor of site-specific substitution rates, accounting, on average, for 39.1% of the rate variation among sites. In contrast, the independent contribution of main-chain contact density (WCN_α) is negligible (0.7%). These results are consistent with the predictions of the stress model, extended to include explicitly main chain and side chains. According to this theory, mutations replace side chains thus changing the parameters of interaction between the mutated side chain and the rest of the protein. WCN_ρ is proportional to the destabilization of the protein's active conformation, which is why it correlates with rates: mutations are accepted or rejected according to the degree of destabilization of the active conformation.

From a practical point of view, regardless of the validity of the stress theory, WCN_ρ outperforms WCN, that was, so far, the best structural predictor of site-specific substitution rates (Yeh et al., 2014a,b). Therefore, at least for the data set of monomeric enzymes used, WCN_ρ is the new best predictor of site-specific substitution rates. WCN_ρ could be used to improve structure-based empirical models of protein evolution and phylogenetic inference (see e.g. (Kleinman et al., 2010)).

REFERENCES

Bustamante, C. D., Townsend, J. P., and Hartl, D. L. (2000). Solvent accessibility and purifying selection within proteins of Escherichia coli and Salmonella enterica. *Molecular biology and evolution*, 17(2):301–308.

Conant, G. C. and Stadler, P. F. (2009). Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Molecular biology and evolution*, 26(5):1155–1161.

Echave, J. (2008). Evolutionary divergence of protein structure: The linearly forced elastic network model. *Chemical physics letters*, 457(4-6):413–416.

Echave, J. and Fernández, F. M. (2010). A perturbative view of protein structural variation. *Proteins*, 78(1):173–180.

Echave, J., Jackson, E. L., and Wilke, C. O. (2014). Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. *bioRxiv*.

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

Huang, T.-T., Marcos, M. L., Hwang, J.-K., and 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.

Katoh, K., Kuma, K.-I., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucl. Acids Res.*, 33:511–518.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, 30:772–780.

Kleinman, C. L., Rodrigue, N., Lartillot, N., and Philippe, H. (2010). Statistical Potentials for Improved Structurally Constrained Evolutionary Models. *Molecular biology and evolution*, 27(7):1546–1560.

Liberles, D. A., Teichmann, S. A., Bahar, I., Bastolla, U., Bloom, J., Bornberg-Bauer, E., Colwell, L. J., de Koning, A. P. J., Dokholyan, N. V., Echave, J., Elofsson, A., Gerloff, D. L., Goldstein, R. A., Grahnen, J. A., Holder, M. T., Lakner, C., Lartillot, N., Lovell, S. C., Naylor, G., Perica, T., Pollock, D. D., Pupko, T., Regan, L., Roger, A., Rubinstein, N., Shakhnovich, E., Sjölander, K., Sunyaev, S., Teufel, A. I., Thorne, J. L., Thornton, J. W., Weinreich, D. M., and Whelan, S. (2012). The interface of protein structure, protein biophysics, and molecular evolution. *Protein science : a publication of the Protein Society*, 21(6):769–785.

Lin, C. P., Huang, S. W., Lai, Y. L., Yen, S. C., Shih, C. H., Lu, C. H., Huang, C. C., and Hwang,

J. K. (2008). Deriving protein dynamical properties from weighted protein contact number. *Proteins*, 72:929–935.

Mayrose, I., Graur, D., Ben-Tal, N., and Pupko, T. (2004). Comparison of site-specific rate-inference methods: Bayesian methods are superior. *Mol. Biol. Evol.*, 21:1781–1791.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsey, D. C., Scherrer, M. P., Zhou, T., and Wilke, C. O. (2011). The Relationship Between Relative Solvent Accessibility and Evolutionary Rate in Protein Evolution. *Genetics*, 188(2):479–488.

Shahmoradi, A., Sydykova, D. K., Spielman, S. J., Jackson, E. L., Dawson, E. T., Meyer, A. G., and Wilke, C. O. (2014). Predicting evolutionary site variability from structure in viral proteins: buriedness, packing, flexibility, and design. *Journal of molecular evolution*, 79(3-4):130–142.

Shih, C.-H. and Hwang, J.-k. (2012). Evolutionary information hidden in a single protein structure. *Proteins*, 80(6):1647–1657.

Sikosek, T. and Chan, H. S. (2014). Biophysics of protein evolution and evolutionary protein biophysics. *Journal of the Royal Society, Interface / the Royal Society*, 11(100):20140419–20140419.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30:1312–1313.

Yang, L., Song, G., and Jernigan, R. L. (2009). Protein elastic network models and the ranges of cooperativity. *Proc Natl Acad Sci USA*, 106:12347–52.

Yeh, S. W., Huang, T. T., Liu, J. W., Yu, S. H., Shih, C. H., Hwang, J. K., and Echave, J. (2014a). Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. *Biomed Res. Int.*, 2014:572409.

Yeh, S. W., Liu, J. W., Yu, S. H., Shih, C. H., Hwang, J. K., and Echave, J. (2014b). Site-specific structural constraints on protein sequence evolutionary divergence: Local packing density versus solvent exposure. *Mol. Biol. Evol.*, 31:135–139.

FIGURES

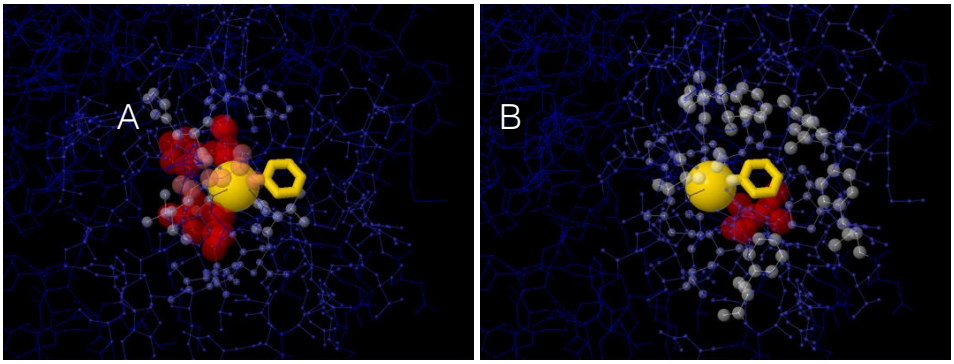

Figure 1. The two environments of a protein residue. Images of the environments of Thr93 of Human Carbonic Anhidrase II (pdb code 1CA2). (A) Environment of the main chain C_α: the size and colors of protein atoms increase with the inverse square distance to Thr93 C_α (gold ball). (B) Environment of the side chain: size and colors of atoms increase with the inverse square distance to the geometric center of Thr93 side chain (gold wireframe).

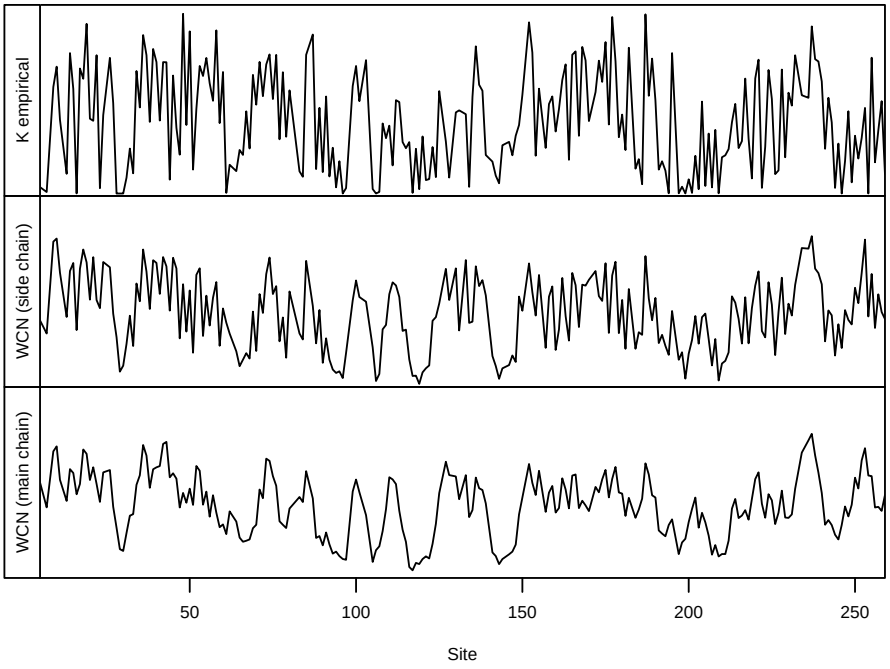

Figure 2. Profiles of site-specific evolutionary rates for 1CA2. (Top) empirical rates K_{R4S} inferred by Rate4Site. (Middle) Rates predicted from the side-chain contact density WCN_ρ. (Bottom) Rates predicted from the main-chain contact density WCN_α. The profile of WCN_ρ-predicted rates looks more similar to the K_{R4S} profile

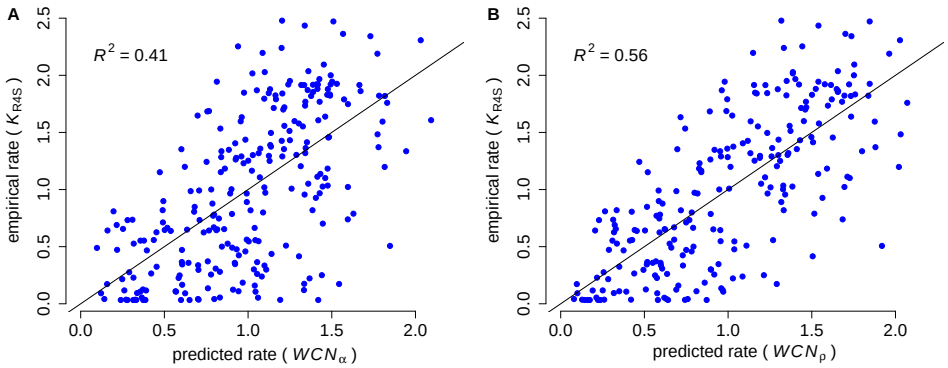

Figure 3. Empirical vs. predicted rates for 1CA2. (A) Empirical rates inferred using Rate4Site vs. rates predicted from the main-chain contact densities WCN_α. (B) Empirical rates vs. rates predicted from side-chain contact densities WCN_ρ. The "x=y" line corresponding to a perfect fit is shown. WCN_α explains $R^2 = 41\%$ of the variation of site-specific empirical rates, WCN_ρ explains 56%.

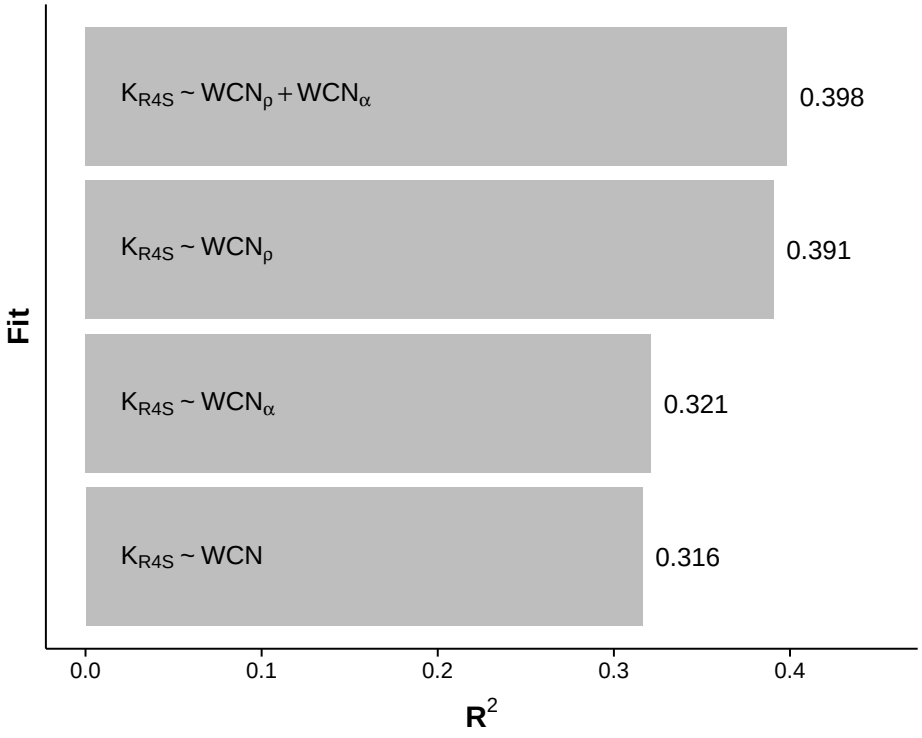

Figure 4. Side chain packing is the sole determinant of site-specific substitution rates. Linear regression of empirical rates (K_{R4S}) using three one-variable fits (WCN, WCN_α, and WCN_ρ) and a two variable fit $K_{\mathrm{R4S}} \sim \mathrm{WCN}_\alpha + \mathrm{WCN}_\rho$. WCN and WCN_α are measures of the contact density of main chain C_αs. WCN_ρ is the contact density of side chains, modeled by their geometric centers. The models are fit for each protein, and R^2 is the average R^2 over the 209 proteins of the data set. WCN_ρ is the best predictor. The independent contribution of $\mathrm{WCN}_\alpha$is very small (0.7%).

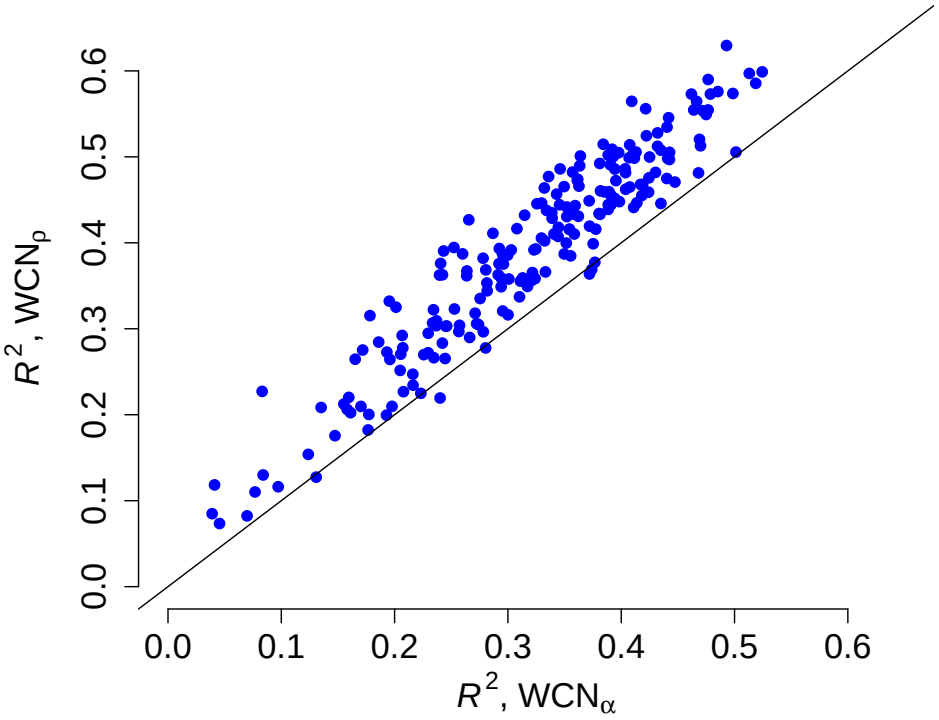

Figure 5. Side chain packing is the best predictor of substitution rates for most proteins. R^2 is the square correlation between empirical rates (K_{R4S}) and either side-chain contact density WCN$_\rho$ (y axis) or main-chain contact density WCN$_\alpha$ (x axis). Each point corresponds to one protein. Empirical rates correlate better with WCN$_\rho$ for 204 out of 209 proteins.

APPENDIX

The stress model

The stress model is based on the idea that a mutant is viable to the extent that it spends time in the active conformation. The fixation probability is modeled as

$$p_{\text{fix}} \propto \frac{C^F_{\text{mutant}}\rho_{\text{mutant}}(\mathbf{r}_{\text{active}})}{C^F_{\text{wt}}\rho_{\text{wt}}(\mathbf{r}_{\text{active}})} \tag{A-1}$$

where C^F is the concentration of folded protein and $\rho(\mathbf{r}_{\text{active}})$ is the probability of it adopting the active conformation. Assuming that $C_{\text{mutant}}/C_{\text{wt}}$ is equal to the ratio of partition functions, from basic statistical physics it follows that:

$$p_{\text{fix}} \propto e^{-\beta\delta V^*}, \tag{A-2}$$

where β can be thought of representing selection pressure rather than temperature and

$$\delta V^* = V_{\text{mutant}}(\mathbf{r}_{\text{active}}) - V_{\text{wt}}(\mathbf{r}_{\text{active}}) \tag{A-3}$$

is the energy difference between mutant and wild-type in the active conformation. Finally, assuming that $\beta\delta V^* \ll 1$ (weak selection), from (A-2) we find:

$$K^i \propto -\langle\delta V^*\rangle^i, \tag{A-4}$$

i.e. the rate of substitution of site i, K^i, is proportional to (minus) the destabilization energy averaged over mutations at i, $\langle\delta V^*\rangle^i$. This is the basic equation of the stress theory.

One-bead-per-site elastic network

To derive substitution rates, we need an energy function. Let us model the protein as an elastic network of nodes placed at C_αs connected by elastic springs. The energy of a conformation $\mathbf{r}$ is given by:

$$V(\mathbf{r}) = \frac{1}{2}\sum_i\sum_{j>i} k_{ij}(d_{ij} - d^0_{ij})^2, \tag{A-5}$$

where $d_{ij} = ||\mathbf{r}_j - \mathbf{r}_i||$ is the distance between C_{α_i} and C_{α_j}, k_{ij} is the force constant of spring $i-j$ and d^0_{ij} its equilibrium length.

The wild-type protein is modeled by using springs that are relaxed at the native conformation: $d^0_{ij} = ||\mathbf{r}^0_j - \mathbf{r}^0_i||$. To model a mutation at site i, we add random perturbations to the spring lengths connecting i other sites: $d^0_{ij} \to d^0_{ij} + \delta_{ij}$. Using (A-5) and (A-3), we find:

$$\delta V^* = \frac{1}{2}\sum_{j\neq i} k_{ij}\delta^2_{ij}, \tag{A-6}$$

where we have assumed that the active conformation is the native conformation of the wild type, $\mathbf{r}_{\text{active}} = \mathbf{r}^0_{wt}$, which is a reasonable assumption for purifying selection. Assuming that δ_{ij} for the different contacts are drawn independently from the same distribution, averaging (A-6) over mutations at site i we find

$$\langle\delta V^*\rangle^i \propto \sum_{j!=i} k_{ij} \tag{A-7}$$

Thus, the mutational destabilization averaged over mutations at a given site (the *mean local mutational stress*) is proportional to the sum of the force constants of the springs connected to the mutated site.

To obtain the site-specific rates, we use the parameter-free Anisotropic Network Model (pfANM):

$$k_{ij} = \frac{1}{{d^0_{ij}}^2}. \tag{A-8}$$

Replacing (A-8) into (A-7), and the result into (A-4), we obtain:

$$K^i \propto -\text{WCN}^i, \tag{A-9}$$

where

$$\mathrm{WCN}^i = \sum_{j \neq i} \frac{1}{{d^0_{ij}}^2} \tag{A-10}$$

is the weighted contact number. Thus, when using the pfANM based on C_α, the stress model predicts that site-specific rates are proportional to (minus) WCN.

Two beads-per-site elastic network

The elastic network of the previous section uses one node per site, thus modeling side chains only implicitly. Since mutations replace side chains, including them explicitly might improve the predictions of the stress theory.

Let us represent each site using two nodes: one for the main chain, α, placed at the residue's C_α as before, and another for the side chain, ρ, placed at the side-chain geometric center (Gly's are represented using only one node at C_α). The elastic energy is:

$$V(\mathbf{r}) = \frac{1}{2}\sum_i \sum_{j>i} k_{\alpha_i\alpha_j}(d_{\alpha_i\alpha_j} - d^0_{\alpha_i\alpha_j})^2 + \frac{1}{2}\sum_i \sum_{j>i} k_{\alpha_i\rho_j}(d_{\alpha_i\rho_j} - d^0_{\alpha_i\rho_j})^2 \tag{A-11}$$

$$+\frac{1}{2}\sum_i \sum_{j>i} k_{\rho_i\alpha_j}(d_{\rho_i\alpha_j} - d^0_{\rho_i\alpha_j})^2 + \frac{1}{2}\sum_i \sum_{j>i} k_{\rho_i\rho_j}(d_{\rho_i\rho_j} - d^0_{\rho_i\rho_j})^2, \tag{A-12}$$

$$\tag{A-13}$$

where d_{n_i,n_j} is the distance between nodes n_i and n_j (n is α or ρ), $k_{n_in_j}$ is the force constant of the spring connecting these nodes, and $d^0_{n_in_j}$ the equilibrium spring length.

A mutation at site i will replace ρ_i, affecting only the parameters of the energy function related to this node. Modeling a mutation at i by adding random perturbations to the springs of ρ_i: $d^0_{\rho_i\rho_j} \to d^0_{\rho_i\rho_j} + \delta_{\rho_i\rho_j}$ and $d^0_{\rho_i\alpha_j} \to d^0_{\rho_i\alpha_j} + \delta_{\rho_i\alpha_j}$, we find:

$$\delta V^* = \frac{1}{2}\sum_{j!=i} (k_{\rho_i\alpha_j}\delta^2_{\rho_i\alpha_j} + k_{\rho_i\rho_j}\delta^2_{\rho_i\rho_j}). \tag{A-14}$$

where δV^* is defined in (A-3). Assuming perturbations are drawn independently from the same distribution, averaging (A-14) over mutations at i we find:

$$\langle \delta V^* \rangle^i \propto \sum_{j\neq i} (k_{\rho_i\alpha_j} + k_{\rho_i\rho_j}). \tag{A-15}$$

Finally, assuming that $k_{n_in_j} = \frac{1}{{d^0_{n_in_j}}^2}$, from (A-4) and (A-15) we obtain:

$$K^i \propto -\mathrm{WCN}^i_\rho, \tag{A-16}$$

where

$$\mathrm{WCN}^i_\rho = \sum_{j\neq i} \left(\frac{1}{{d^0_{\rho_i\alpha_j}}^2} + \frac{1}{{d^0_{\rho_i\rho_j}}^2} \right). \tag{A-17}$$

WCN_ρ, defined here, is the side-chain weighted contact number. Thus, when using the pfANM based on main chain nodes α and side-chain nodes ρ, the stress model predicts that site-specific rates will depend on the contact density of the side chain WCN_ρ.

For the sake of the present study, we also consider whether main-chain packing has any independent effect on rates. For this purpose, we calculate

$$\mathrm{WCN}^i_\alpha = \sum_{j\neq i} \left(\frac{1}{{d^0_{\alpha_i\alpha_j}}^2} + \frac{1}{{d^0_{\alpha_i\rho_j}}^2} \right), \tag{A-18}$$

which is the main chain weighted contact number for the two-beads-per-site network model.