Site-Specific Structural Constraints on Protein Sequence Evolutionary Divergence: Local Packing Density versus Solvent Exposure

So-Wei Yeh,[1,2] Jen-Wei Liu,[1,2] Sung-Huan Yu,[1,2] Chien-Hua Shih,[1,2] Jenn-Kang Hwang,*[,1,2] and Julian Echave*[,3]

[1]Institute of Bioinformatics and Systems Biology, National Chiao Tung University, HsinChu, Taiwan, ROC

[2]Center for Bioinformatics Research, National Chiao Tung University, HsinChu, Taiwan, ROC

[3]Escuela de Ciencia y Tecnología, Universidad Nacional de San Martín, Martín de Irigoyen 3100, San Martín, Buenos Aires, Argentina

***Corresponding author:** E-mail: jkhwang@faculty.nctu.edu.tw; jechave@unsam.edu.ar.

Associate editor: Andrew Roger

Abstract

Protein sequences evolve under selection pressures imposed by functional and biophysical requirements, resulting in site-dependent rates of amino acid substitution. Relative solvent accessibility (RSA) and local packing density (LPD) have emerged as the best candidates to quantify structural constraint. Recent research assumes that RSA is the main determinant of sequence divergence. However, it is not yet clear which is the best predictor of substitution rates. To address this issue, we compared RSA and LPD with site-specific rates of evolution for a diverse data set of enzymes. In contrast with recent studies, we found that LPD measures correlate better than RSA with evolutionary rate. Moreover, the independent contribution of RSA is minor. Taking into account that LPD is related to backbone flexibility, we put forward the possibility that the rate of evolution of a site is determined by the ease with which the backbone deforms to accommodate mutations.

Key words: **protein evolution, site-specific evolutionary rate, protein structure, local packing density, contact number, weighted contact number, relative solvent accessibility.**

Protein evolutionary divergence is subject to functional and biophysical constraints (Pal et al. 2006; Thorne 2007; Worth et al. 2009; Wilke and Drummond 2010; Grahnen et al. 2011; Liberles et al. 2012). Such constraints result in emergent patterns of sequence variability: different sites evolve at different rates. In most current research, relative solvent accessibility (RSA) is considered to be the main determinant of site-specific evolutionary rates (Bustamante et al. 2000; Dean et al. 2002; Franzosa and Xia 2009; Ramsey et al. 2011). However, local packing density (LPD), measured using either the contact number (CN) (Liao et al. 2005; Franzosa and Xia 2009) or the weighted contact number (WCN) (Shih et al. 2012), has also been found to correlate significantly with sequence variability. Of the cited studies, only one compared CN and RSA, finding similar correlations with evolutionary rates, with RSA performing better, and CN having a significant but minor independent contribution (Franzosa and Xia 2009). Untangling the relative contributions of LPD and RSA is important because, being conceptually different, they lead to different pictures of how structure constrains evolutionary sequence divergence. Despite recent advances, further research is needed to settle this issue. Here, we compare the ability of RSA and LPD, measured by CN or WCN, to account for site-specific rates of evolution.

We compiled a data set of 216 monomeric enzymes randomly picked from the Catalytic Site Atlas 2.2.11 (Porter et al. 2004). Protein size varies widely within the data set, and it includes representatives of all six main EC functional classes (Webb 1992) and domains of all main SCOP structural classes (Murzin et al. 1995). A list of the proteins and their properties is included in supplementary table S1, Supplementary Material online.

For each protein of the data set, we calculated three structural profiles: WCN, CN, and RSA. The WCN profile $WCN = (WCN_1, WCN_2 \ldots WCN_N)$ is a local packing density profile with $WCN_i = \sum_{j \neq i}^{N} 1/r_{ij}^2$, where r_{ij} is the distance between the C_α of residues i and j, and N is the protein length (Lin et al. 2008). The CN profile $CN = (CN_1, CN_2 \ldots CN_N)$ is a local packing density profile with CN_i defined as the number C_α within a spherical neighborhood of radius r_0. Here, we used $r_0 = 13$ Å as in Franzosa and Xia (2009). The RSA profile is $RSA = (RSA_1, RSA_2 \ldots RSA_N)$, where the RSA of each residue was obtained by dividing its area accessible to the solvent (ASA), calculated using DSSP (Kabsch and Sander 1983), by the maximum ASA for the given amino acid type (Miller et al. 1987).

To quantify evolutionary constraints at sequence level, we calculated the site-dependent sequence variability profile $CS = (CS_1, CS_2 \ldots CS_N)$, where CS_i is the rate of evolution of site i relative to the mean and N is the number of sites. These rates were calculated as follows. First, we obtained a set of up to 300 homologous sequences from the Clean_Uniprot database following the ConSurf protocol (Goldenberg et al. 2009; Ashkenazy et al. 2010). Second, we obtained the

Letter

multiple sequence alignment (MSA) using MUSCLE (Edgar 2004). Finally, given the MSA, we calculated the site-specific rates using Rate4Site, which builds the phylogenetic tree using the neighbor-joining algorithm and estimates the rates using an empirical Bayesian method and the JTT model of sequence evolution (Pupko et al. 2002; Mayrose et al. 2004).

We compared the sequence and structural profiles for each protein of the data set. For the sake of such comparison, profiles were turned into *z*-scores calculating their difference from the mean and dividing by their standard deviation (SD). For LPD measures, the sign of the *z*-score was changed so that higher scores correspond to higher expected evolutionary rates and expected correlations are positive. To reduce noise, before normalization profiles were smoothed using a sliding window of size 3 as recommended in Pei and Grishin (2001). We quantified the similarity between profiles using Pearson's correlation coefficient ρ, which measures the degree of linear association.

In figure 1, we show the profiles for the threonine synthase of *Saccharomyces cerevisiae* (Baker's yeast), pdb code 1kl7, a protein that we will use throughout this letter as an illustrative example. The qualitative similarity between all structural profiles and the sequence profile for this protein is clear. Quantitatively, for this case, we find $\rho(CS,WCN) = 0.77$, $\rho(CS,CN) = 0.71$, $\rho(CS,RSA) = 0.62$, which are all significant ($P << 10^{-2}$). For this protein, then, LPD profiles (WCN and CN) are more similar to CS than the RSA profile. To see whether this is the case in general, we calculated these three correlations for each protein of the data set. For 213/216 proteins, the three correlations are significant ($P < 0.01$), whereas for 3/216 none of the three correlations are significantly different from 0. We removed these three proteins for the rest of the analysis. Results are shown in figure 2 where we compare LPD–CS correlations (y axis) with RSA–CS correlations (x axis) for all proteins of the data set. Counting the number of cases above and below the $y = x$ diagonal, we find that $\rho(CS,WCN) > \rho(CS,RSA)$ for 173/213 = 81% of cases and $\rho(CS,CN) > \rho(CS,RSA)$ for 154/213 = 72% of cases. A binomial test shows that these values are significantly larger than 50% ($P << 10^{-2}$). Moreover, the mean sequence–structure correlations (±1 SD) are $\langle\rho(CS,WCN)\rangle = 0.62 \pm 0.01$, $\langle\rho(CS,CN)\rangle = 0.59 \pm 0.01$, and $\langle\rho(CS,RSA)\rangle = 0.56 \pm 0.01$. Therefore, both the number of cases and the mean values support that both LPD measures, especially WCN, correlate better than RSA with evolutionary rates.

For the example protein of figure 1, LPD and RSA profiles are clearly similar to each other, which is also the case for all proteins of the data set. This raises the issue of whether LPD and RSA provide overlapping or complementary contributions to evolutionary rate variation among sites. To address this question, we used semipartial correlations to partition the overall variance into overlapping and unique contributions. Given a dependent variable y and two independent variables u and v, the semipartial correlation $\rho(y,u\,|\,v) = \rho(y,u) - \rho(u,v)\rho(y,v) \Big/ \sqrt{1-\rho^2(u,v)}$ is the correlation between y and u from which v has been partialed out. If one performs a linear fit $y \sim u + v$, the squared total correlation coefficient $0 \leq R^2 \leq 1$ represents the proportion of the variance of y accounted for by the linear model: R^2 is the *explained variance* of y. Using semipartial correlations,

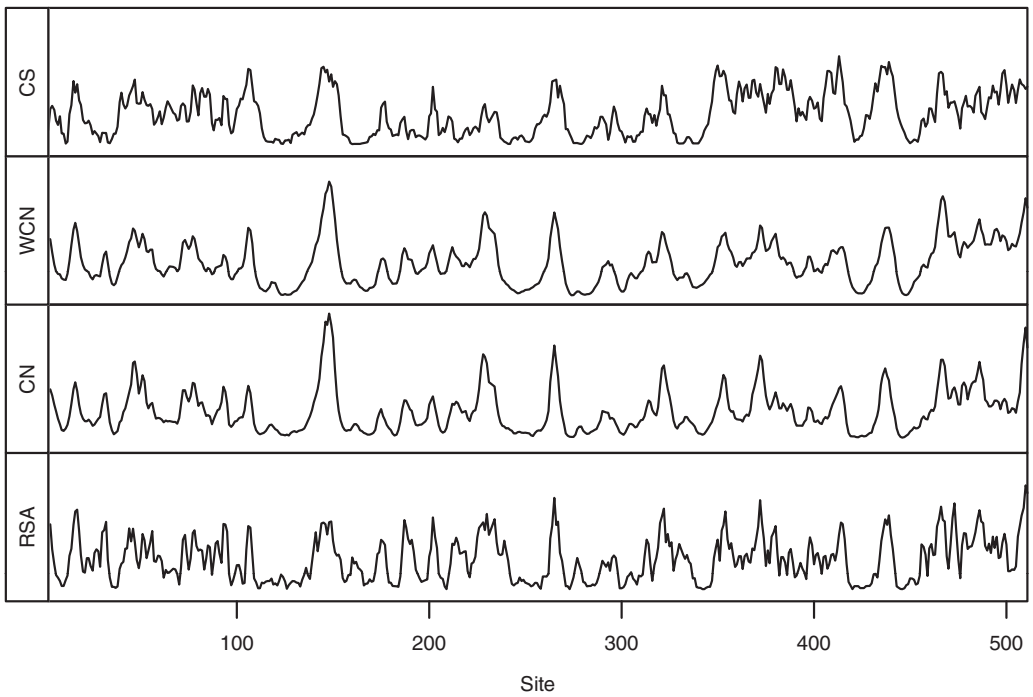

Fig. 1. Comparison of the site-dependent sequence-variability profile (CS) and structural profiles (WCN, CN, RSA) for the threonine synthase of *Saccharomyces cerevisiae*, pdb code 1kl7. Profiles have been normalized by turning them into z-scores. For WCN and CN, the sign of z-scores was reversed so that the expected correlation with rates of evolution is positive.

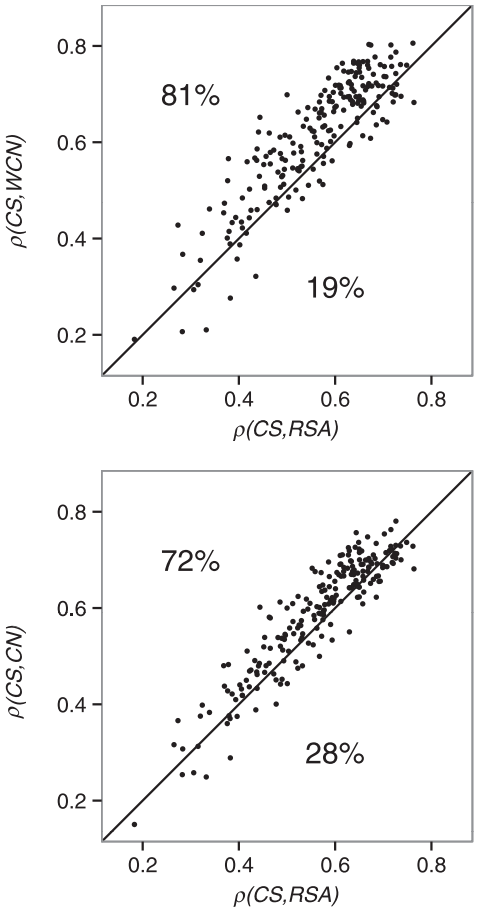

Fig. 2. Comparison between the correlation coefficients of site-specific rates of evolution and different structural measures. Top: $\rho(CS,WCN)$ versus $\rho(CS,RSA)$. Bottom: $\rho(CS,CN)$ versus $\rho(CS,RSA)$. The sign of WCN and CN profiles is changed before calculating the correlations so that significant correlations are expected to be positive. Points above (below) the diagonal are proteins for which LPD (RSA) correlates better than RSA (LPD) with sequence variability. The percentages of points above and below the diagonals are shown.

such explained variance can be partitioned as the sum of three contributions $R^2 = \rho^2(y,u \text{ or } v) + \rho^2(y,u \mid v) + \rho^2(y,v \mid u)$. The first term accounts for the redundancy between the independent variables: it represents the information that the two variables have in *common*. The last two terms represent the *unique* contributions of u and v, respectively, to the explained variance of y. For a more detailed explanation variance partitioning, see Cohen (2003) and Warner (2013).

For the example protein 1kl7 (fig. 1), a variance partitioning analysis of CS~WCN + RSA shields $R^2 = 0.59 = 0.38 + 0.21 + 0.00$ for the common, unique WCN, and unique RSA contributions, respectively. Thus, the variance of CS explained by WCN and RSA is 59% of its total variance. Dividing the previous equation by R^2 and multiplying by 100, we can partition the explained variance as $100\% = 65\% + 35\% + 0\%$, meaning that 65% of the explained variance is accounted commonly by WCN or RSA and 35% uniquely by WCN, but RSA shows no unique contribution (0%). A similar analysis of CS~CN + RSA shows that the explained variance is 51% of the total CS variance ($R^2 = 0.51$) and such explained variance (100%) is partitioned into common, unique CN, and unique RSA contributions of 75%, 25%, and 0%, respectively. Therefore, for this particular case, RSA does not contribute to the variance of CS once either LPD measure, WCN or CN, are controlled for. To see whether this is the case in general, we repeated this analysis for each protein of the data set. The variance partition analysis is summarized in table 1. As a result of the large LPD-RSA correlations, the redundancy term is the largest. For the CS~WCN + RSA case, WCN accounts uniquely for 19% of the explained variance while RSA's unique contribution is 5%. For the CS~CN + RSA case, the unique contribution of CN is 13% and that of RSA is 5%. Therefore, both LPD measures have larger unique contributions to the explained variance than RSA.

Table 1. Variance Partitioning.

Fit	Contribution	R^2	%
CS~WCN + RSA	Total	0.413 ± 0.010	100.0
	Common	0.318 ± 0.008	76.3 ± 0.9
	WCN	0.080 ± 0.004	19.1 ± 0.9
	RSA	0.015 ± 0.001	4.6 ± 0.6
CS~CN + RSA	Total	0.380 ± 0.009	100.0
	Common	0.316 ± 0.008	82.1 ± 0.6
	CN	0.048 ± 0.003	12.7 ± 0.7
	RSA	0.017 ± 0.001	5.2 ± 0.5

Note.—Fit is the bivariate linear fit considered; R^2 is the explained variance averaged over the data set of 213 enzymes ± its SD; % is the proportion of explained variance accounted for by the given contribution.

To complete this study, we compare univariate and bivariate models. For this purpose, we note that the squared semipartial correlation $\rho^2(y,v \mid u)$ represents the increase in squared correlation R^2 resulting from adding variable v to the linear model $y \sim u$ to obtain the model $y \sim u + v$ (Warner 2013). The semipartial correlation's P value can be used to decide whether to include v or not. The univariate model should be discarded in favor of the bivariate model only if such P value is below a given cut-off significance level, which here we choose to be 0.01.

For example, for 1kl7, $\rho(CS,RSA \mid WCN) = -0.03$ and $\rho(CS,RSA \mid CN) = -0.01$ with P values of 0.46 and 0.91, respectively. As both P values are above 0.01, for this case adding RSA will not significantly improve either of the univariate models CS~WCN or CS~CN. Repeating this analysis for all proteins of the data set, for the CS~WCN + RSA case we find that the best model is CS~WCN for 146/213 proteins, the bivariate model for 33/213 cases, and CS~RSA for 34/213 cases. For the CS~CN + RSA case, the best model is CS~CN in 131/213 cases, the bivariate model in 33/213 cases, and CS~RSA in 49/213 cases. Therefore, not only the average unique contributions of both LPD measures are larger than that of RSA, but if one were to choose one variable to model structural constraints, LPD measures are a safer bet with an odds ratio of approximately 3:1.

To assess the robustness of our conclusions, we performed some extra tests. First, because Pearson's correlation coefficient may be sensitive to the existence of outliers or nonlinear

dependency between variables, we repeated the whole analysis using Spearman's rank-based correlations, finding very similar results (not shown). Second, we found the same trends when using other methods to quantify sequence variability. Third, we verified that the conclusions are the same regardless of structural SCOP class or EC functional class. Finally, we verified that there are no effects of protein size and resolution of the X-ray structure.

Summing up, we have found that both LPD measures, WCN and CN, correlate significantly better than RSA with site-specific rates of evolution. Moreover, when LPD measures are adjusted for, RSA is often not an important predictor of CS. This is rather surprising considering that most current work assumes that RSA is the main determinant of structural constraints at site level. More importantly, such assumption is supported by a recent study that compared RSA and CN and found that RSA correlated better than CN, whose unique contribution was significant but minor (Franzosa and Xia 2009). Their data set and methodology is too different from ours for us to perform a detailed comparison. However, we note that the correlations found in that study are much smaller than ours: 0.126 and −0.118 for RSA and CN, respectively, versus our correlations of approximately 0.6, on average. Our results are less noisy, probably due to usage of a much larger number of homologous sequences to estimate rates and not grouping sites of different proteins. Therefore, the present results are less likely to be affected by spurious correlations. We are confident that the present results provide strong support to the notion that local packing density is a better candidate than solvent exposure to quantify structural evolutionary constraints on sequence divergence.

To finish, we consider what mechanism could explain the connection between local packing density and sequence divergence. Two observations may be relevant in this regard. First, local packing density is related to backbone flexibility (Halle 2002; Lin et al. 2008). Second, backbone flexibility is directly connected with the structural change resulting from a mutation (Echave and Fernandez 2010). Thus, we speculate that local packing density of a site could be an approximate measure of its flexibility, which could be the actual causal determinant of its rate of evolution. Mutations at flexible sites (low packing density) would be accommodated more easily than mutations at rigid sites (high packing density). Such a possibility was hinted in Liao et al. (2005), and it is consistent with some recent studies of the correlation between mobility and conservation (Liu and Bahar 2012). However, further work is needed to test this hypothesis.

Supplementary Material

Supplementary table S1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This research was supported in part by Academic Summit Program of National Science Council with grant number 101-2745-B-009-001-ASP and the Center for Bioinformatics Research of Aiming for the Top University Program of the National Chiao Tung University and Ministry of Education, Taiwan, ROC. J.E. is a researcher of CONICET.

References

Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* 38: W529–W533.

Bustamante CD, Townsend JP, Hartl DL. 2000. Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. *Mol Biol Evol.* 17:301–308.

Cohen J. 2003. Applied multiple regression/correlation analysis for the behavioral sciences. Mahwah (NJ), London: L. Erlbaum Associates.

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

Echave J, Fernandez FM. 2010. A perturbative view of protein structural variation. *Proteins* 78:173–180.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

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

Goldenberg O, Erez E, Nimrod G, Ben-Tal N. 2009. The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Res.* 37:D323–D327.

Grahnen JA, Nandakumar P, Kubelka J, Liberles DA. 2011. Biophysical and structural considerations for protein sequence evolution. *BMC Evol Biol.* 11:361.

Halle B. 2002. Flexibility and packing in proteins. *Proc Natl Acad Sci U S A.* 99:1274–1279.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.

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.* 18:59–64.

Liberles DA, Teichmann SA, Bahar I, et al. (33 co-authors). 2012. The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci.* 21:769–785.

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.

Liu Y, Bahar I. 2012. Sequence evolution correlates with structural dynamics. *Mol Biol Evol.* 29:2253–2263.

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.* 21:1781–1791.

Miller S, Janin J, Lesk AM, Chothia C. 1987. Interior and surface of monomeric proteins. *J Mol Biol.* 196:641–656.

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.* 247:536–540.

Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet.* 7:337–348.

Pei J, Grishin NV. 2001. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* 17:700–712.

Porter CT, Bartlett GJ, Thornton JM. 2004. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.* 32:D129–D133.

Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* 18(Suppl 1):S71–S77.

Ramsey DC, Scherrer MP, Zhou T, Wilke CO. 2011. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics* 188:479–488.

Shih CH, Chang CM, Lin YS, Lo WC, Hwang JK. 2012. Evolutionary information hidden in a single protein structure. *Proteins* 80: 1647–1657.

Thorne JL. 2007. Protein evolution constraints and model-based techniques to study them. *Curr Opin Struct Biol.* 17: 337–341.

Warner RM. 2013. Applied statistics: from bivariate through multivariate techniques. Thousand Oaks (CA): SAGE Publications.

Webb EC. 1992. Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. San Diego: Academic Press.

Wilke CO, Drummond DA. 2010. Signatures of protein biophysics in coding sequence evolution. *Curr Opin Struct Biol.* 20:385–389.

Worth CL, Gong S, Blundell TL. 2009. Structural and functional constraints in the evolution of protein families. *Nat Rev Mol Cell Biol.* 10:709–720.