PeerJ

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

María Laura Marcos and Julian Echave

Escuela de Ciencia y Tecnología, Universidad Nacional de San Martín, 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 are solvent accessibility and packing density. The packing density measure that best correlates with rates is the weighted contact number (WCN), the sum of inverse square distances between a site's C_α and the C_α of the other sites. 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*. Here, we 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 contact density. We tested this prediction on a data set of structurally and functionally diverse monomeric enzymes. We compared side-chain contact density with main-chain contact density measures and with relative solvent accessibility (RSA). We found that side-chain contact density is the best predictor of rate variation among sites (it explains 39.2% of the variation). Moreover, the independent contribution of main-chain contact density measures and RSA are negligible. Thus, as predicted by the stress theory, site-specific evolutionary rates are determined by side-chain packing.

Subjects Biophysics, Computational Biology, Evolutionary Studies, Mathematical Biology
Keywords Protein evolution, Structural constraints, Packing, Contact density, Rate variation among sites, Side chain

Submitted 31 December 2014
Accepted 4 April 2015
Published 23 April 2015

Corresponding author
Julian Echave,
julian.echave@unsam.edu.ar

Academic editor
Rafael Najmanovich

Additional Information and Declarations can be found on page 10

DOI **10.7717/peerj.911**

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 & 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? The most studied predictors

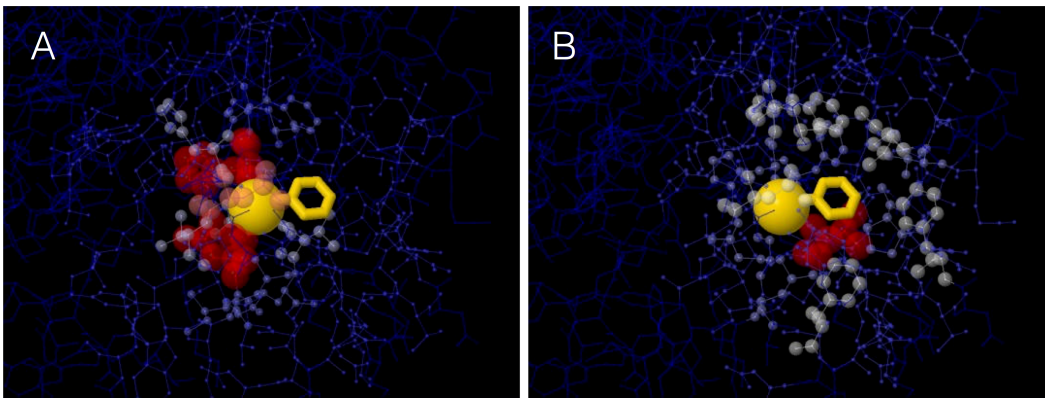

Figure 1 The two environments of a protein residue. Images of the environments of Thr93 of Human Carbonic Anhidrase II (pdb code 1CA2). (A) Environment of the main chain C_α: the size and colors of protein atoms increase with the inverse square distance to Thr93 C_α (gold ball). (B) Environment of the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

are structural site-specific properties (*Franzosa & Xia, 2009*). For years, the main structural predictor was believed to be *solvent accessibility*, as quantified by the Relative Solvent Accessibility (RSA) (*Bustamante, Townsend & Hartl, 2000*; *Conant & Stadler, 2009*; *Franzosa & Xia, 2009*; *Ramsey et al., 2011*; *Shahmoradi et al., 2014*). However, *local packing density*, quantified by the Weighted Contact Number (WCN), predicts evolutionary rates at least as well as RSA (*Shih & Hwang, 2012*; *Yeh et al., 2014a*; *Yeh et al., 2014b*).

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 substitution. Using the energy function of the parameter-free Anisotropic Network Model (*Yang, Song & Jernigan, 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.

A site's WCN is the sum of inverse square distances from its C_α to the C_α of other sites: it is a measure of C_α packing density. Therefore, previous substitution rate vs. WCN studies were based on *main chain* (C_α) packing (*Shih & 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 (Fig. 1A) differs from that of the side chain (Fig. 1B). 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 and we theoretically derived that substitution rates depend only on side-chain of

packing. We tested the theory on a data set of monomeric enzymes. In agreement with predictions, site-specific substitution rates correlate better with side-chain packing than with main-chain packing measures and RSA. Moroever, partialing out the effect of side-chain packing, the independent contributions of main-chain packing and RSA are negligible.

METHODS

Theory

In this section, we show that the mechanistic stress model of protein evolution predicts that the substitution rate of a protein site is determined by the packing density of its side chain. This prediction and its empirical assessment are the point of this paper.

The stress model was proposed by *Huang et al. (2014)* to explain the observed correlation between site-specific substitution rates and packing density. The model is based on the idea that a mutant is viable to the extent that it spends time in the active conformation. In turn, this time will depend on mutational changes of the stability of the active conformation. The fixation probability of a mutant is modeled as

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

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

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

where β represents the selection pressure and

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

is the energy difference between mutant and wild-type in the active conformation.

Assuming that $\beta \delta V^* \ll 1$ (weak selection), from (2) we find:

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

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

In *Huang et al. (2014)*, mutational stability changes were calculated using an elastic network model in which each residue is represented by a single node. Within such a one-node-per-residue representation, there is no differentiation between main chain and side chain. Therefore, we cannot *predict* whether evolutionary rates will be determined by main chain packing or side chain packing. To address this issue, here we represent each residue using *two nodes*: a main-chain node α, placed at the residue's C_α, and a side-chain node ρ, placed at the geometric center of the residue's side chain (Gly's are represented

using only one node at C_α). The energy function is:

$$V(\mathbf{r}) = \frac{1}{2}\sum_i\sum_{j>i} k_{\alpha_i\alpha_j}(r_{\alpha_i\alpha_j} - d_{\alpha_i\alpha_j})^2 + \frac{1}{2}\sum_i\sum_{j>i} k_{\alpha_i\rho_j}(r_{\alpha_i\rho_j} - d_{\alpha_i\rho_j})^2$$
$$+\frac{1}{2}\sum_i\sum_{j>i} k_{\rho_i\alpha_j}(r_{\rho_i\alpha_j} - d_{\rho_i\alpha_j})^2 + \frac{1}{2}\sum_i\sum_{j>i} k_{\rho_i\rho_j}(r_{\rho_i\rho_j} - d_{\rho_i\rho_j})^2, \quad (5)$$

where $r_{n_in_j}$ 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 $d_{n_in_j}$ 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. We emphasize: while the mutation may well induce global structural changes involving the backbone and other side chains, the only *parameters* that will change are those of the mutated side chain. Following *Echave (2008)* and *Echave & Fernández (2010)*, we model a mutation at i by adding random perturbations to the lengths of the springs connected to ρ_i: $d_{\rho_i\rho_j} \to d_{\rho_i\rho_j} + \delta_{\rho_i\rho_j}$ and $d_{\rho_i\alpha_j} \to d_{\rho_i\alpha_j} + \delta_{\rho_i\alpha_j}$, to find, using (3) and (5):

$$\delta V^* = \frac{1}{2}\sum_{j\neq i}(k_{\rho_i\alpha_j}\delta^2_{\rho_i\alpha_j} + k_{\rho_i\rho_j}\delta^2_{\rho_i\rho_j}). \quad (6)$$

Assuming that perturbations are drawn independently from the same distribution, averaging (6) over mutations at i we find:

$$\langle \delta V^* \rangle^i \propto \sum_{j\neq i}(k_{\rho_i\alpha_j} + k_{\rho_i\rho_j}). \quad (7)$$

To finish, we assume, as in the parameter-free Anisotropic Network Model (pfANM) of *Yang, Song & Jernigan (2009)*, that $k_{n_in_j} = \frac{1}{d^2_{n_in_j}}$. Then, from (4) and (7) we obtain:

$$K^i \propto -\mathrm{WCN}^{\alpha\rho}_{\rho}, \quad (8)$$

where

$$\mathrm{WCN}^{\alpha\rho}_{\rho} = \sum_{j\neq i}\left(\frac{1}{d^2_{\rho_i\alpha_j}} + \frac{1}{d^2_{\rho_i\rho_j}}\right). \quad (9)$$

$\mathrm{WCN}^{\alpha\rho}_{\rho}$, derived here, is the side-chain weighted contact number. It depends on contacts between node ρ of the site considered (subscript) and nodes α and ρ of the other sites (superscript). Therefore, the stress model, combined with a two-nodes-per-site pfANM energy function, predicts that site-specific rates will depend on the contact density of the side chain $\mathrm{WCN}^{\alpha\rho}_{\rho}$.

By analogy with (9) we can calculate the main-chain weighted contact number:

$$\mathrm{WCN}^{\alpha\rho}_{\alpha} = \sum_{j\neq i}\left(\frac{1}{d^2_{\alpha_i\alpha_j}} + \frac{1}{d^2_{\alpha_i\rho_j}}\right). \quad (10)$$

We expect $\mathrm{WCN}_{\alpha}^{\alpha\rho}$ to correlate with $\mathrm{WCN}_{\rho}^{\alpha\rho}$, which may result in indirect correlations with substitution rates. However, if the stress model is correct, rates will be determined only by $\mathrm{WCN}_{\rho}^{\alpha\rho}$ and there should not be any *independent* effect of $\mathrm{WCN}_{\alpha}^{\alpha\rho}$.

Other structural predictors

To assess the prediction of the previous section, we also consider the following structural properties. First, the Weighted Contact Number WCN, which was introduced by *Lin et al. (2008)* and found to be among the best structural predictors of site-dependent evolutionary rates (*Yeh et al., 2014a*; *Yeh et al., 2014b*). It is defined as:

$$\mathrm{WCN} = \mathrm{WCN}_{\alpha}^{\alpha} = \sum_{j \neq i} \frac{1}{d_{\alpha_i \alpha_j}^2} \tag{11}$$

where $d_{\alpha_i \alpha_j}$ is the distance between the the alpha carbons of sites i and j. For the sake of clarity, wherever it is convenient we will use the notation $\mathrm{WCN}_{\alpha}^{\alpha}$ to make explicit that the distances between the C_α of a site (subscript) and the C_α of the other sites (superscript) are considered. Therefore, $\mathrm{WCN}_{\alpha}^{\alpha}$ can be considered a measure of main-chain packing density (based only on C_α–C_α interactions).

Second, by analogy with (11) we can use side-chain centers of mass ρ rather than C_α to define:

$$\mathrm{WCN}_{\rho}^{\rho} = \sum_{j \neq i} \frac{1}{d_{\rho_i \rho_j}^2} \tag{12}$$

$\mathrm{WCN}_{\rho}^{\rho}$ quantifies the packing density of the side chain including only ρ–ρ interactions.

Finally, we also consider the Relative Solvent Accessibility, RSA, which is the most studied structural determinant of evolutionary rates. The RSA of a residue is obtained by dividing its area accessible to the solvent (SA) by the maximum SA for the given amino acid type (*Tien et al., 2013*).

Dataset and empirical substitution rates

To test our theory, we used the data set of *Echave, Jackson & Wilke (2015)*. 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, Jackson & Wilke (2015)*. 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 & Standley, 2013*). Second, using the resulting alignments as input, Maximum Likelihood phylogenetic trees were inferred with RAxML (Randomized Axelerated Maximum Likelihood), using the LG substitution matrix (named after Le and Gascuel) 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}.

Comparison of empirical rates with structural properties

For each protein of the dataset, we used the pdb structure to calculate the five site-dependent structural properties defined above: $\mathrm{WCN}_{\rho}^{\alpha\rho}$, $\mathrm{WCN}_{\alpha}^{\alpha\rho}$, $\mathrm{WCN}_{\rho}^{\rho}$, $\mathrm{WCN}_{\alpha}^{\alpha}$ (= WCN), and RSA. For a given predictor x, we quantified its predictive power using the squared Pearson correlation coefficient $R^2(K_{\mathrm{R4S}}, x)$. According to the theoretical predictions, $\mathrm{WCN}_{\rho}^{\alpha\rho}$ should be the sole determinant of site-specific rates. We quantified the *independent* contribution of each of the other structural descriptors by partialing out the effect of $\mathrm{WCN}_{\rho}^{\alpha\rho}$ using semipartial correlations. The squared semipartial correlation $\rho^2(K_{\mathrm{R4S}}, x|\mathrm{WCN}_{\rho}^{\alpha\rho})$ represents the *unique* contribution of predictor x. Also, it is the amount by which the explained variation of K_{R4S} (R^2) would increase when going from the single-variable linear fit $K \sim \mathrm{WCN}_{\rho}^{\alpha\rho}$ to the two-variable fit $K \sim \mathrm{WCN}_{\rho}^{\alpha\rho} + x$. Expected values, standard deviations, and p-values were obtained by averaging protein correlations and semipartial correlations for 10,000 bootstrapped replicas of the dataset of 209 proteins.

For statistical analysis we used R (*R Core Team, 2014*). Correlation coefficients and their p-values were calculated using `cor.test()`. Semipartial correlation coefficients and p-values were calculated using `spcor.test()`. For bootstrapping with used `boot()` with default options.

RESULTS AND DISCUSSION

We theoretically derived a new measure of contact density, the side-chain weighted contact number $\mathrm{WCN}_{\rho}^{\alpha\rho}$ which, according to the stress model, should be the sole structural determinant of site-specific evolutionary rates. We tested this prediction on a dataset of 209 functionally and structurally diverse monomeric enzymes. Empirical site-specific evolutionary rates K_{R4S} were obtained from multiple sequence alignments using Rate4Site. We compared K_{R4S} with $\mathrm{WCN}_{\rho}^{\alpha\rho}$ (side-chain weighted contact number), $\mathrm{WCN}_{\alpha}^{\alpha\rho}$ (main-chain weighted contact number), $\mathrm{WCN}_{\rho}^{\rho}$ (side-chain $\rho-\rho$ weighted contact number), $\mathrm{WCN}_{\alpha}^{\alpha} = \mathrm{WCN}$ (main-chain $\alpha-\alpha$ weighted contact number), and RSA (relative solvent accessibility). For each protein, we calculated correlation coefficients between K_{R4S} and each structural property and semipartial correlations to measure independent contributions. Protein-by-protein results (Table S1) were averaged over all proteins to obtain expected values, using bootstrapping to estimate standard deviations and p-values (see Methods).

Side-chain contact density ($\mathrm{WCN}_{\rho}^{\alpha\rho}$) vs. main-chain contact density ($\mathrm{WCN}_{\alpha}^{\alpha\rho}$)

According to the stress model, site-specific substitution rates depend only on side-chain packing, so that main-chain packing should not be directly related to substitution rates. To test this prediction, we compared empirical substitution rates K_{R4S} with $\mathrm{WCN}_{\rho}^{\alpha\rho}$ and $\mathrm{WCN}_{\alpha}^{\alpha\rho}$.

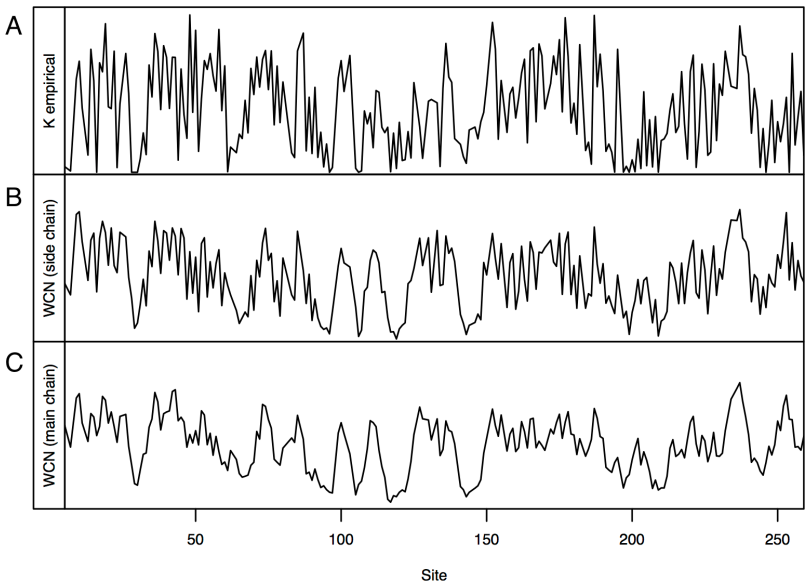

Figure 2 Profiles of site-specific evolutionary rates for 1CA2. (A) empirical rates K_{R4S} inferred by Rate4Site. (B) Rates predicted from the side-chain contact density ($WCN_{\rho}^{\alpha\rho}$) . (C) Rates predicted from the main-chain contact density ($WCN_{\alpha}^{\alpha\rho}$). Both predicted profiles look similar to the K_{R4S} profile. However, the $WCN_{\rho}^{\alpha\rho}$ profile is somewhat better (The $WCN_{\alpha}^{\alpha\rho}$ profile is too smooth).

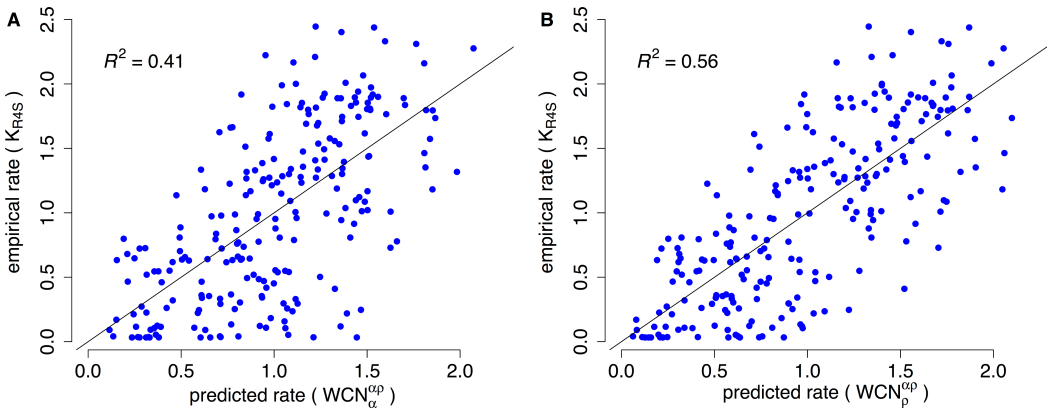

Figure 3 Empirical vs. predicted rates for 1CA2. (A) Empirical rates inferred using Rate4Site vs. rates predicted from the main-chain contact densities ($WCN_{\alpha}^{\alpha\rho}$) (B) Empirical rates vs. rates predicted from side-chain contact densities ($WCN_{\rho}^{\alpha\rho}$). The "$x = y$" line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

Consider, for example, Human Carbonic Anhidrase II (pdb code 1CA2). As we mentioned in the Introduction, main chain environments and side-chain environments are different (Fig. 1). Accordingly, $WCN_{\rho}^{\alpha\rho}$ and $WCN_{\alpha}^{\alpha\rho}$ result in different predicted rates (Fig. 2). The two site-dependent profiles of predicted rates are similar to the empirical K_{R4S} profile. However, $WCN_{\rho}^{\alpha\rho}$-based predictions look somewhat better (Fig. 2) and are better (Fig. 3): the R^2 values are 0.56 for $WCN_{\rho}^{\alpha\rho}$ and 0.41 for $WCN_{\alpha}^{\alpha\rho}$. Thus, for 1CA2 $WCN_{\rho}^{\alpha\rho}$ outperforms $WCN_{\alpha}^{\alpha\rho}$ as predictor of evolutionary rates.

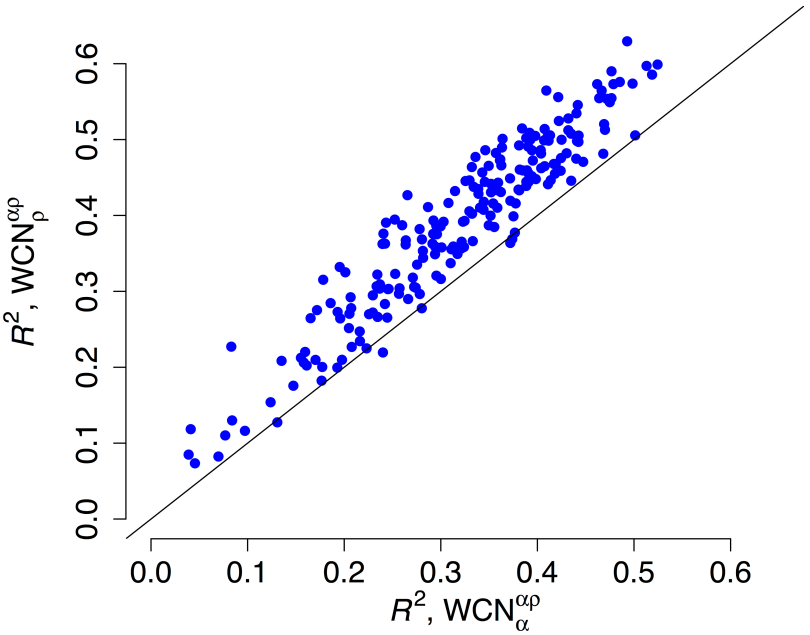

Figure 4 Side chain packing is the best predictor of substitution rates for most proteins. R^2 is the square correlation between empirical rates (K_{R4S}) and either side-chain contact density ($WCN_\rho^{\alpha\rho}$) (y-axis) or main-chain contact density ($WCN_\alpha^{\alpha\rho}$) (x-axis). Each point corresponds to one protein. Empirical rates correlate better with ($WCN_\rho^{\alpha\rho}$) for 204 out of 209 proteins.

We repeated the previous assessment for each of the 209 enzymes of the data set (Fig. 4). Empirical rates correlate with $WCN_\rho^{\alpha\rho}$ better than with $WCN_\alpha^{\alpha\rho}$ for 204 of the 209 proteins studied. Aggregating the data over all proteins, we obtained expected R^2 values of 0.392 ± 0.008 and 0.319 ± 0.008 for $WCN_\rho^{\alpha\rho}$ and $WCN_\alpha^{\alpha\rho}$, respectively. The difference $\Delta R^2 = 0.073 \pm 0.003$ is significantly positive ($p < 10^{-3}$, bootstrapping). Therefore, $WCN_\rho^{\alpha\rho}$ outperforms $WCN_\alpha^{\alpha\rho}$ as predictor of site-specific substitution rates.

The stress model predicts $WCN_\rho^{\alpha\rho}$ to be the *sole* predictor of substitution rates. Any correlation between rates and $WCN_\alpha^{\alpha\rho}$ should be indirect. We measured the direct association between empirical rates and $WCN_\alpha^{\alpha\rho}$ using the squared semipartial correlation $\rho^2(K_{R4S}, WCN_\alpha^{\alpha\rho} | WCN_\rho^{\alpha\rho})$, where the variation of rates due to $WCN_\rho^{\alpha\rho}$ is partialed out. This measure is the *unique* contribution of $WCN_\alpha^{\alpha\rho}$ and it represents how much R^2 would increase when going from the one variable model $K \sim WCN_\rho^{\alpha\rho}$ to the two-variable model $K \sim WCN_\rho^{\alpha\rho} + WCN_\alpha^{\alpha\rho}$. Averaging over the 209 proteins studied, we found $\rho^2(K_{R4S}, WCN_\alpha^{\alpha\rho} | WCN_\rho^{\alpha\rho}) = 0.0024 \pm 0.0005$. This value is statistically significant ($p < 10^{-3}$, bootstrapping), but *very small*: $WCN_\alpha^{\alpha\rho}$'s unique contribution to rate variation among sites is just 0.2%. As predicted by the stress model, the independent contribution of $WCN_\alpha^{\alpha\rho}$ is negligible.

$WCN_\rho^{\alpha\rho}$ vs. WCN_ρ^{ρ}

$WCN_\rho^{\alpha\rho}$, Eq. (9), is based on a two-nodes-per-site network representation of the protein. It considers the contacts between the node ρ that represents the side chain of a site with all

other nodes, ρ and α of the network. $\mathrm{WCN}_{\rho}^{\rho}$, Eq. (12), is an alternative alternative measure of side-chain packing based only on $\rho-\rho$ contacts. $\mathrm{WCN}_{\rho}^{\alpha\rho}$ is a better rate predictor than $\mathrm{WCN}_{\rho}^{\rho}$ for 122 of the 209 proteins. The expected correlations are $R^2(K_{\mathrm{R4S}},\mathrm{WCN}_{\rho}^{\alpha\rho}) = 0.392 \pm 0.008$ and $R^2(K_{\mathrm{R4S}},\mathrm{WCN}_{\rho}^{\rho}) = 0.389 \pm 0.008$. The average difference $\Delta R^2 = 0.0024 \pm 0.007$ is significant ($p < 10^{-3}$, bootstrapping), but very small (just 0.24% of explained variation). Thus, $\mathrm{WCN}_{\rho}^{\rho}$-based predictions are (almost) as good as $\mathrm{WCN}_{\rho}^{\alpha\rho}$ predictions. However, while $\mathrm{WCN}_{\rho}^{\rho}$ was posed *ad hoc*, $\mathrm{WCN}_{\rho}^{\alpha\rho}$ was *theoretically derived*.

$\mathrm{WCN}_{\rho}^{\alpha\rho}$ vs. WCN

Currently, WCN ($=\mathrm{WCN}_{\alpha}^{\alpha}$), the original weighted contact number (*Lin et al., 2008*), is one of the two main structural predictors of site-dependent evolutionary rates (*Yeh et al., 2014a*; *Yeh et al., 2014b*). It is worthwhile to consider whether the new measure presented here, $\mathrm{WCN}_{\rho}^{\alpha\rho}$ provides an improvement over WCN.

We found that $\mathrm{WCN}_{\rho}^{\alpha\rho}$ outperforms WCN for 206 out of the 209 proteins studied. The expected correlations are $R^2(K_{\mathrm{R4S}},\mathrm{WCN}_{\rho}^{\alpha\rho}) = 0.392 \pm 0.008$ and $R^2(K_{\mathrm{R4S}},\mathrm{WCN}) = 0.314 \pm 0.007$. The difference is $\Delta R^2 = 0.078 \pm 0.003$, which is statistically significant ($\mathrm{p} \ll 10^{-3}$, bootstrapping). Thus, not only does $\mathrm{WCN}_{\rho}^{\alpha\rho}$ outperform WCN for almost all proteins, but by a rather large amount: while WCN explains 31.4 % of the variation of evolutionary rates, $\mathrm{WCN}_{\rho}^{\alpha\rho}$ explains 39.2%, an increase by a factor of 1.25. Moreover, $\rho^2(K_{\mathrm{R4S}},\mathrm{WCN}|\mathrm{WCN}_{\rho}^{\alpha\rho}) = 0.0024 \pm 0.005$ ($\mathrm{p} < 10^{-3}$, bootstrapping). Despite statistical significance, the unique contribution of WCN is just 0.2%, which is negligible. Thus, $\mathrm{WCN}_{\rho}^{\alpha\rho}$ is a better predictor and the independent contribution of WCN is negligible.

$\mathrm{WCN}_{\rho}^{\alpha\rho}$ vs. RSA

The most studied structural predictor of site-dependent evolutionary rates is the relative solvent accessibility RSA (*Bustamante, Townsend & Hartl, 2000*; *Conant & Stadler, 2009*; *Franzosa & Xia, 2009*; *Ramsey et al., 2011*; *Shahmoradi et al., 2014*). Therefore, we compare the new measure $\mathrm{WCN}_{\rho}^{\alpha\rho}$ with RSA.

According to protein-by-protein results, $R^2(K_{\mathrm{R4S}},\mathrm{WCN}_{\rho}^{\alpha\rho}) > R^2(K_{\mathrm{R4S}},\mathrm{RSA})$ for 175 of the 209 proteins. The expected square correlations are $R^2(K_{\mathrm{R4S}},\mathrm{WCN}_{\rho}^{\alpha\rho}) = 0.392 \pm 0.008$ and $R^2(K_{\mathrm{R4S}},\mathrm{RSA}) = 0.327 \pm 0.007$. The difference is $\Delta R^2 = 0.065 \pm 0.004$, which is statistically significant ($p < 10^{-3}$, bootstrapping). Thus, $\mathrm{WCN}_{\rho}^{\alpha\rho}$ outperforms RSA as rate predictor for 84% of the proteins and $\mathrm{WCN}_{\alpha}^{\alpha}$ explains 6.5% more of the rate variation among sites, an improvement by a factor of 1.2 over the explaining power of RSA. Moreover, the expected value of the independent contribution of RSA is $\rho(K_{\mathrm{R4S}},\mathrm{RSA}|\mathrm{WCN}_{\rho}^{\alpha\rho}) = 0.005 \pm 0.001$). This is statistically significant ($p < 10^{-3}$, bootstrapping), but very small. Therefore, $\mathrm{WCN}_{\rho}^{\alpha\rho}$ is a better predictor and the independent contribution of RSA is minor.

CONCLUSION

We used the the mechanistic stress model to predict theoretically that site-specific rates of evolution depend solely on the side-chain contact density $\mathrm{WCN}_{\rho}^{\alpha\rho}$. According to the stress

theory, $WCN_{\rho}^{\alpha\rho}$ is proportional to the mutational 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. We tested this prediction on a large dataset of monomeric enzymes. We found that $WCN_{\rho}^{\alpha\rho}$ outperforms WCN_{ρ}^{ρ}, WCN and RSA and that the independent contributions of the latter are negligible, which supports the theoretical prediction.

To finish, we note that the structural properties studied do not explain all of the variation of substitution rates among sites. The best predictor, $WCN_{\rho}^{\alpha\rho}$ explains on average $\sim$ 39% of the variation, leaving 61% unexplained. Further research is needed to gain a full understanding of the variation of substitution rates among protein sites.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work has been supported by CONICET (National Scientific and Technical Research Council) and UNSAM (National University of General San Martín). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
CONICET (National Scientific and Technical Research Council).
UNSAM (National University of General San Martín).

Competing Interests

Julian Echave is a researcher of CONICET (National Scientific and Technical Research Council).

Author Contributions

- María Laura Marcos performed the calculations, analyzed the data, contributed analysis tools, wrote the paper and reviewed drafts of the paper.
- Julian Echave conceived and designed the study, performed the calculations, analyzed the data, contributed analysis tools, wrote the paper, prepared figures and reviewed drafts of the paper.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.911#supplemental-information.

REFERENCES

Bustamante CD, Townsend JP, Hartl DL. 2000. Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. *Molecular Biology and Evolution* **17(2)**:301–308 DOI 10.1093/oxfordjournals.molbev.a026310.

Conant GC, Stadler PF. 2009. Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Molecular Biology and Evolution* **26(5)**:1155–1161 DOI 10.1093/molbev/msp031.

Echave J. 2008. Evolutionary divergence of protein structure: the linearly forced elastic network model. *Chemical Physics Letters* **457(4–6)**:413–416 DOI 10.1016/j.cplett.2008.04.042.

Echave J, Fernández FM. 2010. A perturbative view of protein structural variation. *Proteins* **78(1)**:173–180 DOI 10.1002/prot.22553.

Echave J, Jackson EL, Wilke CO. 2015. Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. *Physical Biology* **12**:025002 DOI 10.1088/1478-3975/12/2/025002.

Franzosa EA, Xia Y. 2009. Structural determinants of protein evolution are context-sensitive at the residue level. *Molecular Biology and Evolution* **26**:2387–2395 DOI 10.1093/molbev/msp146.

Huang T-T, Marcos ML, Hwang J-K, 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 Evolutionary Biology* **14**:78 DOI 10.1186/1471-2148-14-78.

Katoh K, Kuma K-I, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research* **33**:511–518 DOI 10.1093/nar/gki198.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**:772–780 DOI 10.1093/molbev/mst010.

Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, Bornberg-Bauer E, Colwell LJ, de Koning APJ, Dokholyan NV, Echave J, Elofsson A, Gerloff DL, Goldstein RA, Grahnen JA, Holder MT, Lakner C, Lartillot N, Lovell SC, Naylor G, Perica T, Pollock DD, Pupko T, Regan L, Roger A, Rubinstein N, Shakhnovich E, Sjölander K, Sunyaev S, Teufel AI, Thorne JL, Thornton JW, Weinreich DM, Whelan S. 2012. The interface of protein structure, protein biophysics, and molecular evolution. *Protein Science* **21(6)**:769–785 DOI 10.1002/pro.2071.

Lin CP, Huang SW, Lai YL, Yen SC, Shih CH, Lu CH, Huang CC, Hwang JK. 2008. Deriving protein dynamical properties from weighted protein contact number. *Proteins* **72**:929–935 DOI 10.1002/prot.21983.

Mayrose I, Graur D, Ben-Tal N, Pupko T. 2004. Comparison of site-specific rate-inference methods: Bayesian methods are superior. *Molecular Biology and Evolution* **21**:1781–1791 DOI 10.1093/molbev/msh194.

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 DOI 10.1534/genetics.111.128025.

R Core Team. 2014. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

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. *Journal of Molecular Evolution* **79(3–4)**:130–142 DOI 10.1007/s00239-014-9644-x.

Shih C-H, Hwang J-k. 2012. Evolutionary information hidden in a single protein structure. *Proteins* **80(6)**:1647–1657 DOI 10.1002/prot.24058.

Sikosek T, Chan HS. 2014. Biophysics of protein evolution and evolutionary protein biophysics. *Journal of the Royal Society, Interface / the Royal Society* **11(100)**:20140419–20140419 DOI 10.1098/rsif.2014.0419.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313 DOI 10.1093/bioinformatics/btu033.

Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. 2013. Maximum allowed solvent accessibilites of residues in proteins. *PLoS ONE* **8**(**11**):e80635 DOI 10.1371/journal.pone.0080635.

Yang L, Song G, Jernigan RL. 2009. Protein elastic network models and the ranges of cooperativity. *Proceedings of the National Academy of Sciences of the United States of America* **106**:12347–12352 DOI 10.1073/pnas.0902159106.

Yeh SW, Huang TT, Liu JW, Yu SH, Shih CH, Hwang JK, Echave J. 2014a. Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. *BioMed Research International* **2014**:572409 DOI 10.1155/2014/572409.

Yeh SW, Liu JW, Yu SH, Shih CH, Hwang JK, Echave J. 2014b. Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure. *Molecular Biology and Evolution* **31**:135–139 DOI 10.1093/molbev/mst178.