

**A peer-reviewed version of this preprint was published in PeerJ on 30 May 2017.**

[View the peer-reviewed version](https://doi.org/10.7717/peerj.3391) (peerj.com/articles/3391), which is the preferred citable publication unless you specifically need to cite this preprint.

Sydykova DK, Wilke CO. 2017. Calculating site-specific evolutionary rates at the amino-acid or codon level yields similar rate estimates. PeerJ 5:e3391 <https://doi.org/10.7717/peerj.3391>

# Calculating site-specific evolutionary rates at the amino-acid or codon level yields similar rate estimates

Dariya K. Sydykova<sup>1</sup> and Claus O. Wilke<sup>1</sup>

<sup>1</sup>Department of Integrative Biology, Center for Computational Biology and Bioinformatics, and Institute for Cellular and Molecular Biology. The University of Texas at Austin, Austin, TX 78712, USA.

Corresponding author:

Claus Wilke<sup>1</sup>

Email address: wilke@austin.utexas.edu

## ABSTRACT

Site-specific evolutionary rates can be estimated from codon sequences or from amino-acid sequences. For codon sequences, the most popular methods use some variation of the  $dN/dS$  ratio. For amino-acid sequences, one widely-used method is called Rate4Site, and it assigns a relative conservation score to each site in an alignment. How site-wise  $dN/dS$  values relate to Rate4Site scores is not known. Here we elucidate the relationship between these two rate measurements. We simulate sequences with known  $dN/dS$ , using either  $dN/dS$  models or mutation–selection models for simulation. We then infer Rate4Site scores on the simulated alignments, and we compare those scores to either true or inferred  $dN/dS$  values on the same alignments. We find that Rate4Site scores generally correlate well with true  $dN/dS$ , and the correlation strengths increase in alignments with higher sequence divergence and higher number of taxa. Moreover, Rate4Site scores correlate nearly perfectly with inferred  $dN/dS$  values, even for small alignments with little divergence. Finally, we verify this relationship between Rate4Site and  $dN/dS$  in a variety of natural sequence alignments. We conclude that codon-level and amino-acid-level analysis frameworks are directly comparable and yield near-identical inferences.

## INTRODUCTION

Different sites in a protein evolve at different rates (Kimura and Ohta, 1974; Perutz et al., 1965), and these rate differences are shaped by the interplay of functional and structural constraints each site experiences (Echave et al., 2016). For example, protein surface sites tend to evolve faster than interior sites of a protein (Franzosa and Xia, 2009; Shahmoradi et al., 2014; Yeh et al., 2014b,a; Huang et al., 2014; Ramsey et al., 2011; Dean et al., 2002; Scherrer et al., 2012; Mirny and Shakhnovich, 1999; Zhou et al., 2008). Active sites in enzymes tend to be highly conserved (Jack et al., 2016; Dean et al., 2002; Kimura and Ohta, 1973; Huang et al., 2015), and sites involved in protein–protein interactions are somewhat more conserved than other surface sites (Mintseris and Weng, 2005; Kim et al., 2006; Franzosa and Xia, 2009; Jack et al., 2016).

Analyses of sequence variation in a structural context frequently make use of site-specific evolutionary rate estimates, and a wide variety of different methods exist to infer such rates from either codon or amino-acid sequences (Nielsen and Yang, 1998; Yang and Nielsen, 2002; Kosakovsky Pond et al., 2005; Kosakovsky Pond and Muse, 2005; Yang et al., 2000; Murrell et al., 2012; Lemey et al., 2012; Pupko et al., 2002; Fernandes and Atchley, 2008; Huang and Golding, 2014, 2015; Mayrose et al., 2004). The most widely applied methods for codon sequences are based on  $dN/dS$ , the rate of non-synonymous substitutions per non-synonymous site  $dN$  divided by the rate of synonymous substitutions per synonymous site  $dS$ . The  $dN/dS$  ratio is commonly used to infer purifying ( $dN/dS < 1$ ) or positive selection ( $dN/dS > 1$ ) in protein-coding genes (Nielsen and Yang, 1998; Goldman and Yang, 1994). The most popular method for rate inference in amino-acid sequences is Rate4Site (Pupko et al., 2002; Mayrose et al., 2004). Rate4Site assigns a score to a site as a proxy for the rate of evolution at that site. Rate4Site

is typically used to locate active sites on the protein structure, such as protein–protein interaction or protein–ligand interaction sites and catalytic sites (Mousson et al., 2005; Fischer et al., 2008; Tuncbag et al., 2009; Bradford et al., 2006; Guney et al., 2008). How  $dN/dS$  inference methods relate to Rate4Site scores is not known.

The relationship between protein structure and evolutionary variation has been investigated in many different protein families and many different datasets of varying divergence levels and taxonomic origin, and some studies have used codon sequences and codon-related methods to infer the rate of evolution (Franzosa and Xia, 2009; Shahmoradi et al., 2014; Scherrer et al., 2012; Zhou et al., 2008; Kim et al., 2006) while others have used amino-acid sequences (Ramsey et al., 2011; Yeh et al., 2014b,a; Huang et al., 2014, 2015; Jack et al., 2016; Mirny and Shakhnovich, 1999). Because of these differences in datasets and analysis approaches, it is not obvious to what extent results from different studies can be compared. To the extent that different studies produce contradictory results, and they frequently do, are these contradictions due to fundamental differences in the analyzed datasets (e.g., highly diverged sequences from many taxonomic groups vs. weakly diverged sequences from a single population) or in the employed methods to infer evolutionary rates (e.g., inference based on amino-acid sequences vs. on codon sequences)?

Here we address the second question, to what extent analyses at the codon level are comparable to analyses at the amino-acid level. Specifically, we use extensive simulations to ask how similar Rate4Site scores are to site-specific  $dN/dS$  values. We simulate sequence divergence both under  $dN/dS$  models and under mutation–selection models, and we then ask how inferred Rate4Site scores for these simulated alignments compare to (i) the true simulated  $dN/dS$  values at each site and (ii) the inferred  $dN/dS$  values obtained from the simulated alignments. We find that Rate4Site scores generally correlate well with  $dN/dS$ , in particular if both quantities are inferred from sequence data. We verify this observation on rates inferred from natural sequence alignments, and we conclude that amino-acid level and codon-level analysis of rate variation will generally yield comparable results.

## RESULTS

Both Rate4Site scores and per-site  $dN/dS$  values are measures of the extent to which selection acts on individual protein sites. The Rate4Site model decomposes evolutionary distances in amino-acid alignments into a site-specific component  $r_k$  and a branch-specific component  $t_i$ , such that the total divergence at site  $k$  along branch  $i$  can be written as  $r_k t_i$ . Here,  $r_k$  is the Rate4Site score at site  $k$  and  $t_i$  is the branch length of branch  $i$  in the phylogenetic tree. Importantly,  $r_k$  is the same at all branches in the tree and  $t_i$  is the same at all sites for each branch  $i$ . Because the rate decomposition is invariant under a rescaling of  $r'_k = Cr_k$  and  $t'_i = t_i/C$ , Rate4Site scores are not unique unless an additional normalization condition is specified as well. The Rate4Site software solves this uniqueness problem by turning the  $r_k$  into  $z$ -scores. However, the more natural normalization is to divide all  $r_k$  by their mean,  $r'_k = r_k / \sum_j r_j$ , where the sum runs over all sites in the protein. These normalized  $r'_k$  scores have the simple interpretation of providing the relative increase or decrease in substitution rate at site  $k$  compared to the average rate of substitution in the rest of the protein.

In contrast to Rate4Site scores, which are calculated from amino-acid alignments,  $dN/dS$  ratios are calculated on nucleotide alignments. They estimate the rate of non-synonymous divergence relative to the rate of synonymous divergence. However, just like in Rate4Site, in a site-specific  $dN/dS$  model evolutionary divergence is decomposed into a site-specific  $dN/dS$  value and a site-independent branch length. Thus, Rate4Site and site-specific  $dN/dS$  measure fundamentally the same quantity. The main difference is the input data (amino-acid sequences vs. codon sequences) and the normalization (relative to mean across sites vs. relative to the synonymous divergence rate  $dS$ ).

### Relationship between Rate4Site scores and true $dN/dS$

To determine the relationship between Rate4Site and  $dN/dS$  models, we began by simulating sequence evolution with known, site-specific  $dN/dS$  values and then comparing these true  $dN/dS$  values to Rate4Site scores inferred from the simulated alignments (Figure 1). We first considered the case of constant  $dS$  among all sites. Thus, for each site in each alignment, we randomly drew a  $dN$  from a uniform distribution ranging from 0.1 to 1.6. We set  $dS = 1$  for all sites, such that the  $dN/dS$  ratios similarly varied from 0.1 to 1.6. We ran simulations along a set of 25 balanced trees with different branch lengths and numbers of taxa, as used previously in a study of  $dN/dS$  inference (Spielman et al., 2016). Simulated

sequences were 100 codon sites long, and we generated 50 replicate simulations for each simulation condition.

We calculated the correlation between each site's true  $dN/dS$  and its inferred Rate4Site score to assess how well the Rate4Site scores agreed with the simulated rates. We then plotted the mean correlation strengths in replicate simulations against the simulations' branch lengths and number of taxa. We found that correlation strengths systematically increased with both increasing branch lengths and number of taxa (Figure 2A). While correlations were low to moderate for the least-diverged and smallest alignments, for larger and/or more diverged alignments correlations approached values ranging from 0.8 to 1.0.

We also performed a comparison of the magnitude of Rate4Site scores and  $dN/dS$  scores, by calculating root-mean-square deviations (RMSD) between these scores. Because these two types of scores are not measured in the same units, this comparison may not seem meaningful. However, we can convert both types of scores into normalized, relative scores by dividing them by their mean score. These normalized scores have comparable interpretations and RMSDs between them are meaningful quantities.

We found that RMSD values were generally moderate, between 0.1 and 0.6 (Figure 2B). They declined with both increasing number of taxa and increasing sequence divergence. However, overall RMSD depended more strongly on branch length than on the number of taxa. Visual inspection of normalized Rate4Site scores plotted against normalized  $dN/dS$  scores revealed no major systematic differences between these scores (Figure 3). Differences seemed to be driven primarily by the sampling noise inherent in estimating site-specific evolutionary rates.

We repeated the same analysis but now using simulations in which  $dS$  was allowed to vary among sites as well. The  $dN/dS$  range was kept the same as before (0.1 to 1.6), but now each site had its own unique  $dS$ , randomly chosen from a uniform distribution ranging from 0.5 to 2. Overall, we found similar patterns in the variable  $dS$  case as we had seen for constant  $dS$  (compare Figures 2C, D to Figures 2A, B). However, correlations were generally somewhat weaker (Figure 2C) and RMSD values somewhat higher (Figure 2D) than what we had observed for constant  $dS$ . These results were to be expected, since Rate4Site as an amino-acid based metric does not take synonymous variation into account, and thus the  $dS$  variation acts simply as added random noise on the  $dN/dS$  scores compared to Rate4Site scores.

### Relationship between Rate4Site scores and scaled selection coefficients

The  $dN/dS$  model is not a particularly realistic model of sequence evolution, because it does not have the notion of an underlying fitness landscape. A mutation increasing fitness should fix much more rapidly than the reverse mutation decreasing fitness. However, in a  $dN/dS$  model, both mutations fix at the same rate. To increase realism in our analysis, we next investigated Rate4Site in the context of sequences simulated with mutation–selection (MutSel) models. MutSel models are specified by scaled selection coefficients, which describe the relative fitness of different amino acids (or codons) at each site in a sequence. We can derive expected  $dN/dS$  values from these scaled selection coefficients (Spielman and Wilke, 2015b; dos Reis, 2015) and hence we can ask how Rate4Site scores compare to the predicted  $dN/dS$  values in MutSel models.

For this analysis, we employed previously published sequence alignments from Spielman et al. (2016). These alignments had been simulated using the Halpern and Bruno model (HB98) (Halpern and Bruno, 1998) along the same 25 phylogenetic trees we employed in our previous analysis (five branch lengths in all pairwise combinations with five numbers of taxa). As before, there were 50 replicates per simulation condition, and we again had one dataset with constant  $dS$  and one with variable  $dS$ . In the dataset with constant  $dS$ , all synonymous codons have the same fitness (neutral synonymous codons). In the dataset with variable  $dS$ , there are fitness differences among synonymous codons (non-neutral synonymous codons). See Spielman et al. (2016) for details of parameter choices.

Our results for sequences simulated with MutSel models were broadly similar to our results for sequences simulated with  $dN/dS$  models (Figure 4). As before, correlations increased and approached 1 with increasing branch lengths and numbers of taxa, and RMSDs commensurately decreased. However, correlation strengths were consistently lower and RMSD values higher for the MutSel datasets than for the  $dN/dS$  datasets at the same sequence divergence and number of taxa (compare Figure 4 to Figure 2). As before, differences between normalized Rate4Site scores and normalized true  $dN/dS$  values seemed to be driven primarily by the sampling noise inherent in estimating site-specific evolutionary rates (Figure 5). Finally, we found only minor differences between simulations with neutral synonymous codons and simulations with non-neutral synonymous codons. However, in the latter case, correlations were generally

153 slightly lower and RMSD values somewhat higher (compare Figure 4C and D to Figure 4A and B).

# 154 Relationship between Rate4Site scores and inferred $dN/dS$

155 The preceding analyses asked to what extent Rate4Site scores reflect the known underlying parameters  
156 used to generate the sequence alignments. An alternative question, possibly more applicable to practical  
157 sequence analysis, is to what extent Rate4Site scores mirror  $dN/dS$  values inferred on the same sequence  
158 data. To address this second question, we inferred site-wise  $dN/dS$  values for all sites in all alignments  
159 studied in the previous two subsections. The  $dN/dS$  values were inferred using the one-rate fixed-effects  
160 likelihood method (FEL1) implemented in HyPhy (Kosakovsky Pond et al., 2005). The FEL1 method  
161 assigns one  $dN$  value per site and one  $dS$  value across all sites in the sequence (Spielman et al., 2016).  
162 Therefore, the variation in the inferred  $dN/dS$  values is captured entirely by  $dN$ .

163 We found that Rate4Site scores were very highly correlated to inferred  $dN/dS$  across all datasets and  
164 simulation conditions (Figures 6 and 7). For sequences simulated with the  $dN/dS$  model, correlations for  
165 all branch lengths exceeded 0.8 and approached 1 as the sequence divergence increased (Figure 6A, C).  
166 RMSD values were large for the smallest and least-diverged alignments but declined rapidly as either  
167 branch length or number of taxa increased (Figure 6B, D). There was little difference between alignments  
168 simulated with constant  $dS$  and with variable  $dS$  (Figure 6A vs. 6C and Figure 6B vs. 6D). Results  
169 for sequences simulated with MutSel models were similar (Figure 7). The main difference was that  
170 correlation coefficients were more sensitive to the number of taxa. For the lowest number of taxa (128)  
171 correlation coefficients were systematically lower when sequences were simulated with MutSel models  
172 rather than with  $dN/dS$  models. The pattern reversed for the highest number of taxa (2048). RMSDs  
173 however were systematically higher for sequences simulated with MutSel models rather than with  $dN/dS$   
174 models. In all cases, there was little difference between sequences simulated with neutral synonymous  
175 codons and with non-neutral synonymous codons (Figure 7A vs. 7C and Figure 7B vs. 7D).

176 Finally, we asked to what extent the results found for simulated sequences carry over to natural  
177 sequence alignments. We inferred both Rate4Site scores and site-wise  $dN/dS$  in two distinct datasets,  
178 one consisting of several membrane proteins in mammals (taken from Spielman and Wilke (2013)) and  
179 one consisting of several alignments of HIV-1 genes (taken from Meyer and Wilke (2015b)). For both  
180 datasets, we found that Rate4Site scores and  $dN/dS$  were highly correlated, with correlation coefficients  
181 exceeding 0.8 in all cases (Figure 8). We concluded that our simulation results carry over to natural  
182 sequence alignments, and that Rate4Site scores are generally comparable to inferred  $dN/dS$  values.

# 183 DISCUSSION

184 We have compared codon-level site-specific evolutionary rates estimated via  $dN/dS$  to amino-acid level  
185 site-specific rates estimated via Rate4Site. We have found that Rate4Site scores correlate well with the  
186 known true  $dN/dS$  values both in sequences simulated with a  $dN/dS$  model and in sequences simulated  
187 with a MutSel model. Correlations generally increase and approach 1 for more diverged sequence  
188 alignments and for alignments with more taxa. Correlations are generally somewhat stronger when there  
189 is no variation in  $dS$  among sites, though this effect is minor. We have also compared Rate4Site scores to  
190 inferred  $dN/dS$  values and have found nearly perfect correlations between the two measures, even for  
191 less diverged and smaller alignments. Finally, we have verified the relationship between Rate4Site scores  
192 and  $dN/dS$  in a set of natural alignments, and again have found nearly perfect correlations between the  
193 two rate estimates.

194 Surprisingly, even in scenarios of low sequence divergence or few taxa, when Rate4Site scores are  
195 only weakly correlated with the true  $dN/dS$ , we have found that they nevertheless correlate highly with  
196 inferred  $dN/dS$ . For all levels of sequence divergence and numbers of taxa the two rate measurements  
197 had nearly perfect correlations. Moreover,  $dS$  variation seemed to have a negligible influence on the  
198 correlation strengths. A strong relationship between these two rate measurements was also evident in  
199 natural sequence alignments. For the majority of alignments that we considered, correlation coefficients  
200 were in excess of 0.9. For the two alignments with the lowest correlation coefficients, of 0.83 for both  
201 alignments, the number of sequences were 19 and 22 and sequence divergence was low. Thus, unless  
202 alignments are very small and/or have very little divergence, Rate4Site scores and site-specific  $dN/dS$   
203 can be expected to correlate strongly in all cases. These findings demonstrate that Rate4Site and  $dN/dS$   
204 approaches have comparable ability to infer rates from sequence alignments. For sufficiently diverged



alignments, both methods accurately recover the true underlying rates. And for alignments with less divergence, mis-estimate the underlying rates in a similar fashion.

The  $dN/dS$  metric is frequently used to identify sites under positive selection in viruses (Vijaykrishna et al., 2008; Wood et al., 2009; Demogines et al., 2013; Meyer and Wilke, 2015a). By contrast, Rate4Site has been mostly applied to identify conserved sites that correspond to protein–protein interaction sites or active sites in enzyme (Mousson et al., 2005; Fischer et al., 2008; Tuncbag et al., 2009; Bradford et al., 2006; Guney et al., 2008). Our results here show that for purposes of finding the most conserved or most rapidly varying sites in a sequence alignments, both methods would likely identify similar sites. One advantage of the  $dN/dS$  approach, of course, is the ability to test whether  $dN/dS$  is significantly above 1. When using Rate4Site scores, one can identify the most rapidly varying sites but one cannot run a statistical test that would determine whether the site is positively selected or not.

Recently, there has been considerable interest in linking site-specific rate variation to structural features of proteins (Echave et al., 2016). Studies addressing this topic have considered both  $dN/dS$ -based methods (Scherrer et al., 2012; Franzosa and Xia, 2009; Shahmoradi et al., 2014; Kim et al., 2006; Meyer and Wilke, 2015b,a) and Rate4Site scores (Huang et al., 2014; Yeh et al., 2014b,a; Jack et al., 2016; Huang et al., 2015), though these studies have generally been done on disparate datasets. The extent to which results found with  $dN/dS$  carry over to Rate4Site and vice versa has not been clear. Our findings here show that since the two methods infer rates that correlate strongly with each other, either type of inferred rate should produce comparable correlation strengths with structural features such as solvent accessibility.

We note several caveats to our conclusions. First, our simulated alignments were generally large and diverged, even for the smallest number of taxa and lowest branch lengths. Even smaller and/or less diverged alignments will yield more noisy, less reliable Rate4Site inferences. Second, all our simulated alignments were obtained under the assumption that sites evolve independently from each other and that the rate of evolution does not change over time. These assumptions will generally increase the congruence between the true, simulated  $dN/dS$  and the inferred Rate4Site score. However, the strong correlations we observed between Rate4Site scores and  $dN/dS$  in several natural alignments demonstrate that neither of these assumptions and limitations fundamentally invalidate our main findings. Amino-acid level and codon-level analyses of sequence data will generally yield comparable estimates of site-specific rates of evolution.

## METHODS

### Generation of simulated alignments

Our simulation approach was similar to the one employed by Spielman et al. (2016). In brief, we first generated a set of balanced, binary trees with different branch lengths and numbers of taxa, using the R package *ape* (Paradis et al., 2004). We then simulated sequence evolution along these trees using the python library *pyvolve* (Spielman and Wilke, 2015a).

We generated a total of 25 trees, using all pairwise combinations of five different branch lengths and five different numbers of taxa. The branch lengths we used were 0.0025, 0.01, 0.04, 0.16, and 0.64. These numbers indicate the divergence in mutations per site between two nodes in a tree. The numbers of taxa we used were 128, 256, 512, 1024, and 2048.

To generate alignments with site-specific  $dN/dS$  values, we simulated sequences with 100 codon sites using a site-specific Muse-Gaut model (Muse and Gaut, 1994). To simulate sequences with constant  $dS$ , we set  $dS = 1$  at all sites and set  $dN$  at each site to a different value randomly drawn from a uniform distribution between 0.1 and 1.6. To simulate sequences with variable  $dS$ , we assigned each site a distinct  $dN$  and  $dS$  value, by first choosing a randomly drawn  $dN/dS$  value, then choosing a randomly drawn  $dS$  value, and then setting  $dN = dS \times (dN/dS)$ . The  $dN/dS$  values were drawn from a uniform distribution between 0.1 and 1.6, and the  $dS$  values were drawn from a uniform distribution between 0.5 and 2. We generated 50 replicate sequence alignments for each combination of branch length, number of taxa, and choice of  $dS$  (constant or variable), for a total of 2500 sequence alignments.

For sequences simulated according to MutSel models, we used sequence alignments previously published in Spielman et al. (2016), specifically the alignments simulated with unequal nucleotide frequencies. These sequences were simulated using the Halpern and Bruno model (HB98) (Halpern and Bruno, 1998), and we had alignments for the same tree parameters,  $dS$  variation (constant/variable), and replicate numbers as our simulations of the  $dN/dS$  model, again for a total of 2500 sequence alignments.

## Rate inference

To acquire the Rate4Site scores, the simulated sequences were translated into amino acids using biopython. The translated sequences were inputted into Rate4Site along with their corresponding trees. We ran Rate4Site with the following options:

```
rate4site -s aln_file -t tree_file -o norm_rates_file \
-y orig_rates_file
```

Here, `aln_file` is the input fasta file with aligned sequences. The file `tree_file` contains the phylogenetic tree. The file `norm_rates_file` is the output file into which Rate4Site writes  $z$ -normalized rate scores, and `orig_rates_file` is the output file into which Rate4Site writes original rate scores. The option `-y` causes Rate4Site to output original scores. (By default, Rate4Site only outputs  $z$ -transformed scores.) In our analysis we used only the original scores, renormalized such that they had a mean of 1.

We inferred site-specific  $dN/dS$  using the one-parameter fixed-effects likelihood method (FEL1) implemented in HyPhy (Kosakovsky Pond et al., 2005). We ran HyPhy using the FEL1 script provided in Spielman et al. (2016). After running the  $dN/dS$  inference, we explicitly set  $dN/dS = 0$  at all sites that did not experience any amino-acid changes. We did so because the FEL1 method assigns a site-wise  $dN/dS$  of 1 to completely conserved sites that contain no synonymous and no non-synonymous mutations. However,  $dN/dS$  should equal 0 at such sites in a one-parameter model, which implicitly assumes that  $dS$  is the same at all sites and hence will be non-zero even at completely conserved sites.

## Analysis of natural sequence alignments

For analysis of natural sequence alignments, we used data from Spielman and Wilke (2013) and Meyer and Wilke (2015b). From Spielman and Wilke (2013), we acquired unaligned sequence data for six arbitrarily chosen membrane proteins: Mannose-6-phosphate receptor M6PR (Ensembl transcript ID: ENST00000000412), CD74 (Ensembl transcript ID: ENST000000009530), CD4 (Ensembl transcript ID: ENST00000011653), G protein-coupled receptor class C, GPRC5A (Ensembl transcript ID: ENST00000014914), Gamma-aminobutyric acid type A receptor, GABRA1 (Ensembl transcript ID: ENST00000023897), and TNF receptor superfamily member 17, TNFRSF17 (Ensembl transcript ID: ENST00000053243). We aligned the amino-acid sequences using MAFFT 7.305b (Multiple Alignment using Fast Fourier Transform) (Katoh and Standley, 2013). We ran MAFFT using default options with:

```
mafft input_fasta_file > output_fasta_file
```

Here, `input_fasta_file` is the fasta file containing sequences to be aligned and `output_fasta_file` is the output file into which the alignment is written. The aligned amino-acid sequences were subsequently back-translated to codon sequences using the unaligned nucleotide sequences.

For sequence data from Meyer and Wilke (2015b), we used the aligned amino and nucleotide sequence files for all of the proteins used in the paper.

We inferred phylogenetic trees from amino-acid sequences using RAxML (Stamatakis, 2014). We ran RAxML with the following options:

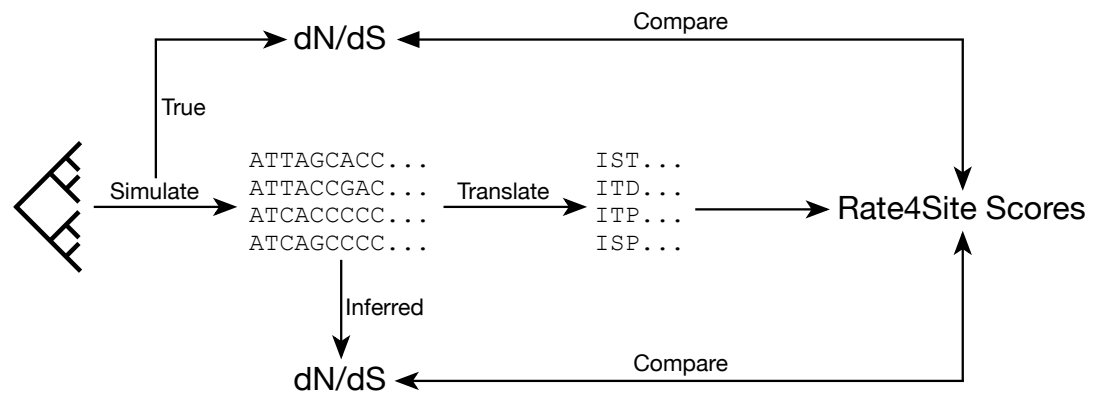
```
raxmlHPC-PTHREADS-SSE3 -T 48 -s fasta_file -w output_directory \
-n tree_name -m PROTCATLG -p 1
```

Here, `fasta_file` is the input file containing the aligned sequences. RAxML outputs all output files into the directory indicated by `output_directory`, and `tree_name` is the name for the output tree files. The option `-m PROTCATLG` picks a protein CAT model with LG matrix for the tree inference, and the option `-p 1` generates a random number seed for parsimony inference.

Finally, for all natural sequence alignments, we inferred Rate4Site scores and per-site  $dN/dS$  values as described in the subsection “Rate inference.”

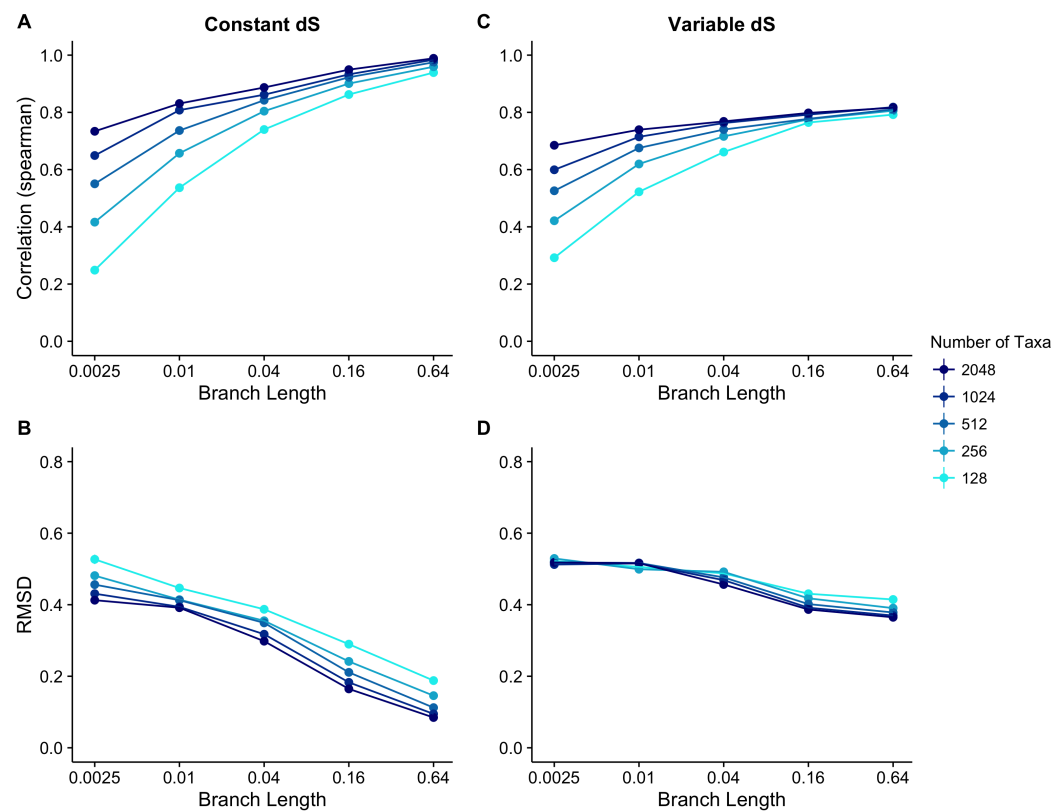
## Data availability

Processed data and code can be found at [https://github.com/wilkelab/r4s\\_benchmark](https://github.com/wilkelab/r4s_benchmark)

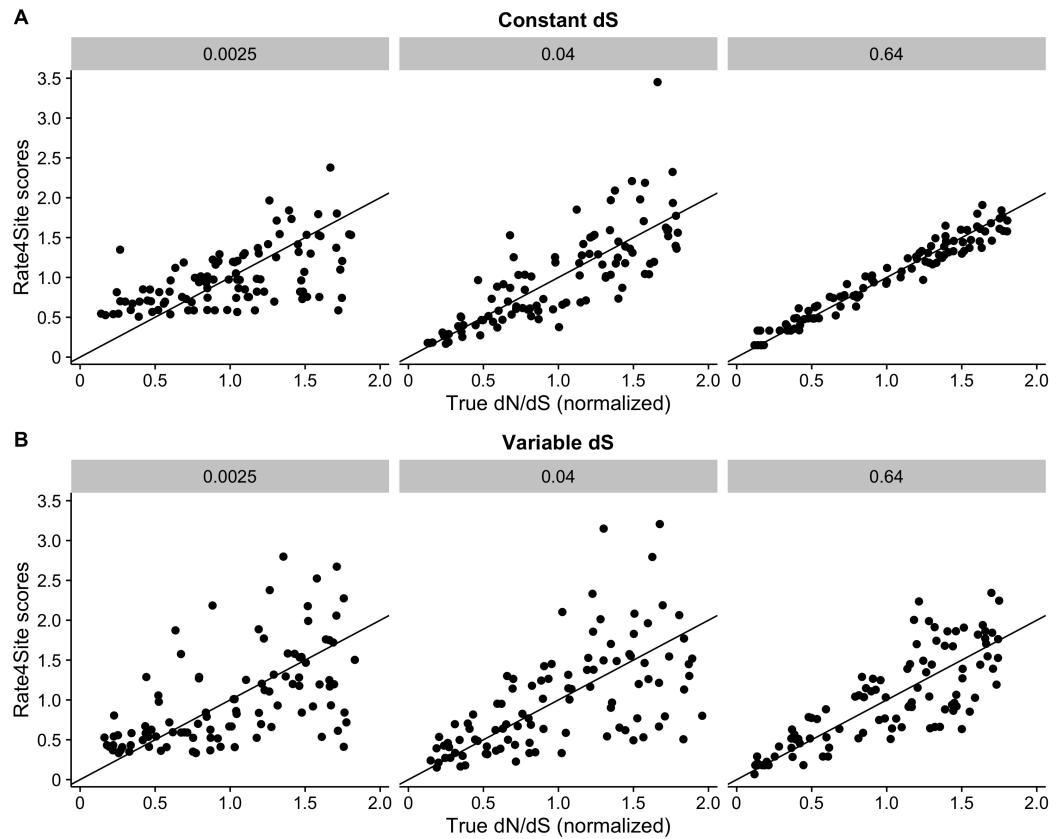


**Figure 1.** Analysis approach. We assess the relationships between  $dN/dS$  values and Rate4Site scores by simulating sequences with known  $dN/dS$  values and then comparing either known or inferred  $dN/dS$  values for these simulated alignments to the Rate4Site scores inferred on the same alignments.

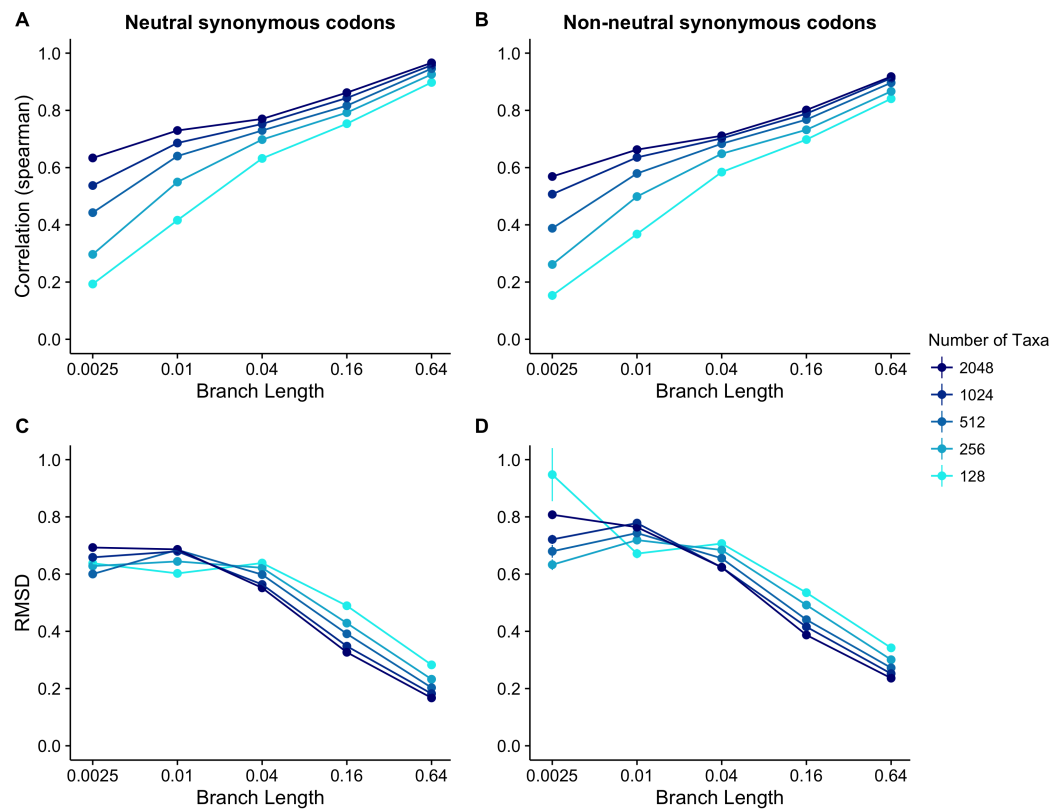




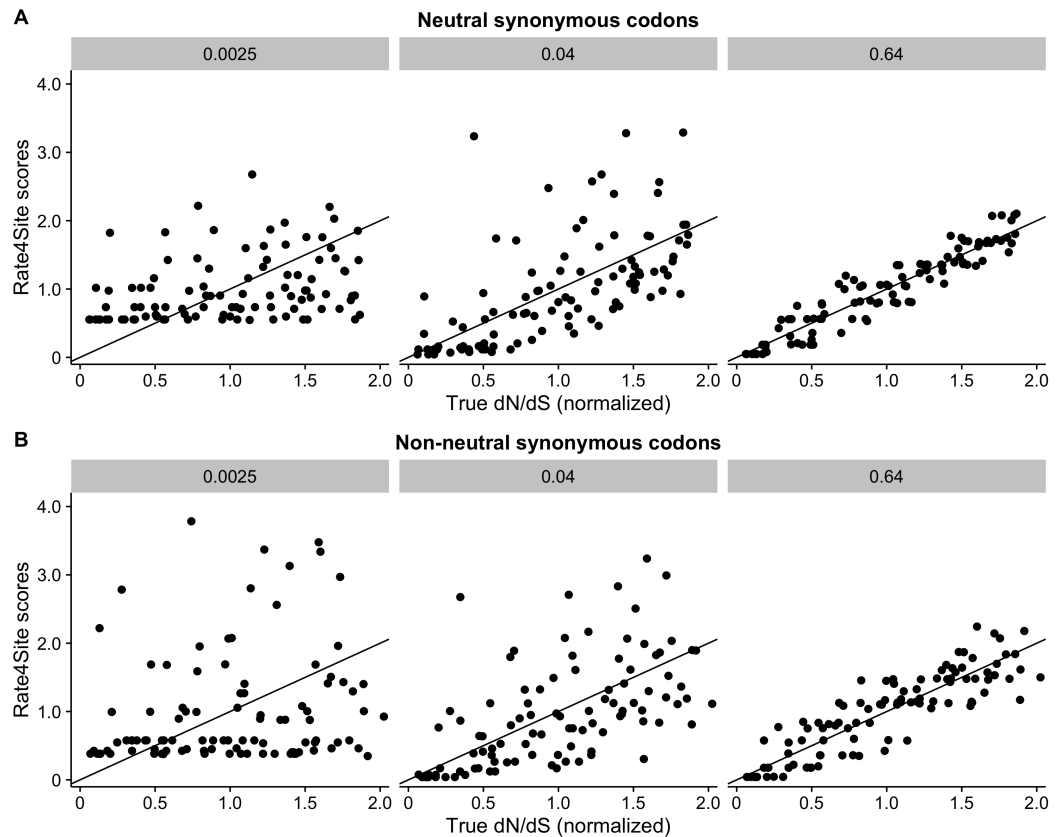
**Figure 2.** Relationship between Rate4Site scores and true site-specific  $dN/dS$  for simulations performed with  $dN/dS$  (Muse-Gaut) models. Each point represents the mean over 50 replicate simulations. The error bars represent the standard error. In nearly all cases, error bars are smaller than the symbol size. (A) Correlations and (B) RMSD values between Rate4Site scores and true  $dN/dS$ , for the sequences simulated with constant  $dS = 1$ . (C) Correlations and (D) RMSD values between Rate4Site scores and true  $dN/dS$ , for the sequences simulated with variable  $dS$ .



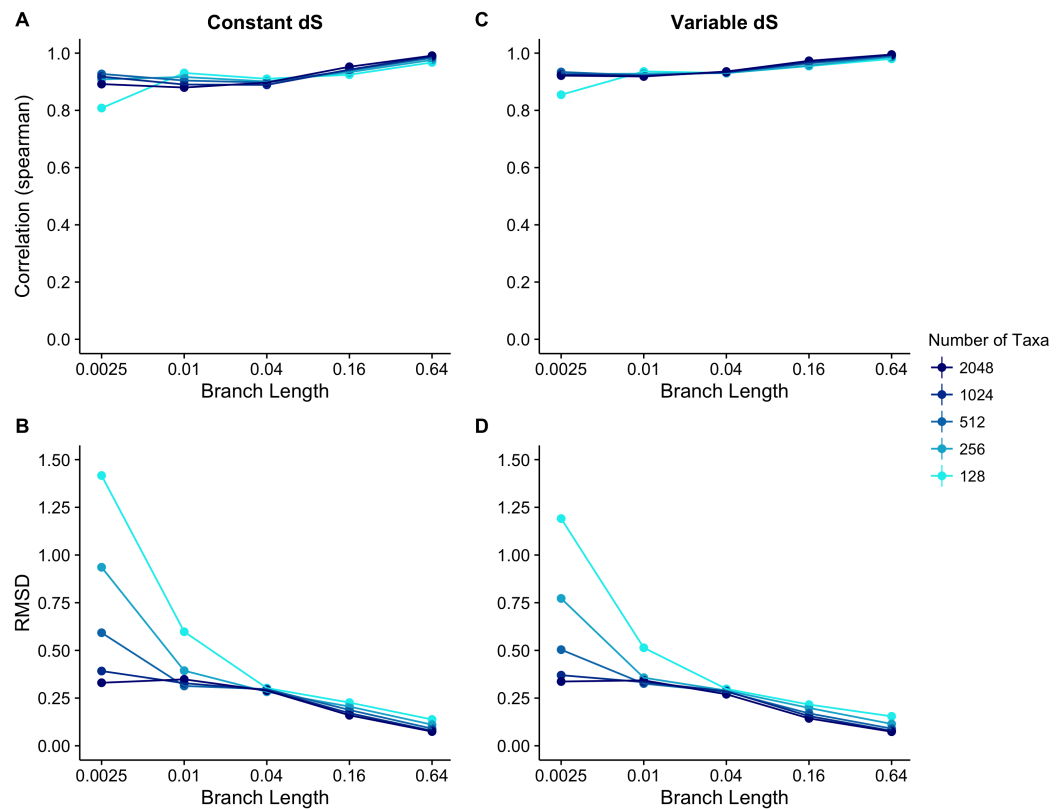
**Figure 3.** Rate4Site scores vs. true normalized  $dN/dS$  for a few example alignments simulated with a  $dN/dS$  model. Each point represents one site in the simulated alignment, and the diagonal line represents the  $x = y$  line. Numbers above each subplot indicate the branch length of the alignment, and the number of taxa was 512 in all cases. (A) Simulations with constant  $dS = 1$ . (B) Simulations with variable  $dS$ .



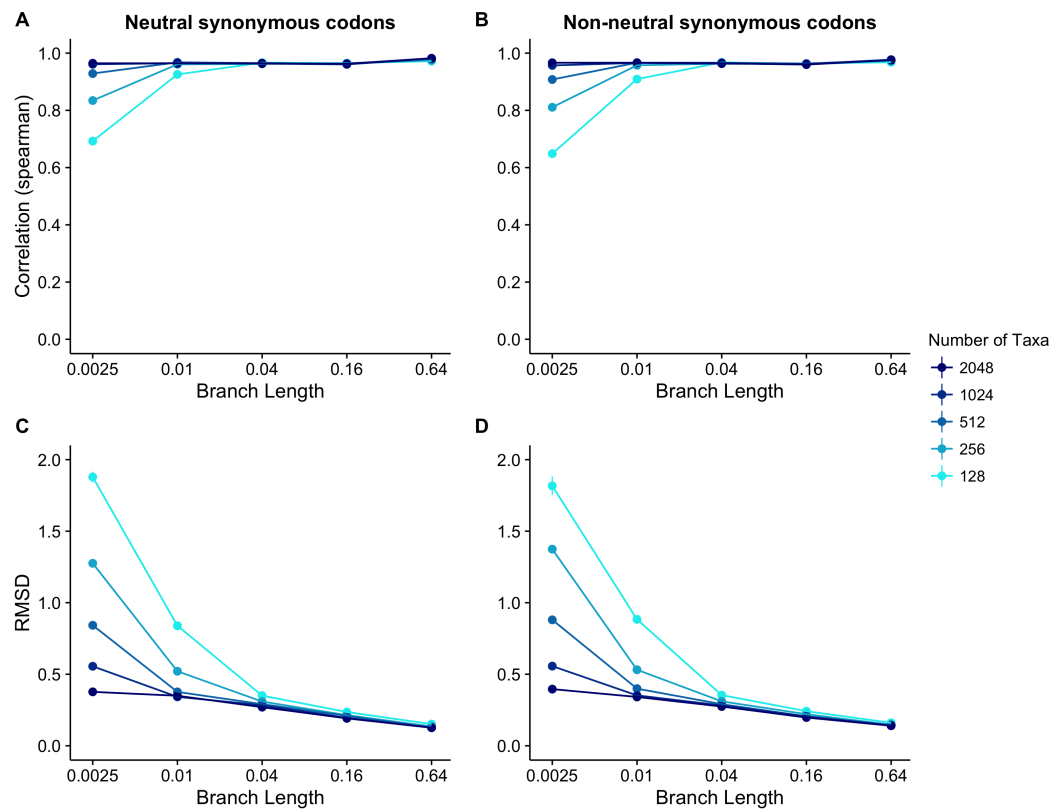
**Figure 4.** Relationship between Rate4Site scores and true site-specific  $dN/dS$  for simulations performed with MutSel (Halpern-Bruno) models. Each point represents the mean over 50 replicate simulations. The error bars represent the standard error. In nearly all cases, error bars are smaller than the symbol size. (A) Correlations and (B) RMSD values between Rate4Site scores and true  $dN/dS$ , for the sequences simulated without codon bias (neutral synonymous codons). (C) Correlations and (D) RMSD values between Rate4Site scores and true  $dN/dS$ , for the sequences simulated with codon bias (non-neutral synonymous codons).



**Figure 5.** Rate4Site scores vs. true normalized  $dN/dS$  for a few example alignments simulated with a MutSel (Halpern-Bruno) models. Each point represents one site in the simulated alignment, and the diagonal line represents the  $x = y$  line. Numbers above each subplot indicate the branch length of the alignment, and the number of taxa was 512 in all cases. (A) Simulations without codon bias (neutral synonymous codons). (B) Simulations with with codon bias (non-neutral synonymous codons).

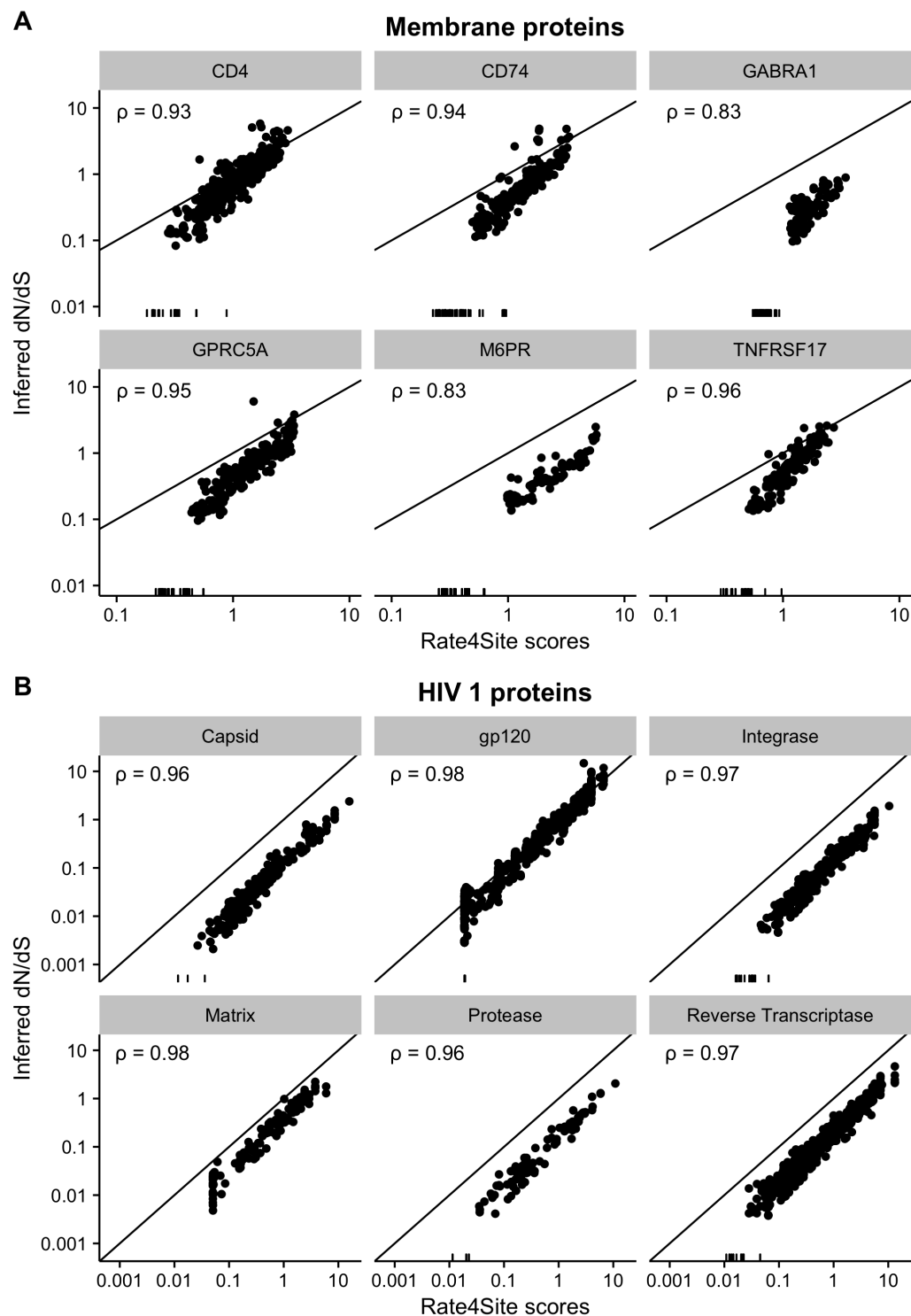


**Figure 6.** Relationship between Rate4Site scores and inferred site-specific  $dN/dS$  for simulations performed with  $dN/dS$  (Muse-Gaut) models. Each point represents the mean over 50 replicate simulations. The error bars represent the standard error. In nearly all cases, error bars are smaller than the symbol size. (A) Correlations and (B) RMSD values between Rate4Site scores and inferred  $dN/dS$ , for the sequences simulated with constant  $dS = 1$ . (C) Correlations and (D) RMSD values between Rate4Site scores and inferred  $dN/dS$ , for the sequences simulated with variable  $dS$ .



**Figure 7.** Relationship between Rate4Site scores and inferred site-specific  $dN/dS$  for simulations performed with MutSel (Halpern-Bruno) models. Each point represents the mean over 50 replicate simulations. The error bars represent the standard error. In nearly all cases, error bars are smaller than the symbol size. (A) Correlations and (B) RMSD values between Rate4Site scores and inferred  $dN/dS$ , for the sequences simulated without codon bias (neutral synonymous codons). (C) Correlations and (D) RMSD values between Rate4Site scores and inferred  $dN/dS$ , for the sequences simulated with codon bias (non-neutral synonymous codons).





**Figure 8.** Inferred  $dN/dS$  vs. Rate4Site scores for natural sequence alignments. Each dot represents one site in the respective alignment, and the diagonal line represents the  $x = y$  line. Rugs along the  $x$ -axis show sites with  $dN/dS < 0.001$ . Correlation coefficients are Spearman  $\rho$ , and all correlations are significant ( $p < 10^{-15}$  throughout). Note that  $dN/dS$  values were not normalized to a mean of one here, unlike Figures 3 and 5. (A) Inferred  $dN/dS$  vs. Rate4Site scores for six membrane proteins taken from Spielman and Wilke (2013). (B) Inferred  $dN/dS$  vs. Rate4Site scores for six HIV-1 proteins taken from Meyer and Wilke (2015b).

## REFERENCES

- Bradford, J. R., Needham, C. J., Bulpitt, A. J., and Westhead, D. R. (2006). Insights into protein–protein interfaces using a Bayesian network prediction method. *J. Mol. Biol.*, 362:365–386.
- Dean, A. M., Neuhauser, C., Grenier, E., and Golding, G. B. (2002). The pattern of amino acid replacements in  $\alpha/\beta$ -barrels. *Mol. Biol. Evol.*, 19:1846–1864.
- Demogines, A., Abraham, J., Choe, H., Farzan, M., and Sawyer, S. L. (2013). Dual host-virus arms races shape an essential housekeeping protein. *PLoS Biol.*, 11:e1001571.
- dos Reis, M. (2015). How to calculate the non-synonymous to synonymous rate ratio of protein-coding genes under the Fisher-Wright mutation-selection. *Biol. Lett.*, 11:20141031.
- Echave, J., Spielman, S. J., and Wilke, C. O. (2016). Causes of evolutionary rate variation among protein sites. *Genetics*, 17:109–121.
- Fernandes, A. D. and Atchley, W. R. (2008). Site-specific evolutionary rates in proteins are better modeled as non-independent and strictly relative. *Bioinformatics*, 24:2177–2183.
- Fischer, J. D., Mayer, C. E., and Söding, J. (2008). Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics*, 24:613–620.
- 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.
- Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, 11:725–736.
- Guney, E., Tuncbag, N., Keskin, O., and Gursoy, A. (2008). HotSpring: database of computational hot spots in protein interfaces. *Nucleic Acids Research*, 36:D662–D666.
- Halpern, A. L. and Bruno, W. J. (1998). Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.*, 15:910–917.
- Huang, T.-T., Del Valle 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.
- Huang, Y. F. and Golding, G. B. (2014). Phylogenetic Gaussian process model for the inference of functionally important regions in protein tertiary structures. *PLoS Comput. Biol.*, 10:e1003429–e1003412.
- Huang, Y. F. and Golding, G. B. (2015). FuncPatch: a web server for the fast bayesian inference of conserved functional patches in protein 3D structures. *Bioinformatics*, 31:523–531.
- Huang, Y.-W., Chang, C.-M., Lee, C.-W., and Hwang, J.-K. (2015). The conservation profile of a protein bears the imprint of the molecule that is evolutionarily coupled to the protein. *Proteins*, 83:1407–1413.
- Jack, B. R., Meyer, A. G., Echave, J., and Wilke, C. O. (2016). Functional sites induce long-range evolutionary constraints in enzymes. *PLoS Biol.*, 14:e1002452.
- 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.
- Kim, P. M., Lu, L. J., Xia, Y., and Gerstein, M. B. (2006). Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, 314:1938–1941.
- Kimura, M. and Ohta, T. (1973). Mutation and evolution at the molecular level. *Genetics*, 73:19–35.
- Kimura, M. and Ohta, T. (1974). On some principles governing molecular evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 71:2848–2852.
- Kosakovsky Pond, S. L., Frost, S. W. D., and Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenetics. *Bioinformatics*, 21:676–679.
- Kosakovsky Pond, S. L. and Muse, S. V. (2005). Site-to-site variation of synonymous substitution rates. *Mol. Biol. Evol.*, 22:2375–2385.
- Lemey, P., Minin, V. N., Bielejec, F., Kosakovsky Pond, S. L., and Suchard, M. A. (2012). A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. *Bioinformatics*, 28:3248–3256.
- Mayrose, I., Graur, D., Ben-Tal, N., and Pupko, T. (2004). Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.*, 21:1781–1791.
- Meyer, A. G. and Wilke, C. O. (2015a). Geometric constraints dominate the antigenic evolution of influenza H3N2 hemagglutinin. *PLoS Pathog.*, 11:e1004940.
- Meyer, A. G. and Wilke, C. O. (2015b). The utility of protein structure as a predictor of site-wise  $dN/dS$

- varies widely among HIV-1 proteins. *J. R. Soc. Interface*, 12:20150579.
- Mintseris, J. and Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc. Natl Acad. Sci. USA*, 102:10930–10935.
- Mirny, L. A. and Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.*, 291:177–196.
- Mousson, F., Lautrelle, A., Thuret, J. Y., Agez, M., Courbeyrette, R., Amigues, B., Becker, E., Neumann, J. M., Guerois, R., Mann, C., and Ochsenbein, F. (2005). Structural basis for the interaction of Asf1 with histone H3 and its functional implications. *Proc. Natl. Acad. Sci. U.S.A.*, 102:5975–5980.
- Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., and Kosakovsky Pond, S. L. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.*, 8(7):e1002764.
- Muse, S. V. and Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.*, 11:715–724.
- Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148:929–936.
- Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290.
- Perutz, M. F., Kendrew, J. C., and Watson, H. C. (1965). Structure and function of haemoglobin. II. Some relations between polypeptide chain configuration and amino acid sequence. *J. Mol. Biol.*, 13:669–678.
- Pupko, T., Bell, R. E., Mayrose, I., Glaser, F., and Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in the proteins by surface mapping evolutionary determinants within their homologues. *Bioinformatics*, 18:S71–S77.
- 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*, 188:479–488.
- Scherrer, M. P., Meyer, A. G., and Wilke, C. O. (2012). Modeling coding-sequence evolution within the context of residue solvent accessibility. *BMC Evol. Biol.*, 12:179.
- 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. *J. Mol. Evol.*, 79:130–142.
- Spielman, S. J., Wan, S., and Wilke, C. O. (2016). A comparison of one-rate and two-rate inference frameworks for site-specific  $dN/dS$  estimation. *Genetics*, 204:499–511.
- Spielman, S. J. and Wilke, C. O. (2013). Membrane environment imposes unique selection pressures on transmembrane domains of G protein-coupled receptors. *J. Mol. Evol.*, 76:172–182.
- Spielman, S. J. and Wilke, C. O. (2015a). Pyvolve: A flexible python module for simulating sequences along phylogenies. *PLoS ONE*, 10:e0139047.
- Spielman, S. J. and Wilke, C. O. (2015b). The relationship between  $dN/dS$  and scaled selection coefficients. *Mol. Biol. Evol.*, 32:1097–1108.
- Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30:1312–1313.
- Tuncbag, N., Gursoy, A., and Keskin, O. (2009). Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics*, 25:1513–1520.
- Vijaykrishna, D., Bahl, J., Riley, S., Duan, L., Zhang, J. X., Chen, H., Malik Peiris, J. S., Smith, G. J. D., and Guan, Y. (2008). Evolutionary dynamics and emergence of panzootic H5N1 influenza viruses. *PLoS Pathog.*, 4:e1000161.
- Wood, N., Bhattacharya, T., Keele, B. F., Giorgi, E., Liu, M., Gaschen, B., Daniels, M., Ferrari, G., Haynes, B. F., McMichael, A., Shaw, G. M., Hahn, B. H., Korber, B., and Seio, C. (2009). HIV evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of APOBEC. *PLoS Pathog.*, 5:e1000414.
- Yang, Z. and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.*, 19:908–917.
- Yang, Z. H., Nielsen, R., Goldman, N., and Pedersen, A. M. K. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155:431–449.
- 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.

- 415 Yeh, S.-W., Liu, J.-W., Yu, S.-H., Shih, C.-H., Hwang, J.-K., and Echave, J. (2014b). Site-specific
- 416 structural constraints on protein sequence evolutionary divergence: local packing density versus solvent
- 417 exposure. *Mol. Biol. Evol.*, 31:135–139.
- 418 Zhou, T., Drummond, D. A., and Wilke, C. O. (2008). Contact density affects protein evolutionary rate
- 419 from bacteria to animals. *J. Mol. Evol.*, 66:395–404.