MBE Advance Access published September 11, 2016

Extensively Parameterized Mutation–Selection Models Reliably Capture Site-Specific Selective Constraint

Stephanie J. Spielman,*[1,2,3] and Claus O. Wilke[1,2]

[1]Department of Integrative Biology, Center for Computational Biology and Bioinformatics, The University of Texas at UT Austin, Austin

[2]Institute for Cellular and Molecular Biology, The University of Texas at UT Austin, Austin

[3]Present address: Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia

***Corresponding author:** E-mail: stephanie.spielman@gmail.com.

Associate editor: Sergei Kosakovsky

Abstract

The mutation–selection model of coding sequence evolution has received renewed attention for its use in estimating site-specific amino acid propensities and selection coefficient distributions. Two computationally tractable mutation–selection inference frameworks have been introduced: One framework employs a fixed-effects, highly parameterized maximum likelihood approach, whereas the other employs a random-effects Bayesian Dirichlet Process approach. While both implementations follow the same model, they appear to make distinct predictions about the distribution of selection coefficients. The fixed-effects framework estimates a large proportion of highly deleterious substitutions, whereas the random-effects framework estimates that all substitutions are either nearly neutral or weakly deleterious. It remains unknown, however, how accurately each method infers evolutionary constraints at individual sites. Indeed, selection coefficient distributions pool all site-specific inferences, thereby obscuring a precise assessment of site-specific estimates. Therefore, in this study, we use a simulation-based strategy to determine how accurately each approach recapitulates the selective constraint at individual sites. We find that the fixed-effects approach, despite its extensive parameterization, consistently and accurately estimates site-specific evolutionary constraint. By contrast, the random-effects Bayesian approach systematically underestimates the strength of natural selection, particularly for slowly evolving sites. We also find that, despite the strong differences between their inferred selection coefficient distributions, the fixed- and random-effects approaches yield surprisingly similar inferences of site-specific selective constraint. We conclude that the fixed-effects mutation–selection framework provides the more reliable software platform for model application and future development.

Key words: **mutation–selection models, selection coefficients, protein evolution, dN/dS, sequence simulation, molecular evolution.**

Introduction

Proteins are subject to a variety of structural, functional, and physiochemical constraints that influence their evolutionary trajectories. A growing body of research has demonstrated that these constraints lead individual protein sites to have distinct tolerances to different amino acids (Porto et al. 2004; Ramsey et al. 2011; Pollack et al. 2012; Ashenberg et al. 2013; Bloom 2014a, 2014b; Risso et al. 2014; Abriata et al. 2015; Doud et al. 2015; Echave et al. 2016). Recent experimental studies have further demonstrated that, for at least several proteins, site-wise amino-acid preferences are broadly conserved over evolutionary time (Ashenberg et al. 2013; Risso et al. 2014; Doud et al. 2015).

To achieve a complete picture of protein evolutionary dynamics, it is critical that we employ robust sequence evolution frameworks which explicitly incorporate site-specific amino acid propensities. One such evolutionary model that achieves this goal, known as the mutation–selection model, is an implementation of the classical population genetics Fisher–Wright model (Fisher 1930; Wright 1931) applied to protein-coding sequences. By modeling the joint forces of selection and mutation in protein-coding sequences along a phylogeny, the mutation–selection framework considers site-specific amino-acid and/or codon propensities as its focal parameters (Halpern and Bruno 1998; McCandlish and Stoltzfus 2014). Specifically, the mutation–selection model estimates the scaled fitness, $F = 4N_e f$ (or $F = 2N_e f$ for haploid organisms), where N_e is the effective population size and f represents the Malthusian fitness, of each amino acid at a given position in a protein-coding sequence. These fitnesses are often used to infer the distribution of scaled selection coefficients $S_{ij} = F_j - F_i$, where F_i and F_j are the scaled fitnesses of amino acids i and j. The distribution of S values indicates the range of selective responses to new mutations across a given protein sequence.

Importantly, the term "fitness", as used in the context of mutation–selection models, refers to the overall time-averaged propensities of amino-acids at particular sites, and not necessarily to exact fitness effects incurred by mutations. Indeed, fitness effects at individual sites may fluctuate over

Article

time, e.g., due to epistatic interactions (Weinreich et al. 2006; Pollack et al. 2012; Ashenberg et al. 2013; Draghi and Plotkin 2013; Gong et al. 2013; McCandlish and Stoltzfus 2014; Shah et al. 2015). As such, our use of the word fitness throughout this study should be interpreted primarily as a mutation–selection model parameter indicating conserved site-specific properties.

Recently, two alternative implementations of site-specific mutation–selection models have been released. The first implementation, known as swMutSel, estimates site-specific fitness parameters as fixed-effect variables through a maximum penalized-likelihood (MPL) approach (Tamuri et al. 2012, 2014). The second implementation, available in the PhyloBayes software package, instead employs a Dirichlet Process (DP) Bayesian framework and models site-specific fitness parameters as random effects (Rodrigue et al. 2010; Rodrigue and Lartillot 2014). For simplicity, we will refer to the latter implementation as "pbMutSel" throughout this article. Both platforms are based on the mutation–selection models introduced by Halpern and Bruno (1998) and Yang and Nielsen (2008), and they make nearly identical assumptions about the evolutionary process. For instance, both swMutSel and pbMutSel assume that sites evolve independently, that there is no selection on synonymous codons (i.e., all synonymous codons have the same fitness), and that nucleotide mutation rates are shared across all sites. Furthermore, both frameworks require a fixed phylogeny topology to compute fitness parameters and are not currently suitable for co-estimation of fitnesses and phylogeny.

Although the mutation–selection model provides a promising framework for modeling protein sequence evolution in a mechanistic context, it is not yet clear how one might use its estimates to gain insight into the evolutionary process. Whether the amino-acid fitnesses estimated by either swMutSel or pbMutSel truly reflect evolutionary constraint remains an open question, in particular because these two implementations produce seemingly incompatible results: swMutSel infers S distributions with two peaks representing nearly neutral (centered at $S = 0$) and highly deleterious changes, commonly defined as $S < -10$ in the context of mutation–selection models (Tamuri et al. 2012, 2014; Rodrigue 2013). In contrast, pbMutSel infers unimodal distributions centered at $S = 0$, without a peak of highly deleterious changes.

The relative accuracy between these two distinct approaches has sparked a lively debate in the literature (Rodrigue 2013; Rodrigue and Lartillot 2014; Scheffler et al. 2014; Tamuri et al. 2014). Specifically, Rodrigue (2013) critiqued early swMutSel implementations as suffering from overparameterization, as swMutSel's fixed-effects framework requires estimating 19 parameters per site. He argued that the characteristic peak at $S < -10$ in swMutSel-inferred scaled selection-coefficient distributions is an erroneous artifact of model overparameterization. Rodrigue (2013) additionally contended that, by modeling fitnesses as random effects, pbMutSel avoids overfitting and certain statistical inconsistencies that extensive parameterization might introduce. In response, Tamuri et al. (2014) argued that experimental evidence from population genetics literature supports swMutSel's recovery of a prominent peak of highly deleterious $S < -10$ changes. To ameliorate potential overfitting artifacts, swMutSel has been updated with several likelihood penalty functions that regularize extreme fitness estimates (Tamuri et al. 2014).

Previous quantitative comparisons of swMutSel and pbMutSel inferences have focused nearly exclusively on asking how well they recapitulate the gene-wide distribution of S, or similarly the gene-wide proportions of deleterious and beneficial substitutions (Rodrigue et al. 2010; Tamuri et al. 2012, 2014; Rodrigue 2013; Rodrigue and Lartillot 2014). In spite of these efforts, however, there remains no conclusive evidence supporting either swMutSel or pbMutSel as the more reliable inference approach. Indeed, support for either approach currently rests on theoretical arguments regarding either pbMutSel's more desirable statistical properties or swMutSel's general agreement with population-genetics literature. However, statistical consistency does not necessarily correspond to empirical accuracy, and phylogenetic data may not be directly comparable to population data. As such, neither argument presents strong evidence in favor of either pbMutSel or swMutSel.

We posit that no consensus regarding mutation–selection implementation accuracy has emerged specifically because performance has been assessed using whole-gene S distributions. Pooling all site-specific S values into a single distribution makes it impossible to conduct a systematic analysis of differences between inference methods, especially given that these methods were implemented to estimate amino-acid fitness values at individual sites. As a consequence of this approach, it remains unknown how well inferred parameters capture site-specific evolutionary processes.

Therefore, in this study, we have investigated the relative performance of these two mutation–selection model implementations by directly comparing how well each infers evolutionary constraints at individual sites, rather than focusing primarily on S distributions. We have found that swMutSel, specifically run with a weak likelihood penalty function, consistently estimates the most accurate site-specific fitness values. By contrast, pbMutSel and strongly penalized swMutSel parameterizations systematically underestimate the strength of natural selection across sites, most notably at slowly evolving sites.

Results

Simulation and Inference Approach

We simulated protein-coding sequence alignments wherein each position evolved according to a distinct mutation–selection model parameterization. We ensured that each simulation reflected evolutionary heterogeneity seen in real proteins by deriving simulation parameterizations from two different empirical data sources. The first simulation data set derived codon fitness parameters from site-specific amino acid frequencies in structurally curated natural amino-acid alignments (Ramsey et al. 2011). We obtained site-specific fitness parameters for the second simulation data set using

amino-acid propensities measured experimentally using deep-mutational scanning (DMS) (Bloom 2014a; Firnberg et al. 2014; Stiffler et al. 2014; Thyagarajan and Bloom 2014; Doud et al. 2015; Kitzman et al. 2015). Derivation of simulation parameters is described in depth in *Materials and Methods*. A total of 11 natural alignments and four DMS data sets were used, resulting in a total of 15 alignment parameter sets, as described in table 1. We refer to each alignment simulated using parameters derived from natural alignments as "natural simulations", and similarly to each alignment simulated using DMS-derived parameters as "DMS simulations".

We assumed that all codons for a given amino acid had the same fitness, and we assumed globally equal mutation rates. For each gene, we simulated two alignments, each along a balanced 512-taxon tree, with all branch lengths set equal to either 0.5 or 0.01. We refer to these simulation conditions as BL = 0.5 and BL = 0.01 for simulations with branch lengths of 0.5 and 0.01, respectively. Note that, in the context of mutation–selection models, branch lengths refer to the expected number of neutral substitutions per unit time (Spielman and Wilke 2015b). The BL = 0.5 simulation condition yielded alignments at evolutionary equilibrium, meaning that each simulation should contain sufficient information to discern the underlying stationary amino-acid fitnesses (Spielman et al. 2016). Under the BL = 0.01 condition, on the other hand, simulated sequences will not have diverged enough to reflect their stationary states. Therefore, we expect inferences performed on BL = 0.5 simulations to yield results that are more comparable to true parameters than results from inferences on BL = 0.01 simulations are.

Table 1. Description of Data Sets Used to Derive Simulation Parameters.

Name	Type	Length	Mean Site Entropy	Mean Site *dN*/*dS*
1B4T_A	Natural	115	1.12	0.27
1W7W_B	Natural	125	0.95	0.23
2CFE_A	Natural	151	0.92	0.21
2BCG_Y	Natural	156	1.01	0.26
1G58_B	Natural	165	1.06	0.25
1GV3_A	Natural	176	1.08	0.26
1V9S_B	Natural	177	1.10	0.27
2FLI_A	Natural	190	1.20	0.30
1RII_A	Natural	195	1.13	0.27
1R6M_A	Natural	203	1.13	0.28
1IBS_A	Natural	291	1.07	0.28
Gal4[a]	DMS	64	1.92	0.47
LAC[b]	DMS	262	2.08	0.61
NP[c]	DMS	498	2.38	0.70
HA[d]	DMS	564	2.25	0.63

NOTE.—The *Type* column indicates whether the source of simulation parameters was a natural alignment ("Natural") or deep-mutational scanning data ("DMS"). Natural alignments are named according to their corresponding PDB [Protein Data Bank (Berman et al. 2000)] ID and chain, i.e., the name 1B4T_A corresponds to PBD ID 1B4T, chain A.
[a]Yeast Gal4.
[b]TEM-1 β-lactamase.
[c]Influenza nucleoprotein.
[d]Influenza H1N1 hemgglutinin.

Importantly, natural and DMS simulations featured distinct evolutionary pressures: all natural *S* distributions featured relatively high proportions of strongly deleterious changes ($|S| \geq 10$), whereas all DMS *S* distributions were unimodal, with varying degrees of spread (fig. 1). Similarly, natural simulations contained stringent levels of selective constraint, with most sites under strong purifying selection, but sites in the DMS simulations were subject to moderate-to-weak purifying selection (supplementary fig. S1, Supplementary Material online). We note that the effective population size for DMS experiments is likely smaller than the effective population size in natural settings, partially explaining why these data sets feature weaker selection pressures. That said, all DMS preferences employed here have corrected to account for selection's relatively weaker stringency in DMS experiments, thereby partially ameliorating artifacts caused by differences in population size (Bloom 2014b, 2016).

Findings from previous mutation–selection model studies would suggest that natural simulations would favor the swMutSel platform, which is known to estimate large proportions of deleterious changes, and conversely DMS simulations would favor the pbMutSel platform, which tends to infer strictly unimodal *S* distributions (Rodrigue et al. 2010; Tamuri et al. 2012, 2014; Rodrigue 2013). Therefore, the different features across our simulation sets allowed us to directly contrast how each mutation–selection inference platform behaves on data with realistic levels of evolutionary heterogeneity, without biasing results towards one particular implementation.

We processed each simulated alignment with both swMutSel and pbMutSel. For swMutSel, we processed each alignment both without a penalty and under four penalty

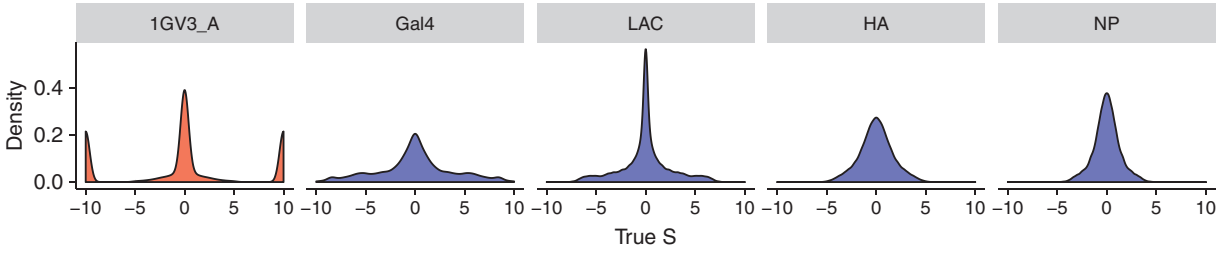

FIG. 1. Distributions of true scaled selection coefficients, S, for a representative natural simulation (1GV3_A) and each DMS simulation. Scaled selection coefficients have been binned at $S \geq 10$ and $S \leq -10$ for visualization. Histograms depicting the true S distribution for all other natural simulations are provided in supplementary figures S8–S10, Supplementary Material online. S distributions shown represent the selection coefficients among all possible single-nucleotide changes, across all sites.

functions (Tamuri et al. 2014). Penalty functions examined included the multivariate normal penalty function with the σ^2 parameter equal to either 10 or 100 (referred to as mvn10 and mvn100, respectively), as well as the Dirichlet-based penalty function with the α parameter equal to either 0.1 or 0.01 (referred to as d0.1 and d0.01, respectively). Each set of penalty-function parameterizations represents stronger to weaker penalties, i.e., mvn10 and d0.1 are stronger penalties than are mvn100 and d0.01, respectively. We refer to each swMutSel inference using its respective penalty specification and to each swMutSel inference without a penalty function as "unpenalized".

Distance between True and Inferred Parameters Depends on Method and Data set

We first assessed how the inferred site-specific fitness values compared with the true fitness values. We derived, for each site-specific set of inferred fitnesses, the corresponding equilibrium amino-acid frequencies (Sella and Hirsh 2005; Spielman and Wilke 2015a). We calculated the Jensen–Shannon distance (JSD) between the inferred and true equilibrium frequency distributions. JSD is defined as

$$\mathrm{JSD}(P,Q) = \sqrt{\frac{D(P,M) + D(Q,M)}{2}}, \qquad (1)$$

where $P = (p_1, \ldots, p_{20})$ and $Q = (q_1, \ldots, q_{20})$ are the amino-acid frequency distributions to be compared, $M = (P+Q)/2$ is the element-wise average between P and Q, and $D(A,B) = \sum_i a_i \ln(a_i/b_i)$ is the Kullback–Leibler divergence between distributions $A = (a_1, \ldots, a_{20})$ and $B = (b_1, \ldots, b_{20})$. JSD values range from 0 for completely identical distributions to 1 for completely dissimilar distributions.

Across all data sets, JSD values were 1.5–2 times larger for BL = 0.5 than for BL = 0.01 simulations (fig. 2). This result reflects that the true amino-acid frequencies represent those present under an evolutionary equilibrium, which is not reached under such short time scales of 0.01 branch lengths.

Trends for results from natural simulations were consistent between branch-length conditions: unpenalized swMutSel and multivariate normal penalties displayed the lowest JSD values, of roughly 0.15 on an average. In fact, their JSD distributions were statistically indistinguishable for a given data set ($P > 0.99$, mixed-effects linear model). Under Dirichlet penalities and pbMutSel, JSD for natural simulations sharply increased, with pbMutsel universally showing the highest JSD.

By contrast, DMS simulations showed different trends between branch length conditions. For BL = 0.5, DMS simulations had low mean JSD values under unpenalized swMutSel and multivariate normal penalties, but their JSD values either slightly decreased or remained unchanged under swMutSel Dirichlet penalties and pbMutSel. Moreover, Gal4 consistently showed larger JSD values across unpenalized and multivariate normal swMutSel penalties, and it further showed increased JSD under pbMutSel, similar to natural simulations. This outlying result may be due to shorter length of Gal4 compared with the other simulated genes (table 1). The JSD values for DMS and natural simulations were most comparable under the d0.01 penalty in swMutSel (fig. 2A), suggesting that this parameterization may be least sensitive to selection pressures in the data. The DMS simulations under BL = 0.01, however, displayed the opposite trends from BL = 0.5: JSD was decreased from unpenalized swMutSel to reach its lowest values under pbMutSel. Again, the Gal4 DMS simulation yielded an increased JSD for pbMutSel.

Why were JSD results consistent between branch length conditions for natural simulations but not for DMS simulations? We suggest that this finding directly resulted from different selective constraints operating between data sets. First, note that careful statistical analysis on swMutSel and pbMutSel has shown that swMutSel "considers unobserved amino acids as highly deleterious", whereas pbMutSel is "less conclusive in this regard" (Rodrigue 2013). The DMS data sets featured far weaker selective pressure across sites, meaning that more amino acids were selectively permitted per site. However, under the short time scale of BL = 0.01, relatively few substitutions will have occurred. DMS sites will therefore appear far more conserved than they truly are under stationarity. By contrast, the relatively higher selective constraint in natural simulations means that fewer amino acids will be tolerated per site. While the full stationary distribution of states will still not be reached at BL = 0.01, enough changes

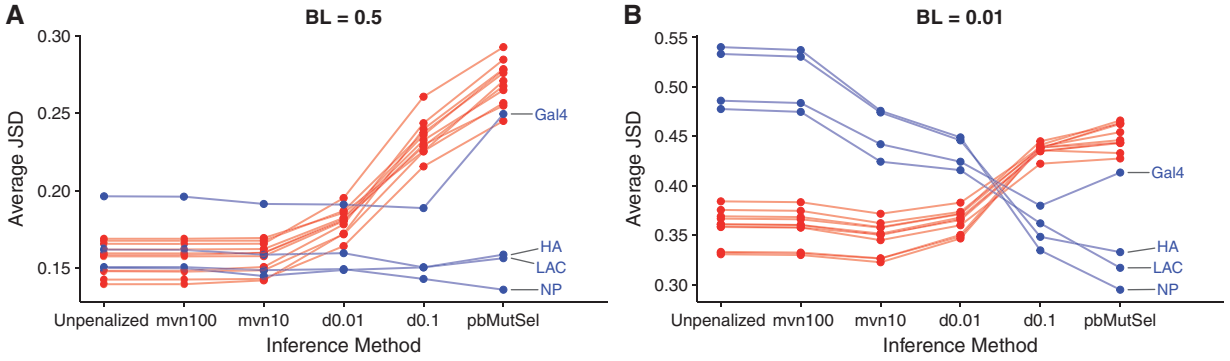

FIG. 2. Jensen–Shannon distance between true and inferred amino-acid frequency distributions. (A) BL = 0.5 simulations. (B) BL = 0.01 simulations. Each point represents the mean JSD across sites for a given simulation, and labeled points correspond to DMS simulations. Note that the *y*-axis of each panel has a different range.

will likely have occurred to reveal the most fit amino acids, leading to lower JSD at low divergence.

We further considered that JSD might be an overly sensitive metric wherein distances between small values will appear inflated. For example, comparing frequencies of 10^{-4} and 10^{-5} may yield fairly large JSD, but in fact these frequencies are comparable in terms of evolutionary pressures. To test this possibility, we additionally calculated the sum of absolute differences of site-specific frequencies, and we recovered broadly the same trends as for the JSD analysis (supplementary fig. S2, Supplementary Material online), and thus JSD results did not appear to be artifactual.

Extensively Parameterized Models Best Infer Evolutionary Constraint

Importantly, JSD is not an explicit evolutionary measure. For instance, while a large JSD indicates high dissimilarity, it is neither possible to tell how this dissimilarity relates to selection pressure nor whether high JSD corresponds to systematically biased or randomly distributed error in estimates. Furthermore, distance metrics like JSD may obscure the true evolutionary constraint. Consider an amino-acid whose presence is not tolerated at a given site: Whether this amino-acid has an associated scaled selection coefficient of −100 or −200 amounts to the same evolutionary pressure, although its JSD may instead be quite large.

Therefore, we next asked whether site-specific inferences from swMutSel and pbMutSel corresponded to the true selective constraint at each site. We measured selective constraint at individual sites using two metrics: predicted dN/dS and Shannon entropy H. We predicted a dN/dS rate ratio for each site's set of mutation–selection parameters as described in Spielman and Wilke (2015a, 2015b) [see also dos Reis (2015) for an alternative method of dN/dS calculation]. The predicted dN/dS value indicates the expected substitution-rate ratio under evolutionary equilibrium. Further, because our simulations assumed symmetric nucleotide mutation rates and no codon bias, all true dN/dS ratios are constrained to $dN/dS \in [0, 1]$ (Spielman and Wilke 2015a).

Entropy is calculated for a given alignment column as

$$H = -\sum_i P_i \ln P_i, \qquad (2)$$

where P_i is the frequency of amino-acid i, and the sum runs over all 20 amino-acids. Entropy is bounded by $H \in [0, 3.0]$, and the value $3.0 = \ln(1/20)$ indicates that each amino acid is equally frequent. Both metrics have clear, widely accepted interpretations: Lower values indicate stronger selective constraint, and higher values indicate progressively weaker constraint. Moreover, dN/dS uniquely provides an evolutionarily aware summary statistic for the selection pressure acting at a given site. While entropy calculations consider only amino-acid frequencies, dN/dS is calculated directly from substitution rates between codons. As such, dN/dS is geared more specifically towards evolutionary analysis than is entropy.

We calculated site-specific dN/dS and entropy for each true and inferred distribution of site-specific amino-acid fitnesses and nucleotide mutation rates (Spielman and Wilke 2015a), and we compared the resulting true and predicted values across inference methods (figs. 3 and 4 and supplementary figs. S3–S7, Supplementary Material online). Specifically, we measured r^2 between inferred and true parameters, the estimator bias of each inference method, and finally the slope of the linear relationship between inferred and true parameters. r^2 indicates the percent of variance in the true parameters explained by inferred parameters, estimator bias indicates whether an inference method tends to overestimate or underestimate true parameters, and the slope indicates whether an inference method tends to overestimate (slope > 1) or underestimate (slope < 1) larger parameters relative to smaller parameters. Note that we performed hypothesis tests on the slope using the null hypothesis of slope equal to 1, rather than the more traditional null of slope equal to 0, to test specifically for this deviation.

Unlike JSD, the results for dN/dS and H comparisons showed similar trends across data sets and branch-length conditions. For BL = 0.5 simulations, we found excellent agreement between true and predicted quantities for natural, LAC, and Gal4 simulations when run with unpenalized swMutSel, mvn100, mvn10, and d0.01 (figs. 3 and 4*A* and *B*). However, unpenalized swMutSel, mvn100, and mvn10 tended to slightly underestimate dN/dS and entropy, i.e., overestimate selective constraint. On the other hand, d0.01 mostly showed no estimator bias for natural simulations, and finally d0.1 and pbMutSel overestimated dN/dS and entropy (figs. 3 and 4*C* and *D*). The NP and HA DMS simulations yielded similar patterns to the other simulations, although these simulations were associated with generally lower r^2 values. Further, the estimator bias for the NP simulation, under pbMutSel inference as measured using entropy, was not statistically significant (Bonferroni-corrected $P > 0.05$). Finally, no true-inferred slopes, for either dN/dS or entropy, showed a statistically significant deviation from 1 (Bonferroni-corrected $P > 0.05$) for BL = 0.5 (fig. 4*E* and *F*).

Unexpectedly, even though DMS simulations (particularly NP and HA) featured unimodal selection coefficient distributions which we suspected would be more suited to pbMutSel analysis, weakly penalized swMutSel in fact gave the best performance across all data sets. In addition, all metrics considered here showed that unpenalized swMutSel in fact outperformed pbMutSel for DMS simulations, in spite of the paucity of highly deleterious amino acids in these simulations. pbMutSel, and to a lesser extent d0.1, systematically overestimated dN/dS and entropy, thereby inferring much weaker selection pressure than was truly present. This trend was pronounced for highly constrained, i.e., low dN/dS, sites, explaining why estimator bias was generally larger for natural simulations. Indeed, all sites in the NP and HA DMS simulations had $dN/dS \geq 0.24$, and roughly 90% of sites in Gal4 and LAC simulations had $dN/dS \geq 0.2$ (supplementary fig. S1, Supplementary Material online). As such, a small minority of sites, if any, were subject to very strong purifying selection in DMS simulations. By contrast, between 40% and 60% of sites in natural simulations had $dN/dS \leq 0.2$ (supplementary fig. S1, Supplementary Material online), and hence

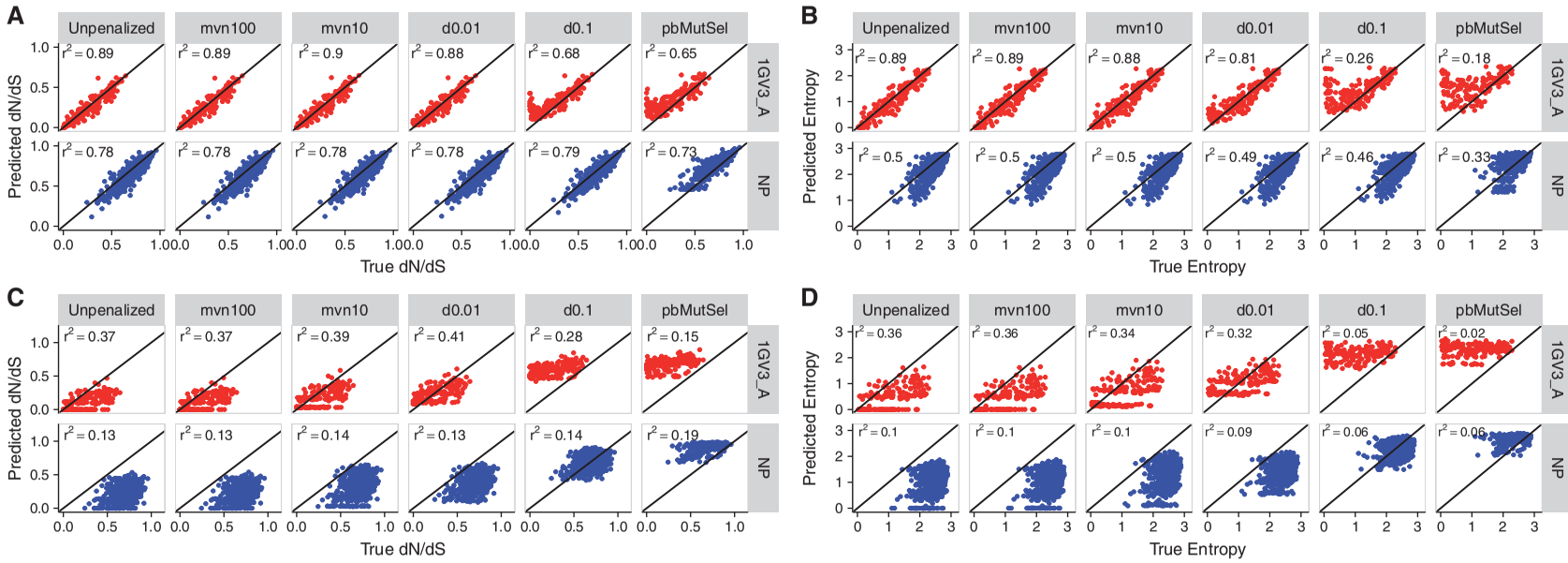

Fig. 3. Performance of mutation–selection model inference platforms for one representative natural (1GV3_A) and DMS (NP) simulation each. (*A*) *dN*/*dS* predicted from model inference versus true *dN*/*dS*, under branch lengths of 0.5. (*B*) Entropy calculated from model inferences versus true entropy, under branch lengths of 0.5. (*C*) *dN*/*dS* predicted from model inference versus true *dN*/*dS*, under branch lengths of 0.01. (*D*) Entropy calculated from model inferences versus true entropy, under branch lengths of 0.01. In each panel, the straight line indicates the $y = x$ line. Scatterplots for all simulated data sets are provided in supplementary figures S3–S6, Supplementary Material online.

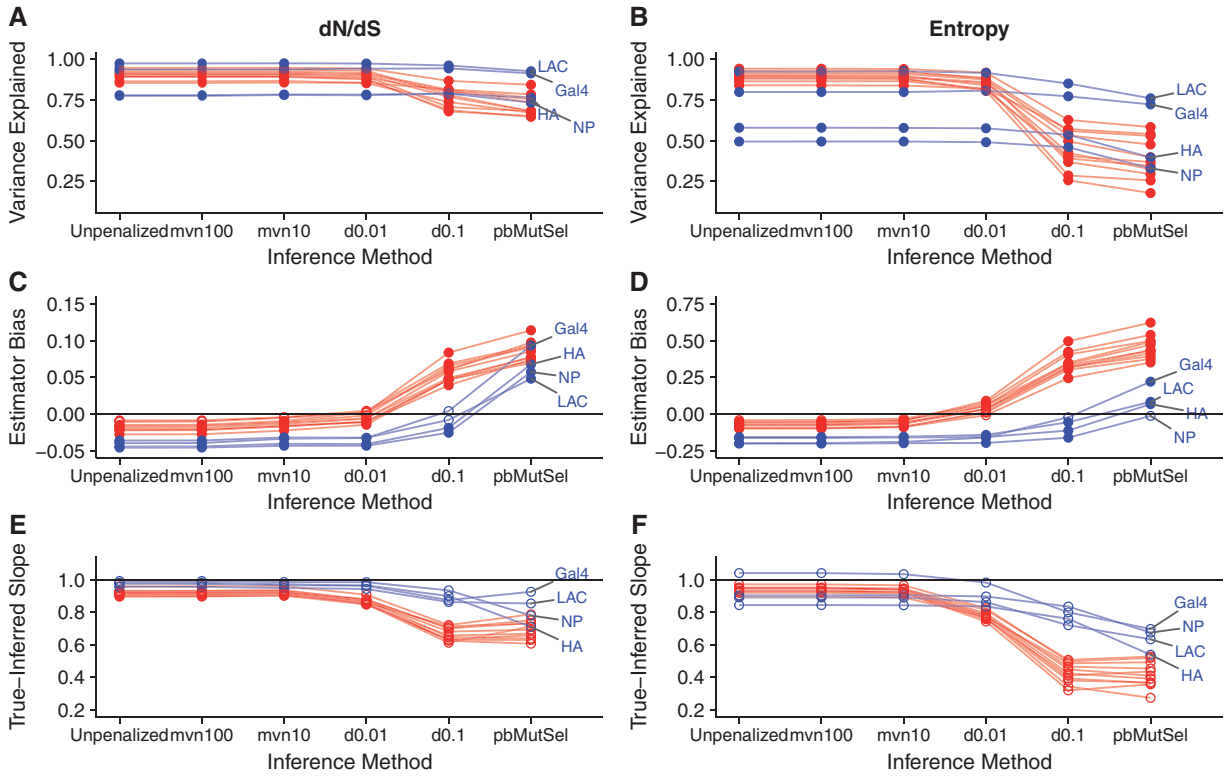

Fig. 4. Performance of mutation–selection model inference platforms on simulations with branch lengths of 0.5. Labeled points correspond to DMS simulations. (*A* and *B*) r^2 between true and inferred *dN*/*dS* (*A*) and entropy (*B*) across inference methods, for all simulated data sets. (*C* and *D*) Estimator bias of inference methods relative to true *dN*/*dS* (*C*) and entropy (*D*) values, for all simulated data sets. Open points indicate biases that were not significantly different from 0 (Bonferroni-corrected $P > 0.05$, test for intercept in linear model), and solid points indicate biases that were significantly different from 0 (Bonferroni-corrected $P < 0.05$). The straight line indicates an estimator bias of 0, meaning an unbiased predictor. Note that panels (*C* and *D*) use different *y*-axis ranges, due to the different scales between *dN*/*dS* and entropy. (*E* and *F*) Slope for the linear relationship of inferred regressed on true *dN*/*dS* (*E*) and entropy (*F*) values. Open points indicate slopes that were not significantly different from 1 (Bonferroni-corrected $P > 0.05$, test for slope in linear model not equal to 1), and solid points indicate biases that were significantly different from 1 (Bonferroni-corrected $P < 0.05$). The straight line indicates the null slope of 1. A corresponding figure for simulations using branch lengths of 0.01 is in supplementary figure S7, Supplementary Material online.

overestimation by d0.1 and pbMutSel was more apparent for natural simulations.

For the BL = 0.01 simulations, all methods performed poorly, likely because sequences did not attain the evolutionary equilibrium reflected by the true parameter values (fig. 3*C* and *D* and supplementary figs. S4, S6, and S7, Supplementary Material online). Specifically, unpenalized swMutSel, mvn100, and mvn10 strongly underestimated *dN*/*dS* and entropy, meaning that they inferred far more stringent evolutionary constraint than existed. Further, d0.01 showed the least estimator bias for natural simulations, and d0.1 showed the least estimator bias for DMS simulations, likely resulting from the different selection pressures between simulation sets. pbMutSel greatly overestimated *dN*/*dS* and entropy, often to the point where virtually no relationship existed between true and inferred metrics. These results suggest that mutation–selection models might be unreliable for analyzing data sets with low divergence. Even so, the overall patterns observed for r^2, estimator bias, and slope were consistent between branch length conditions, implying that *dN*/*dS* and entropy, moreso than JSD, served as robust indicators of mutation–selection model performance.

Causes of Site-Specific Inference Error across Methods

We next asked whether a given site's underlying selective constraint, as represented by the true site-specific *dN*/*dS*, influenced error in the inferred fitness values, as represented by site-specific JSD. In other words, we examined whether the selection pressure at individual sites biased fitness inferences within a given gene. Given the broad comparability between *dN*/*dS* and entropy metrics (fig. 4), we considered only the more evolutionarily aware *dN*/*dS*. In addition, we studied only the BL = 0.5 simulations.

We regressed site-specific JSD against *dN*/*dS*, and we analyzed the slope of each regression (fig. 5*A* and *B*). For natural simulations, unpenalized swMutSel, mvn100, mvn10, and d0.01 JSD increased with decreasing selection pressure, i.e., increasing *dN*/*dS*, as indicated by positive slopes. However, 5 of the 11 natural data sets yielded slopes that did not significantly differ from 0 (Bonferroni-corrected $P > 0.05$) when run with d0.01, suggesting that this swMutSel parameterization may be less biased by selection pressures. By contrast, d0.1 and pbMutSel displayed the opposite trend from the other approaches: JSD was lowest for these approaches at sites with weak selective constraint, i.e., high *dN*/*dS*.

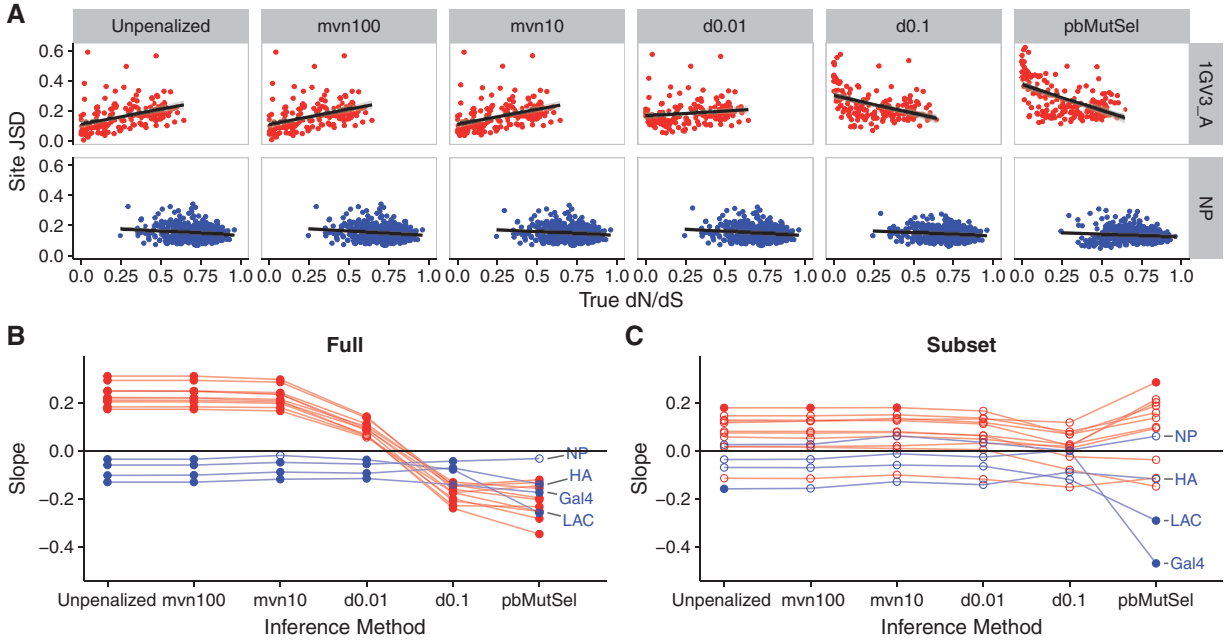

FIG. 5. The site-specific Jensen–Shannon distance between true and inferred amino-acid frequencies depends both on selective constraint and inference method. Results are shown for simulations with branch lengths of 0.5. Labeled points correspond to DMS simulations. (*A*) Site JSD regressed on true site dN/dS. The line in each panel indicates the linear regression line. (*B*) Slope of relationship shown in panel (*A*) for all simulated data sets. (*C*) Slope of relationship shown in panel (*A*) for all simulated data sets, considering only a subset of sites whose true dN/dS falls in the range $dN/dS \in [0.3, 0.6]$. For panels (*B*) and (*C*), the straight line indicates the $y = 0$ line, meaning no linear relationship between JSD and dN/dS. Open points indicate slopes that were not significantly different from 0 (Bonferroni-corrected $P > 0.05$), and solid points indicate slopes that were significantly different from 0 (Bonferroni-corrected $P < 0.05$).

For DMS simulations, on the other hand, all slopes were weakly negative (fig. 5*A* and *B*), meaning that all inference approaches yielded more precise fitness estimates for sites with weaker selection pressure. Moreover, many of the slopes for DMS simulation comparisons were not statistically different from 0 (fig. 5*B*), namely when run with mvn10 and d0.01. Therefore, fitness estimates made by the d0.01 swMutSel parameterization were least influenced by underlying site-specific selection pressure across both natural and DMS data sets.

We hypothesized that the source of discrepancy between natural and DMS simulations (fig. 5*A* and *B*), could be traced back to the different selective landscapes between data sets. We therefore again regressed site-specific JSD on true dN/dS, but using only a subset of each data set so that each gene had fully comparable distributions of selective constraint. In particular, for each regression, we included only sites whose true dN/dS was in the range $0.3 \leq dN/dS \leq 0.6$. This analysis indeed showed that nearly all slopes were not significantly different from zero (Bonferroni-corrected $P > 0.05$, fig. 5*C*). Thus, it appeared that swMutSel had specific difficulty estimating fitnesses at sites with low selective constraint, and conversely pbMutSel had specific difficulty estimating fitnesses at sites with high selective constraint. The platforms performed comparably, in terms of site-specific error, for sites subject to moderate purifying selection.

Inferred Selection Coefficient Distributions Depend on Method, Not on Data set

Previously, it has been an open question whether observed features of inferred S distributions, namely the presence of large proportions of deleterious changes, were primarily caused by the data being analyzed or instead by the statistical properties of the specific inference approach applied (Rodrigue 2013). We therefore next asked whether comparing true and inferred S distributions revealed similar patterns about methodological performance as dN/dS and entropy comparisons did.

In fact, we found instead that the inference approach, not the underlying data set, seemed to predict the shape of the inferred S distribution (fig. 6 and supplementary figs. S8–S10, Supplementary Material online). For example, across all DMS simulations, pbMutSel estimated S distributions that were most similar to the true S distributions, yet for natural simulations, S distributions estimated by unpenalized swMutSel most resembled the true distributions. Our analysis of site-specific selective constraint with dN/dS and entropy, however, did not find that either of these two approaches inferred the most reliable selection pressures. Instead, unpenalized swMutSel tended to underestimate dN/dS entropy, and conversely pbMutSel substantially overestimated these quantities (fig. 4 and supplementary fig. S7, Supplementary Material online).

Discussion

We have investigated the utility of mutation–selection model inference platforms for inferring site-specific selective constraints from coding sequences. We found that swMutSel, run specifically with a weak-to-moderate Dirichlet penalty function, consistently inferred site-specific fitness values that reliably captured each site's evolutionary constraint, as

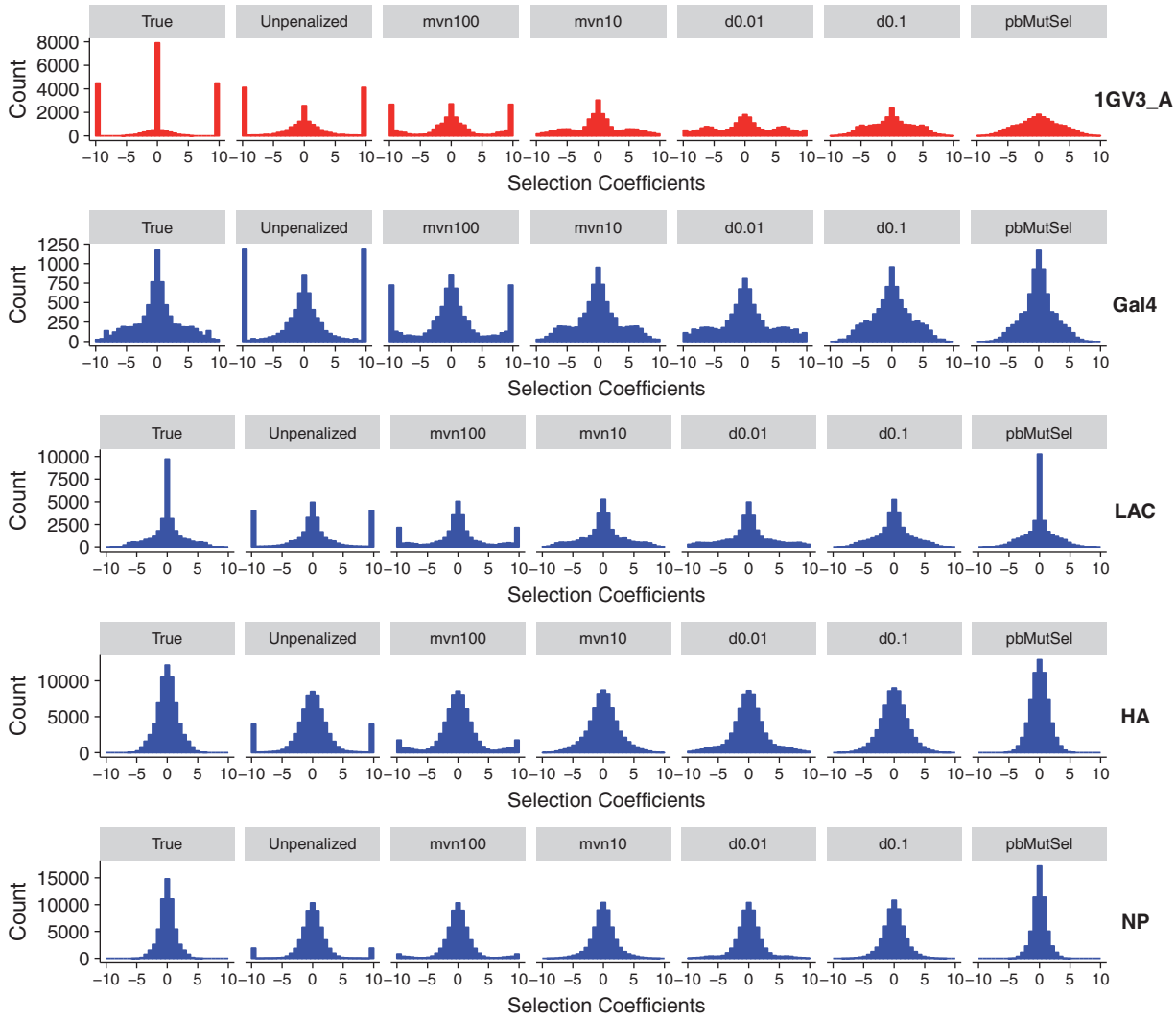

Fig. 6. True and inferred distributions of scaled selection coefficients for a subset of simulations, under branch lengths of 0.5. Histograms for simulations not shown here are in supplementary figures S8–S10, Supplementary Material online. *S* distributions shown represent the selection coefficients among all possible single-nucleotide changes, across all sites.

represented by *dN*/*dS* and entropy. pbMutSel, as well as swMutSel run with a strong Dirichlet penalty function, systematically underestimated the strength of natural selection across sites. In addition, swMutSel multivariate normal penalties estimated fitness values that were nearly identical to unpenalized swMutSel, suggesting that these penalties may not substantially reduce overparameterization. Importantly, our results were robust to the proportion of deleterious changes in the data: d0.1 swMutSel appeared most suited for genes with moderate-to-weak purifying selection (i.e., unimodal *S* distributions), and d0.01 swMutSel was the best performing method for genes subject to strong purifying selection (i.e., *S* distributions with large proportions of deleterious changes). We therefore recommend selecting one of these swMutSel parameterizations for data analysis, depending on the strength of natural selection suspected to act on the gene being analyzed.

Rather than focusing our analysis on *S* distributions, we instead analyzed mutation–selection inferences on a site-specific basis using site entropy as well as the evolutionarily meaningful summary statistic *dN*/*dS*. This strategy allowed for a more fine-grained analysis of inferred parameters compared with whole-gene *S* distributions that can obscure site-specific evolutionary processes. Furthermore, our approach highlighted a considerable disconnect between *S* distributions and site-specific evolutionary constraint: The mutation–selection implementation that provided the best *S* estimates did not necessarily provide the best estimates of site-specific selection pressure, and vice versa. Instead, inferred *S* distributions appeared to be driven primarily by the inference method applied and not by features of the data set. This discordance reveals why previous studies focusing almost exclusively on *S* as a litmus test to compare performance of swMutSel and pbMutSel have been unable to reach a consensus.

We additionally emphasize that, while weakly penalized swMutSel emerged here as the more reliable mutation–selection inference platform, *dN*/*dS* ratios and entropy predicted from all inferences showed strong relationships with their corresponding true parameters (figs. 3 and 4), and indeed with one another. For example, *dN*/*dS* and entropy predicted

from unpenalized swMutSel and pbMutSel were, on an average, correlated with $r^2 = 0.81$ and $r^2 = 0.56$, respectively, across all BL = 0.5 simulations. These high correlations contrast with conclusions drawn from previous studies that swMutSel and pbMutSel make fundamentally distinct, even incompatible, inferences. Therefore, while performance differences between swMutSel and pbMutSel were clearly present, they were smaller than one might assume based on S distributions alone.

Moreover, the larger r^2 associated with dN/dS, compared with entropy, suggests that entropy is a much more sensitive measurement, specifically in terms of selection pressure. For example, consider a given amino acid whose stationary frequency is estimated by different platforms as 10^{-6} and 10^{-8}. In evolutionary terms, these frequencies amount to virtually the same result: Natural selection strongly disfavors this amino acid, which is not likely to fix if it arises by mutation. dN/dS calculations will recognize the similar consequences of these frequencies and yield similar values. By contrast, entropy calculations will be much more sensitive to the two-order of magnitude difference in frequencies. For this reason, the r^2 between unpenalized swMutSel and pbMutSel was higher for dN/dS than for entropy.

We suggest that some modifications to pbMutSel's default settings, such as changing the fixed dispersion parameter for its Dirichlet prior, may produce more reliable inferences. Although such efforts may be helpful, there remained salient differences in runtime between swMutSel and pbMutSel. For example, each swMutSel inference required between 6 and 72 h to converge (with unpenalized swMutSel inferences on the longer HA and NP DMS simulations taking the most time), whereas each pbMutSel inference required between 1 and 3 weeks. In other words, each swMutSel inference converged nearly 10 times more quickly than did each pbMutSel inference. From a practical standpoint, swMutSel's relatively short runtime and reliable inferences make it the preferred inference platform. We therefore recommend the use of swMutSel with a weak (d0.01) Dirichlet penalty for highly constrained genes or with a moderate (d0.1) Dirichlet penalty for more weakly constrained genes.

Materials and Methods

Generation of Simulated Data

Sequences were simulated according to the mutation–selection model in Halpern and Bruno (1998), which assumes a reversible Markov model of sequence evolution. For each site k, this model's rate matrix is given by

$$q_{ij}^{(k)} = \begin{cases} \mu_{ij} u_{ij}^{(k)} & \text{single nucleotide change} \\ 0 & \text{multiple nucleotide changes} \end{cases}, \quad (3)$$

where μ_{ij} is the site-invariant mutation rate between codons i and j, and $u_{ij}^{(k)}$, the site-specific relative fixation probability from codon i to j, is defined as

$$u_{ij}^{(k)} = \frac{S_{ij}^{(k)}}{1 - e^{-S_{ji}^{(k)}}}, \quad (4)$$

where $S_{ij}^{(k)}$ is the scaled selection coefficient from codon i to j at site k (Halpern and Bruno 1998). Note that $u_{ij}^{(k)}$ can also be expressed as

$$u_{ij}^{(k)} = \ln\left(\frac{\pi_j^{(k)}\mu_{ij}}{\pi_i^{(k)}\mu_{ji}}\right) \Big/ \left(1 - \frac{\pi_i^{(k)}\mu_{ji}}{\pi_j^{(k)}\mu_{ij}}\right), \quad (5)$$

where $\pi_i^{(k)}$ is the equilibrium frequency of codon i at site k (Halpern and Bruno 1998; Spielman and Wilke 2015a).

For all simulations, we specified equal mutation rates, $\mu_{ij} = \mu = \text{const}$. We determined each alignment's site-specific codon frequencies from two sources. First, we used a set of structurally curated natural amino-acid alignments, with each sequence homologous to a given PDB structure, compiled by Ramsey et al. (2011). For each of those alignments that contained at least 150 taxa, we calculated each site's amino acid frequencies, which we converted to codon frequencies under the assumption that all synonymous codons for a given amino acid had the same frequency. In addition, sites which contained fewer than 150 amino acids (e.g., a column in an alignment with 200 taxa but half of whose characters are gaps) were discarded. A total of 11 natural alignments, with a number of codon positions ranging from 115 to 291, remained after this procedure. We additionally set the equilibrium frequency of all unobserved amino acids to 10^{-9}.

Second, we used four sets of experimentally determined amino acid propensities from deep-mutational scanning (DMS) experiments. The genes used were influenza H1N1 hemagglutinin (Thyagarajan and Bloom 2014), influenza nucleoprotein (Bloom 2014a; Doud et al. 2015), TEM-1 β-lactamase (Firnberg et al. 2014; Stiffler et al. 2014), and yeast Gal4 (Kitzman et al. 2015). We specifically used scaled experimental amino-acid propensities, as given by and described in Bloom (2016). Because we simulated all alignments with symmetric nucleotide mutation rates, the amino-acid propensities obtained from DMS experiments were equivalent to stationary amino-acid frequencies (Sella and Hirsh 2005; Bloom 2016), which we used for simulation.

For all derived codon frequency parameters, we computed codon fitness parameters to calculate selection coefficient distributions, where $F_i = \log(\pi_i)$ for a given codon i (Sella and Hirsh 2005). This relationship holds specifically in the presence of symmetric mutation rates. Using the resulting fitness parameters and equal mutation rates, we then simulated an alignment corresponding to each of the 11 natural alignments and four DMS profiles using Pyvolve (Spielman and Wilke 2015b). We conducted all simulations along a 512-taxon balanced tree with all branch lengths equal to either 0.5 or 0.01, yielding a total of 30 simulated alignments.

Mutation–Selection Model Inference

We processed all alignments, both simulated and empirical, with swMutSel v1.6 (Tamuri et al. 2014) and pbMutSel, specifically, PhyloBayes-MPI v1.5a (Rodrigue and Lartillot 2014). swMutSel inference was carried out under five specifications, including without the use of a penalty function, and two parameterizations each for both the multivariate normal

and the Dirichlet penalty functions. For the multivariate normal penalty, we set σ^2 to either 10 or 100, and for the Dirichlet penalty, we set α to either 0.1 or 0.01.

For inference with pbMutSel, we followed the inference approach given in Rodrigue (2013). We ran each chain for 5500 iterations, saving every five cycles until a total sample size of 1100 was obtained. The first 100 samples were discarded as burnin, and hence the final posterior distribution from which fitnesses were calculated contained 1000 MCMC draws. Convergence was assessed visually using Tracer (Rambaut et al. 2014). Note that for inferences on NP and HA simulations we saved every three, rather than five, cycles for computational tractability.

We further note that we computed *dN*/*dS* and entropy from the posterior mean of all MCMC cycles for a given inference. An alternative approach might instead compute these quantities for each MCMC cycle, and finally, average these quantities across all draws. However, this procedure is not currently possible with the PhyloBayes software.

Statistical Analysis and Data Availability

All statistical analyses were conducted in the R programming language (R Core Team 2015). All statistical tests were performed with a significance value of $\alpha = 0.05$, with correction for multiple testing using the Bonferroni correction. Simulated data, statistical analyses, and all code used are freely available from the github repository https://github.com/sjspielman/mutsel_benchmark, last accessed August 16, 2016.

Supplementary Material

Supplementary figures S1–S10 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported in part by NIH grant F31 GM113622-01 to S.J.S., NIH grant R01 GM088344 to C.O.W., ARO grant W911NF-12-1-0390 to C.O.W., and NSF Cooperative Agreement No. DBI-0939454 (BEACON Center) to C.O.W. Computational resources were provided by the University of Texas at Austin's Center for Computational Biology and Bioinformatics (CCBB).

References

Abriata LA, Palzkill T, Dal Peraro M. 2015. How structural and physicochemical determinants shape sequence constraints in a functional enzyme. *PLoS One* 10:e0118684–e0118615.

Ashenberg O, Gong LI, Bloom JD. 2013. Mutational effects on stability are largely conserved during protein evolution. *Proc Natl Acad Sci U S A*. 110:21071–21076.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. *Nucleic Acids Res*. 28:235–242.

Bloom JD. 2014a. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Mol Biol Evol*. 31:1956–1978.

Bloom JD. 2014b. An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs. *Mol Biol Evol*. 31:2753–2769.

Bloom JD. 2016. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. bioRxiv doi:10.1101/037689.

dos Reis M. 2015. How to calculate the non-synonymous to synonymous rate ratio of protein-coding genes under the Fisher-Wright mutation-selection framework. *Biol Lett*. 11:20141031.

Doud MB, Ashenberg O, Bloom JD. 2015. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. *Mol Biol Evol*. 32:2944–2960.

Draghi JA, Plotkin JB. 2013. Selection biases the prevalence and types of epistatis along adaptive trajectories. *Evolution* 67:3120–3131.

Echave J, Spielman SJ, Wilke CO. 2016. Causes of evolutionary rate variation among protein sites. *Nat Rev Genet*. 17:109–121.

Firnberg E, Labonte JW, Gray JJ, Ostermeier M. 2014. A comprehensive, high-resolution map of a genes fitness landscape. *Mol Biol Evol*. 31:1581–1592.

Fisher RA. 1930. The genetical theory of natural selection. Oxford: Oxford University Press Pub. Co.

Gong LI, Suchard MA, Bloom JD. 2013. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife* 2:e00631.

Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol*. 15:910–917.

Kitzman JO, Starita LM, Lo R, Fields S, Shendure J. 2015. Massively parallel single-amino-acid mutagenesis. *Nat Methods* 12:203–206.

McCandlish DM, Stoltzfus A. 2014. Modeling evolution using the probability of fixation: history and implications. *Q Rev Biol*. 89:225–252.

Pollack DD, Thiltgen G, Goldstein RA. 2012. Amino acid coevolution induces an evolutionary Stokes shift. *Proc Natl Acad Sci U S A*. 109:E1352–E1359.

Porto M, Roman HE, Vendruscolo M, Bastolla U. 2004. Prediction of site-specific amino acid distributions and limits of divergent evolutionary changes in protein sequences. *Mol Biol Evol*. 22:630–638.

R Core Team. 2015. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

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.

Risso VA, Manssour-Triedo F, Delgado-Delgado A, Arco R, Barroso-delJesus A, Ingles-Prieto A, Godoy-Ruiz R, Gavira JA, Gaucher EA, Ibarra-Molero B, Sanchez-Ruiz JM. 2014. Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. *Mol Biol Evol*. 32:440–455.

Rodrigue N. 2013. On the statistical interpretation of site-specific variables in phylogeny-based substitution models. *Genetics* 193:557–564.

Rodrigue N, Lartillot N. 2014. Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. *Bioinformatics* 30:1020–1021.

Rodrigue N, Philippe H, Lartillot N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc Natl Acad Sci U S A*. 107:4629–4634.

Scheffler K, Murrell B, Kosakovsky Pond SL. 2014. On the validity of evolutionary models with site-specific parameters. *PLoS One* 9:e94534.

Sella G, Hirsh AE. 2005. The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci U S A*. 102:9541–9546.

Shah P, McCandlish DM, Plotkin JB. 2015. Contingency and entrenchment in protein evolution under purifying selection. *Proc Natl Acad Sci U S A*. 112:E3226–E3235.

Spielman SJ, Wilke CO. 2015a. The relationship between *dN*/*dS* and scaled selection coefficients. *Mol Biol Evol*. 32:1097–1108.

Spielman SJ, Wilke CO. 2015b. Pyvolve: a flexible python module for simulating sequences along phylogenies. *PLoS One* 10:e0139047.

Spielman SJ, Wan S, Wilke CO. Forthcoming 2016. A comparison of one-rate vs. two-rate inference frameworks for site-specific *dN*/*dS* estimation. Genetics.

Stiffler MA, Hekstra DR, Ranganathan R. 2014. Evolvability as a function of purifying selection in TEM-1 β-lactamase. *Cell* 160:882–892.

Tamuri AU, dos Reis M, Goldstein RA. 2012. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics* 190:1101–1115.

Tamuri AU, Goldman N, dos Reis M. 2014. A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. *Genetics* 197:257–271.

Thyagarajan B, Bloom JD. 2014. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife* 3:e03300.

Weinreich DM, Delaney ND, DePristo MA, Hartl DL. 2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* 312:111–114.

Wright S. 1931. Evolution in Mendelian populations. *Genetics* 16:97–159.

Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol*. 25:568–579.