STRUCTURAL METRICS PREDICT SITE-SPECIFIC EVOLUTIONARY RATE IN MEMBRANE BOUND PROTEINS

Presented by Xuezhen Du

In partial fulfillment of the requirements for graduation with the

Dean’s Scholars Honors Degree in Biology

_____ _____

(Name) Date

Supervising Professor

_____ _____

(Name) Date

Honors Advisor in Biology

Abstract

Membrane proteins are involved in many critical biological processes and mutations are linked to various diseases. We examined how the properties of the location of an amino acid residue within the protein structure dictates the rate at which it evolves. We tested 3 structural metrics: WCNSC, WCNCA and RSA for their effectiveness at predicting evolutionary rates within membrane proteins. WCNSC performed better than WCNCA in almost all cases and better than RSA in the majority. However, for some classes of proteins, especially those where the pore is a major feature, the effectiveness of WCN greatly diminished while that of RSA decreased by a lesser degree.

Acknowledgements

I would like to thank all the people who made this project possible. Thank you to my professor, Dr. Wilke, for taking me into his lab. Thank you to my honors advisor, Dr. DeLozanne, for encouraging and guiding my interests. And finally, thank you to my mentors, Austin Meyer and Ben Jack, for showing me the ropes and supporting me during this long journey.

Introduction

Rates of evolution among the genes encoding proteins can vary widely. Within the protein, individual residues evolve at different rates as well. A residue evolves when one amino acid is replaced by another at a site. The location of a site in the protein structure greatly influences its evolutionary rate (Echave *et al.* 2016). Two common structural metrics used to predict evolutionary rates are packing density and relative solvent accessibility.

Relative solvent accessibility has been the most commonly used structural metric although recent research suggests that it may not have the best predictive power. The amount of exposure to solvent is linearly correlated with the mutation rate of a site (Ramsay *et al.* 2011). More solvent-exposed sites evolve faster than less exposed sites. Less exposed sites are buried in the protein and are more sterically hindered. Mutations in these sites tend to be more disruptive to the structure of the protein.

Packing density measures how tightly packed the site is within the protein (Echave *et al.* 2016). Mutations in tightly packed sites are expected to be more destabilizing to the proteins, and thus we expect sites with higher packing density to be more conserved. Packing density is correlated with weighted contact number. Contact number is the number of atoms around the site. Weighted contact number takes all of the residues in the proteins and weights them by taking sum of inverse square distances from the site considered to other sites. Amino acids are composed of a main chain forming the backbone of the polypeptide and a side chain which protrudes out from the backbone. Weighted contact number can be calculated using the alpha carbon of the main chain WCNCA or the side chain (WCNSC) at each site. WCNSC has been shown to be a better predictor than the backbone WCNCA.

Studies on enzymes have shown that WCN is a better predictor of mutation rates than RSA. We were interested to see if this trend also applied to membrane bound proteins. Membrane bound proteins make up 30% of the proteins in eukaryotes, and have a variety of functions such as molecular transport and cell signaling (Moraes *et al.* 2014). Mutations in membrane proteins are involved in diseases including cystic fibrosis, obesity, and cancer. Membrane bound proteins differ from unbound proteins in several respects. Most unbound proteins are roughly globular while membrane bound proteins have a range of shapes. Alpha-helical proteins are composed of several helical transmembrane domains stacked together, whereas beta-barrel proteins form by coiling of anti-parallel beta-sheets (von Heijne 1997). Also, whereas the surface of unbound proteins are exposed to the aqueous environment, membrane bound proteins have transmembrane regions which are exposed to a hydrophobic environment. It is possible that different environments have different constraints on evolution rates.

We have several goals in this study. We test whether the patterns we observed in enzymes hold true to other classes of proteins. We also seek to determine how the membrane and different structures affects the predictive power of RSA and WCN.

Methodology

Sequence collection and pre-processing

We used a dataset of 654 membrane-bound protein structures collected from the Membrane proteins of known 3D structure database (http://blanco.biomol.uci.edu/mpstruc/exp/list, Stephen White Lab at UC Irvine). We made an effort to select non-enzymatic proteins in order to better contrast with results from globular enzymes. However, this selection was cursory as it was difficult to determine whether a protein

had enzymatic properties and it is likely that a portion of the dataset is enzymatic. The proteins were primarily beta-barrel and alpha-helical transmembrane proteins.

We recoded whether each protein was monomeric and multimeric. To calculate rates, it was necessary to select a single chain. In this case, the chain chosen was the first labeled chain in the pdb structure. The amino acid sequence of the chain was extracted from the protein structure.

Calculation of evolutionary rates

We used PSI-BLAST (Altschul *et al.* 1997) to collect homologous sequences. PSI-BLAST was run using the Uniref 90 database for two iterations with a e-value of 1e-10, percent identity threshold of 30% and length error of 20%. We used Mafft (Katoh *et al.*2013) to align these sequences. A sample of 300 sequences from each alignment to calculate evolutionary rates. A phylogenic tree was constructed from each alignment using RAXML (Stamatakis 2014). Using these trees, we calculated raw and normalized rates of evolution at each site using Rate4Site (Pupko *et al.*2002).

Structural metrics

We calculated the weighted contact number (WCN) and residue solvent accessibility (RSA) for both the monomer and the multimer. RSA is based on exposure to solvent at a site. WCN is calculated by taking the sum of inverse square distances from the residue to all other residues in the protein. We used both the main chain and side-chain as reference points on the protein. Correlations were calculated between side-chain WCN (WCNSC), C-atom WCN (WCNCA) and RSA and evolutionary rate at each site for each protein. WCN and rate are inversely correlated whereas RSA and rate are directed correlated. To account for this discrepancy, we used 1/WCN instead of WCN in our linear model so that we could direct

compare rate correlations between RSA and WCN. When we say WCN in this context, we are referring to 1/WCN. At this point, 1 protein, 1M0K, was determined to have incomplete data and was removed from the sample.

Construction of linear models

Using the data for RSA, WCNSC, WCNCA an evolutionary rate at each site, we built linear models in R of each of the first three variable with rate for each protein. We compared the coefficient of determination (R^2) of WCNSC, WCNCA, and RSA with evolutionary rate. Because rate can take negative values and we were interested in the magnitude of effects of structural metrics on a site, we use R^2 to describe the correlation rather than R. A negative or positive rate are equally indicative of the sites permissibility to evolve.

Results

Structural Metrics Predict Evolutionary Rate in Membrane Bound Proteins

RSA and WCN are two commonly used metrics used to predict evolutionary rate in proteins. However, both of these metrics have flaws when used on membrane bound proteins. We calculate RSA based on the protein structure, which assumes that residues on the surface are exposed to aqueous solvent. The membrane adds a wrinkle to this assumption, as residues may be exposed to either the solvent, a hydrophilic environment or the membrane, a hydrophobic one. These different environments likely have different influences on evolutionary. Many membrane proteins are transporters which complicates the use of WCN. These proteins tend to have exposed spaces in the interior. A residue on inside surface of a channel tends to be hydrophilic not hydrophobic, and is much less sterically hindered than a residue buried in the core of an enzyme. WCN which considers location relative to other residues in the protein, may not fully

account for these differences. We tested how useful RSA and WCN were in predicting rates in membrane proteins, in light of these potential flaws.

We plotted the coefficient of determination (R^2) of WCNSC and rate against that of RSA and rate. Our results supported the use of WCNSC over RSA. Values of R^2 ranged between 0 and 0.7 for both metrics. For 359 out of 572 proteins analyzed, WCNSC was a better predictor of evolutionary rate than RSA. The coefficient of determination of WCNSC and rate was significantly greater than that of RSA and rate (paired t-test, $p<2.2e-16$) by a mean difference of 0.0322. It appears that, in general, WCNSC is a superior metric. However, for over a third of proteins RSA had greater predictive power. Additionally, neither WCNSC nor RSA had a R^2 greater than 0.6,

Side-chains have Greater Predictive Power than Alpha carbons when used in calculation of WCN

Echave *et al* (2016) argued that WCNSC consistently outperformed WCNCA as a rate determinant. We examined if this relationship also existed in membrane proteins. Alpha carbons have traditionally been used as the reference point in calculations of WCN, but consistent outperformance would recommend side chains be used instead. Values were close to equal for both metrics, with WCNSC generally being higher (WCNSC was a better predictor of evolutionary rate than WCNCA for 470 out of 572 proteins analyzed)(Fig. 2). We found that the correlation of WCNSC and rate was greater than that of RSA and rate (paired t-test, $p<2.2e-16$) by a mean difference of 0.0305. Since WCNSC explained more of the variance in rate than RSA, we were interested to see how WCNCA compared to RSA. However, we did not find a significant difference in correlation with evolutionary rate between RSA and WCNCA (paired t-

test, p=0.6578) (Fig. 3). It appears that the loss in predictive power from using the alpha carbon rather than the side chain in WCN calculations nullified the advantage of using WCN over RSA.

Beta-Barrel Proteins show Discrepancies from the General Trend

We categorized the proteins then compared the R^2 values for WCNSC and RSA with rate to determine if one metric outperformed the other for specific classes of proteins. First, we divided the set based on secondary structure into 410 alpha-helical and 130 beta-barrel proteins. Alpha helical proteins, the most common type of membrane proteins, are comprised of many transmembrane helices, held together by van der Waals forces and hydrogen bonding (Xiong 2006). Beta-barrel proteins, which are mainly found in bacteria, are weaved from anti-parallel beta-sheets into a cylindrical structure. Figures 4-7 show the distribution of R^2 for WCNSC and RSA for alpha-helical and beta-barrel proteins.

For alpha-helical proteins, the distribution resembled the distribution for all membrane proteins. R^2 values ranged from 0 to 0.7 for WCN with a mean of 0.309 and standard deviation 0.16 and 0 to 0.6 for RSA with a mean of 0.263 and standard deviation 0.133. The distributions appear to be approximately normally distributed except for a large amount of proteins with values close to zero for both metrics. The number of these proteins with extremely low correlations is higher for RSA.

For beta-barrel proteins, values for R^2 fell between 0-0.5 for WCN with a mean of 0.169 and standard deviation 0.125 and 0-0.4 for RSA with a mean of 0.172 and standard deviation 0.097. The distribution for RSA appears to be normally distributed, however a large amount of proteins showed R^2 values close to zero with WCNSC.

Figure 8 compares R^2 values for RSA and WCNSC. For 287 of 410 proteins, WCNSC explained more of the variance in evolutionary rate than did RSA. WCNSC had significantly higher R^2 values than RSA, as predicted by previous studies (Paired t-test, $p<2.2e-16$). However, for beta-barrel proteins, RSA unexpectedly outperformed WCNSC for over half (73/130) of proteins, in reverse of expectations. The difference between R^2 values was not significant (paired t-test, $p=0.69$). We also noted a cluster of proteins where WCNSC had very low R^2 values and was greatly outperformed by RSA.

Comparison of WCNSC and RSA among Subsets of Beta-Barrel Proteins

We further divided the beta-barrel proteins to determine whether the performance of RSA over WCNSC was true for all beta-barrel proteins or influenced by a specific subset of outliers. We also wanted to examine if the previously observed cluster represented a specific group. We identified two major subsets of beta-barrel proteins: porins (n=35), monomeric-dimeric (n=69), outer membrane autotransporters (n=12), outer membrane carboxylate channels (n=12), Omp85-TpsB outer membrane transporter superfamily proteins (n=2). Figure 9 compares the distributions of these subgroups. Because of their small counts, the last three groups were put together in an "other" category for readability.

The distribution of monomeric-dimeric proteins (Fig. 10) was fairly similar to the general distribution for membrane proteins. WCNSC explained more of the variance in evolutionary rates than RSA for 43 of 69 proteins. R^2 values were significantly greater for WCNSC than RSA (one tailed t-test, $p=0.008$). R^2 values ranged between 0 to 0.5 for both RSA and WCN and appeared to be fairly evenly distributed along this range.

The distribution of porins (Fig.11) showed most fell within a cluster for which R^2 values for both RSA and WCNSC were low, around 0.1 for RSA and 0.05 for WCNSC. Outside of this cluster there were cases where R^2 values were higher for both metrics. For the entire subset, RSA explained significantly more variance than WCNSC (one tailed t-test, p=0.008). RSA explained more variance than WCNSC for 8 out of 35 proteins. It is unclear if most of these proteins are clustered due to an intrinsic property of porins, or because of a bias due the small size of our subset.

For the outer membrane autotransporter proteins (Fig.12), WCNSC and RSA appeared to be fairly equal. A paired t-test did not find a difference between R^2 values between WCNSC and RSA (p=0.8646), possibly because of the small sample size. RSA explained more of the variation for 7 of 12 proteins. The average difference was 0.003.

Like the porins, variance in outer membrane carboxylate channel proteins (Fig.13) was significantly better explained by RSA than WCNSC (one tailed t-test, p=7.8E-5). RSA outperformed WCNSC for 11 out of 12 proteins, usually by a large margin. The mean difference between RSA and WCNSC was 0.63. Compared to the cluster seen in porins where RSA also greatly outperformed WCNSC, R^2 values for both metrics were greater, falling between 0 -0.2 for WCNSC and 0.1 -0.3 for RSA although the overall range was lower than for the porins. It is possible that with a larger sample size we would have seen a comparable or greater range of R^2 values.

Discussion

Although alpha carbons have traditionally been used in calculation of WCN, WCNSC explained slightly more variance in evolutionary rates than WCNCA in most cases. This

relationship held true for all subsets of membrane proteins. Discrepancies were usually small. This was expected as both metrics are measuring the same behavior using the same method, only with different reference points. From these findings, we conclude that WCNSC is a strictly better predictor of evolutionary rates than WCNCA and WCNCA has no discernible advantage to recommend its use over WCNSC. It appears that side chains better capture the structural changes which influence protein evolution that do the alpha carbons. Side chains likely have greater predictive power because they project out of the backbone. Changes in side chain conformation are more disruptive to the protein structure and therefore a larger effect on fitness.

According to Echave *et al*. (2016), WCNSC performed better than RSA in most proteins and RSA made little independent contribution to rate. In general, this was true also when looking at membrane proteins. However, this was not absolute as RSA outperformed WCNSC in a third of proteins. WCNSC performed better relatively for alpha-helical proteins than beta-barrel proteins. WCNSC faired especially particularly poorly as a predictor for evolutionary rates in the porin family, where correlations with RSA were significantly higher. While WCNSC appears to do fairly well as a predictor for most proteins, it's being outperformed in some classes indicates that RSA does make an independent contribution and better captures some attribute of these proteins. Alpha helical proteins had higher correlations for both WCNSC and RSA. The sample size for alpha-helical proteins was three times greater than for beta-barrel proteins (410 vs 130) so we would expect to see a larger spread. However, if this increase was to the effect of more outliers alone, we would expect to see only a few individuals at the maxima and similar centers which is not the case. We therefore conclude that both structural metrics tested are better suited to describe the properties of alpha-helical proteins than beta-barrel proteins. A key difference in the general structure of alpha-helical and beta-barrel proteins is the size of the pore. Both RSA

and WCN were first used to study enzymes. These proteins tend to be globular without exposed interior. We hypothesize that deviation from the globular shape, especially cavities into the interior of the protein decrease the reliability of the structural metrics tested. Another factor which could play a role is size of the protein. Beta-barrel are larger, on average, than alpha-helical proteins. As size increases the predictive power seems to decrease. Porins which are among the largest proteins had very low correlations between RSA and especially WCNSC with rate. However, size of the protein is correlated with size of the pore. Intuitively, it makes more sense for the pore to hinder the predictive power of these metrics.

This could result from several factors. Normally, sites with high WCN values tend to be hydrophobic and buried within the protein. However, the presence of a pore disrupts several of these assumptions. The residues lining the pore are exposed to an aqueous environment and are usually hydrophilic. Additionally, amino acids across the pore are close enough to contribute to the WCN but may not be close enough to hinder mutations at the site. As the size of the pore increases, the residues lining the pore makes up a greater proportion of the total residues, so the predictive power of WCN decreases. Both RSA and WCN have limitations when applied to membrane proteins. Because we only use the protein structure in our calculations, our method of calculating these metrics is blind to the effects of the membrane. For WCN, the surrounding membrane is in contact with the protein and constrains the structure but is not included in the calculation. Additionally, WCN may be less applicable to sites on the surface of the pore. These sites would have relatively high WCN values due to being in the center of the protein but are exposed to solvent. This difference in external environment could lead to different rates than a site with similar WCN values that is embedded inside the protein and not exposed to solvent.

In regards to RSA, because we are ignoring the membrane, sites on surface of the protein are assumed to be exposed to aqueous solvent, when in reality they are in a hydrophobic environment. Because we are assuming an opposite environment, we would expect predictions at these sites to be highly inaccurate. By accounting for membrane, we can make more accurate models, however even these would not be completely accurate because lipid-amino acid interactions at the interface of the protein and the membrane have different thermodynamic effects then amino acid-amino acid interactions inside the protein. Although we expected misrepresenting exposure to solvent to lead to inaccuracies, RSA performed almost as well as WCN. This could suggest that hydrophobicity is not a major determinant of evolutionary rate or that RSA is actually correlated with some other property unrelated to the external environment.

Our metrics explained at best 70% of the variance evolutionary rates, and often much less. While they can be useful predictors, both RSA and WCN have flaws when applied to irregular structures. In the future, we would examine how the predictive power of these metrics vary among regions of the protein, for example if WCN is especially poor for residues along the pore. We would also compare the relationship between protein size or pore diameter and predictive power. We are also interested in analyzing and comparing the properties of specific cases where R^2 values were extremely high or low.

Figures

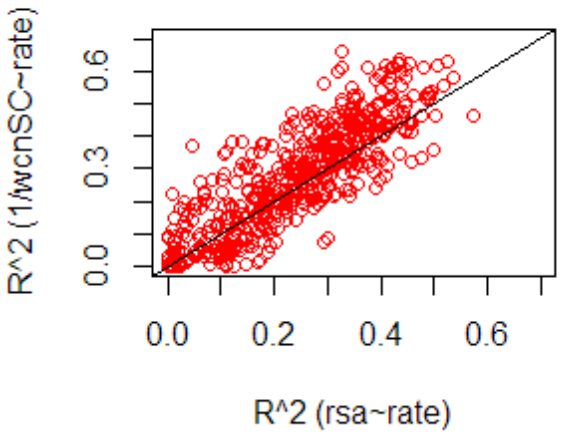

Figure 1. Comparison of R^2 values for RSA and 1/WCNSC with evolutionary rates for the entire sample of 540 membrane proteins. Each circle represents an individual protein. The line shows X=Y. WCNSC performed better than RSA for a majority of proteins

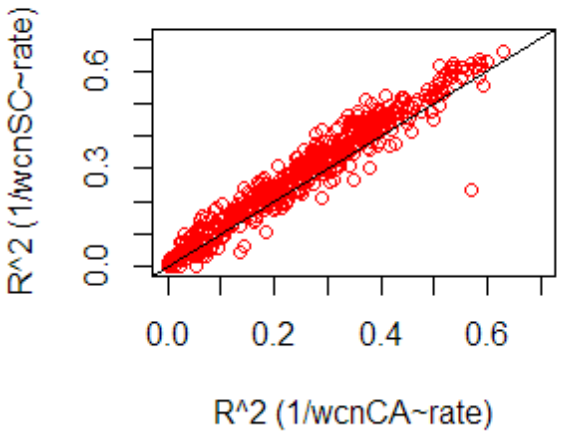

Figure 2. Comparison of R^2 values for 1/WCNCA and 1/WCNSC with evolutionary rates for the entire sample of of 540 membrane proteins. Each circle represents an individual protein. The line shows X=Y. Values are similar but mostly higher for WCNSC.

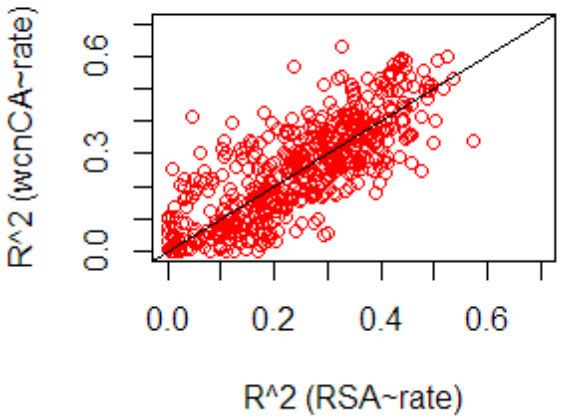

Figure 3. Comparison of R^2 values for 1/WCNCA and RSA with evolutionary rates for the entire sample of of 540 membrane proteins. Each circle represents an individual protein. The line shows X=Y. WCNCA and RSA performed about equally well as predictors for rate.

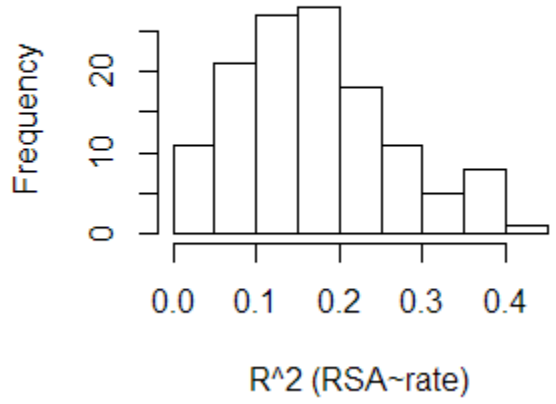

Figure 4. Histogram of R^2 values for RSA with rate in Beta-Barrel Proteins

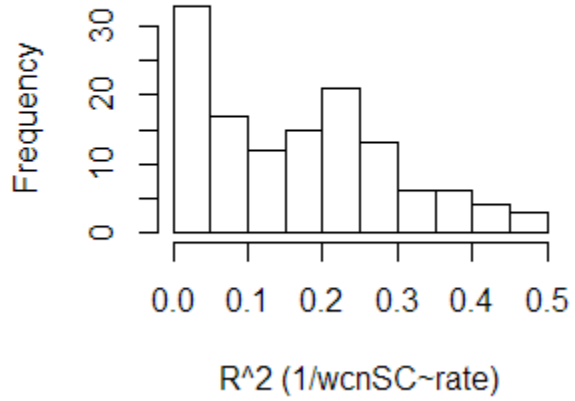

Figure 5. Histogram of R^2 values for 1/WCNSC with rate in Beta-Barrel Proteins. We see a large number of values between 0 and 0.05 then another peak between 0.2 and 0.25.

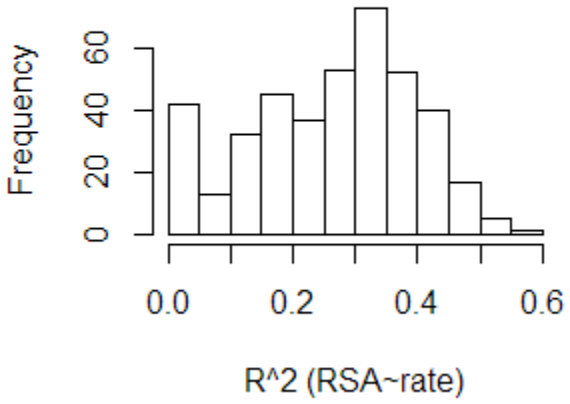

Figure 6. Histogram of R^2 values for RSA with rate in 410 Alpha-Helical Proteins. We see a mostly normal distribution with a peak around 0.3 and a secondary peak between 0 and 0.05.

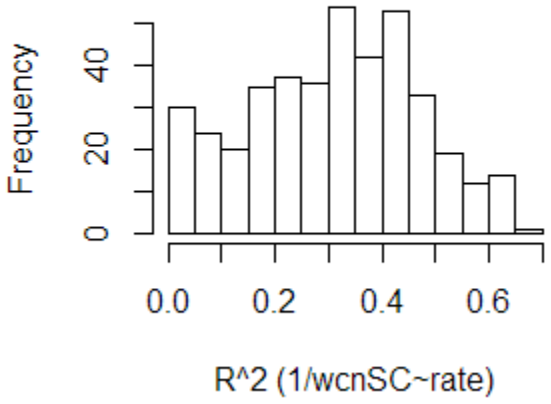

Figure 7. Histogram of R^2 values for 1/WCNSC with rate in Alpha-Helical Proteins.

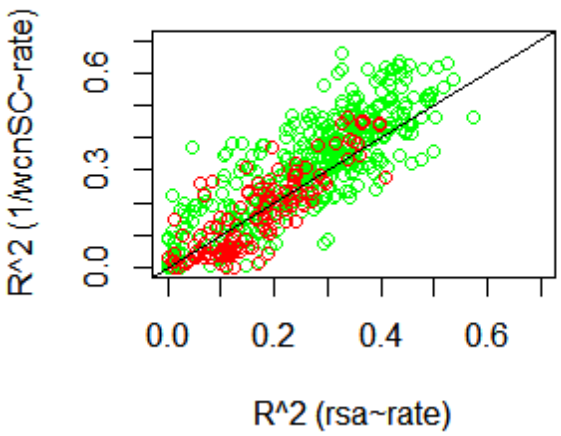

Figure 8. Comparison of R^2 values for 1/WCNSC and RSA with rate in 410 alpha-helical (green), and 130 beta-barrel (red) proteins. Each circle represents an individual protein. The line shows X=Y.Alpha-helical proteins had a higher range and maximum correlation for both metrics.

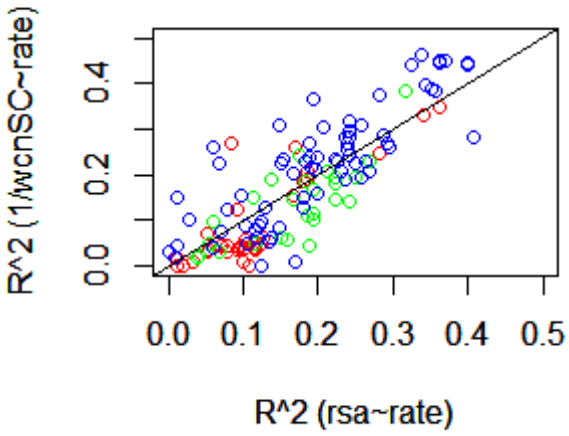

Figure 9. Comparison of R^2 values for 1/WCNSC and RSA with rate in subcategories of beta-barrel proteins: Porins (red, n=35), Monomeric/Dimeric (blue, n=69), and Others (green, n=26). Each circle represents an individual protein. The line shows X=Y.

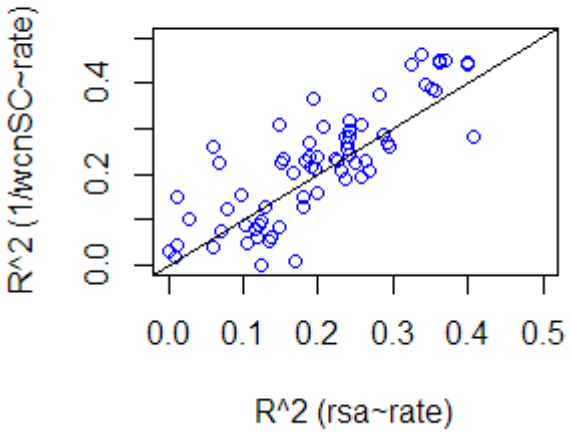

Figure 10. Distribution of R^2 values for 1/WCNSC and RSA with rate in 69monomeric/dimeric proteins. Each circle represents an individual protein. The line shows X=Y.

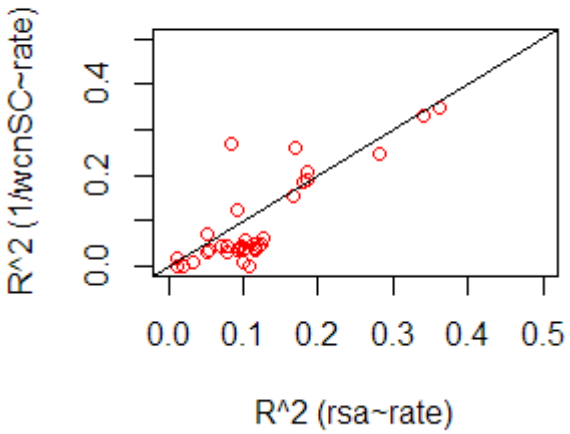

Figure 11. Distribution of R^2 values for 1/WCNSC and RSA with rate in 35 porins. We see a cluster of proteins around the 0.1 mark for RSA and 0.03 for 1/ WCNSC. Each circle represents an individual protein. The line shows X=Y.

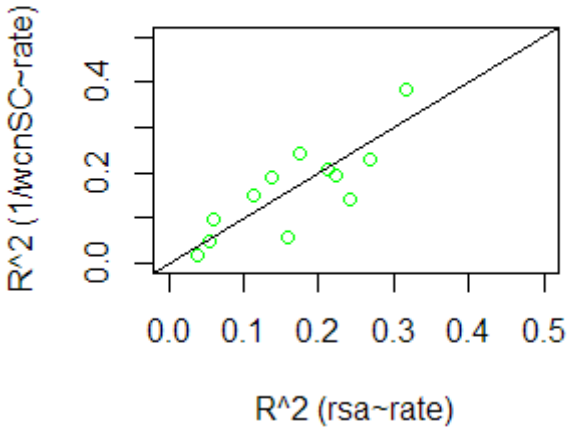

Figure 12. Distribution of R^2 values for 1/WCNSC and RSA with rate in 12 outer membrane autotransporters. Each circle represents an individual protein. The line shows X=Y.

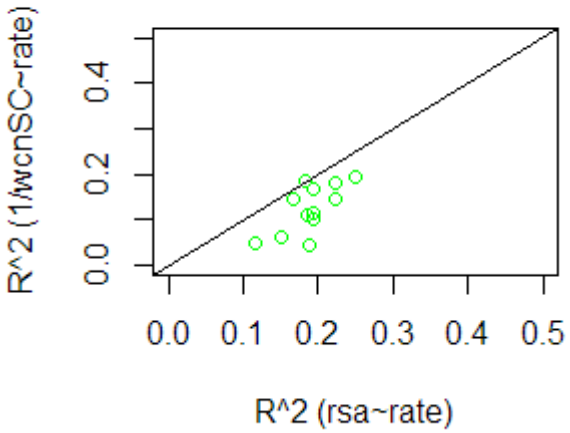

Figure 13. Distribution of R^2 values for 1/WCNSC and RSA in 12 outer membrane carboxylate channels. Each circle represents an individual protein. The line shows X=Y. Although the sample size is small, RSA outperformed WCNSC in almost all of these proteins

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402.

Echave, J., Spielman, S. J., & Wilke, C. O. (2016). Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetic*, *17*,109-121

Katoh, K. & Standley D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*,*30*(4), 772-780.

Marcos, M. L. & Echave, J. (2015). Too packed to change: side-chain packing and site-specific substitution rates in protein evolution. *PeerJ, 3*, e911

Moraes, I., Evans,G., Sanchez-Weatherby, J., Newstead, S., & Stewart, P. D. S. (2014). Membrane protein structure determination — The next generation. *BBA-Biomembranes, 1838* (1), 78-87

Pupko, T., Bell, R. E., 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,* 71-77.

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

Stamatakis, A. (2014). RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics, 30*. 1312-1313

von Heijne, G. (1997). Principles of membrane protein assembly and structure. *Progr.Biophys.Mol.Biol*., *6*6, 113-139.

Xiong , Jin. (2006). *Essential bioinformatics*. Cambridge University Press. pp. 208–. ISBN 978-0-521-84098-9.