Looking for Darwin in genomic sequences: validity and success depends on the relationship between model and data

C. T. Jones, Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia

E. Susko, Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia

J. P. Bielawski, Department of Biology, Dalhousie University, Halifax, Nova Scotia

Abstract

Codon substitution models (CSMs) are commonly used to infer the history of natural section for a set of protein coding sequences, often with the explicit goal of detecting the signature of positive Darwinian selection. However, the validity and success of CSMs used in conjunction with the maximum likelihood (ML) framework is sometimes challenged with claims that the approach might too often support false conclusions. In this chapter we use a case study approach to identify four legitimate statistical difficulties associated with inference of evolutionary events using CSMs. These include (1) model misspecification; (2) low information content; (3) the confounding of processes, and (4) phenomenological load, or PL. While past criticisms of CSMs can be connected to these issues, the historical critiques were often misdirected, or over-stated, because they failed to recognize that the success of any model-based approach depends on the relationship between model and data. Here we explore this relationship and provide a candid assessment of the limitations of CSMs to extract historical information from extant sequences. To aid in this assessment we provide a brief overview of (1) a more realistic way of thinking about the process of codon evolution framed in terms of population genetic parameters, and (2) a novel presentation of the ML statistical framework. We then divide the development of CSMs into two broad phases of scientific activity, and show that the latter phase is characterized by increases in model complexity that can sometimes negatively impact inference of evolutionary mechanisms. Such problems are not yet widely appreciated by the users of CSMs. These problems can be avoided by using

a model that is appropriate for the data; but, understanding the relationship between the data and a fitted model is a difficult task. We argue that the only way to properly understand that relationship is to perform in silico experiments using a generating process that can mimic the data as closely as possible. The mutation-selection modeling framework (MutSel) is presented as the basis of such a generating process. We contend that if complex CSMs continue to be developed for testing explicit mechanistic hypotheses, then additional analyses such as those described in here (e.g., penalized LRTs and estimation of PL) will need to be applied alongside the more traditional inferential methods.

Introduction

Codon substitution models (CSMs) fitted to an alignment of homologous protein-coding genes are commonly used to make inferences about evolutionary processes at the molecular level (see Chapter X for examples of different applications of CSMs). Such processes (e.g., mutation and selection) are represented by a vector of parameters θ that can be estimated using maximum likelihood (ML) or Bayesian statistical methods. Here we focus on ML and for convenience use CSM to indicate a model that is used in conjunction with the ML approach (see Huelsenbeck and Dyer, 2004, for an example of the Bayesian approach). Considerable apprehension was expressed about the statistical validity of CSMs during their initial phase of development. In particular were concerns over the risk of falsely inferring that a sequence or codon site evolved by adaptive evolution (Suzuki and Nei, 2001, 2002, 2004; Zhang, 2004; Hughes, 2007; Friedman and Hughes, 2007; Hughes and Friedman, 2008; Suzuki, 2008; Nozawa *et al.*, 2009). Many of the studies employed in the critique of CSMs were later shown to be flawed due to statistical errors or incorrect interpretation of results (Wong *et al.*, 2004; Yang, 2006; Yang and dos Reis, 2011; Zhai *et al.*, 2012). In their re-analysis of the iconic MHC dataset (Hughes and Nei, 1988), for example, Suzuki and Nei (2001) based their criticism of the ML approach on results that were incorrect due to computational issues (Wong *et al.*, 2004). And in simulation studies by Suzuki (2008) and Nozawa *et al.* (2009), the branch-site model of Yang and Nielsen (2002) was criticised as being too liberal because it falsely inferred positive selection at 32 out of 14,000 simulated sites, despite that this rate (0.0023) was well below the level of significance of the test ($\alpha = 0.05$) (Yang and dos Reis, 2011). Concerns about the ML approach were eventually mollified by numerous simulation studies showing that

the false positive rate is no greater than the specified level of significance of the LRT under a wide range of evolutionary scenarios (Anisimova *et al.*, 2001, 2002; Wong *et al.*, 2004; Zhang, 2004; Kosakovsky Pond and Frost, 2005; Yang *et al.*, 2005; Zhang *et al.*, 2005; Yang and dos Reis, 2011; Kosakovsky Pond *et al.*, 2011; Lu and Guindon, 2013). The validity and success of the approach is now well established (Zhai *et al.*, 2012), and this has led to the formulation of CSMs of ever-increasing sophistication (Rodrigue *et al.*, 2010; Kosakovsky Pond *et al.*, 2011; Tamuri *et al.*, 2012, 2014; Rodrigue and Lartillot, 2014; Murrell *et al.*, 2015; Smith *et al.*, 2015; Rodrigue and Lartillot, 2016).

The most common use of a CSM is to infer whether a given process, such as adaptive evolution somewhere in the gene, the fixation of double and triple mutations, or variations in the synonymous substitution rate, actually occurred when the alignment was generated. Several factors can potentially undermine the reliability of such inferences. These include:

1. **model misspecification**, which can result in biased parameter estimates;

2. **low information content**, which can cause parameter estimates to have large sampling errors and can lead to excessive false positive rates;

3. **confounding**, which can cause patterns in the data generated by one evolutionary process to be attributed to a different process;

4. **phenomenological load**, which can cause a model parameter to be statistically significant even if the process it represents did not actually occur when the data was generated.

These same factors can impact any model based effort to make inferences from data generated by complex biological processes, not only to the CSMs described here. The possibility of false inference due to any combination of these factors does not imply that the CSM approach is unreliable in principle. As has been demonstrated by numerous successful applications, CSMs generally extract accurate and useful information provided the model is well suited for the data at hand (Yang and Bielawski, 2000; Yang, 2005; Anisimova and Kosiol, 2009). We maintain that the validity of inferences is not a function of the model in and of itself, but is a consequence of the relationship between the model and the data.

Here we explore this relationship via case studies taken from the historical development of CSMs. Our objective is to be candid about the limitations of CSMs to reliably extract information

from an alignment. But we emphasize that the impact of these limitations (i.e., false positives and confounding) are a consequence of a mismatch between the parameters included in the model and the often limited information contained in the alignment. The case studies are divided into two parts, each corresponding to a distinct phase in the development of CSMs. Phase I is characterized by pioneering efforts to formulate CSMs to account for the most prominent components of variation in an alignment (Muse and Gaut, 1994; Goldman and Yang, 1994). These include the M-series models that were among the first CSMs to account for variations in selection effects across sites (Yang *et al.*, 2000), and the branch site model of Yang and Nielsen (2002) (hereafter, YN-BSM) formulated to account for variations in selection effects across both sites and branches. The first pair of case studies exemplifies concerns about the impact of low information content (Case Study A) and model misspecification (Case Study B) on the probability of falsely detecting positive selection in a gene or at a particular codon site. We also include a description of methods recently developed to mitigate the problem of false inference.

Phase II in the historical development is characterized by the general increase in the complexity of CSMs aimed to account for more subtle components of variation in an alignment [1]. Models used to detect temporal changes in site-specific selection effects (e.g., Guindon *et al.*, 2004; Kosakovsky Pond *et al.*, 2011; Smith *et al.*, 2015) or "heterotachy" (Lopez *et al.*, 2002) are representative. The movement toward complex parameter-rich models has resulted in a new set of concerns that are not yet widely appreciated. Principal among these is an increase in the possibility of confounding. Two components of the alignment-generating process are confounded if they can produce the same or similar patterns in the data. Such components can be impossible to disentangle without the input of further biological information, and their existence can lead to a statistical pathology that we call phenomenological load (PL). The second pair of case studies illustrates the possibility of false inference due to confounding (Case Study C) and PL (Case Study D). An essential feature of these studies is the use of a much more realistic generating model to produce alignments for the purpose of model evaluation.

Recent discoveries made using the mutation-selection (MutSel, Yang and Nielsen, 2007) framework of Halpern and Bruno (1998), which is based on a realistic approximation of population dynamics at individual codon sites, have challenged the way we think about the relationship be-

[1]The original CSM proposed by Goldman and Yang (1994) was in fact quite complex in that it adjusted substitution rates between nonsynonymous codons to account for differences in physicochemical properties using the Grantham matrix (Grantham, 1974). This approach was later abandoned in favor of the simpler formulation now known as M0 (Nielsen and Yang, 1998), e.g., the first M-series model (Yang *et al.*, 2000).

tween parameters of traditional CSMs and components of the process of molecular evolution they are meant to summarize (e.g., Spielman and Wilke, 2015a,b; Jones *et al.*, 2017, 2018). Previously, there has been a tendency to think about alignment-generating processes as if they occur in the same way they are modelled by a CSM. This way of thinking can be misleading because mechanisms of protein evolution can differ in important and substantial ways from traditional CSMs. To redress this issue, we begin this article with a brief overview of the conceptual foundations of MutSel as a more realistic way of thinking about the actual process of molecular evolution. This material is followed by a novel presentation of the ML statistical framework intended to illustrate potential limitations in what can reasonably be inferred when a CSM is fitted to data.

Conceptual Foundations

How should we think about the alignment-generating process?

A codon substitution model represents an attempt to explain the way a target protein-coding gene changed over time by a combination of mutation, selection (purifying as well as adaptive), and drift. Adaptive evolution occurs at each site within a protein in response to a hierarchy of effects, including, but not limited to, changes in the network of the protein's interactions, changes in the functional properties of that network, and changes in both the cellular and organismal environment over time. The result of the complex interplay between these effects is typically viewed through the narrow lens of an alignment of homologous sequences X obtained from extant species, possibly accompanied by a tree topology τ (for our purposes, it will always assumed that τ is known). The information contained in X is evidently insufficient to resolve all of the effects of the true generating process, which would in any case be difficult or even impossible to parameterize with any accuracy. It is therefore necessary to base the formulation of a CSM on a number of simplifying assumptions. The usual assumptions include that:

1. sites evolved independently;

2. each site evolved via a homogenous substitution process over the tree (formally, by a Markov process governed by a substitution rate matrix Q);

3. the selection regime at a site is determined by Q_j drawn from a small set of possible substitution rate matrices $\{Q_1, ..., Q_k\}$;

4. all sites share a common vector of stationary frequencies and evolved via a common mutation process.

The elements q_{ij} of a substitution rate matrix Q are typically defined for codons $i \neq j$ as follows (Nielsen and Yang, 1998):

$$q_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ by more than one nucleotide} \\ \pi_j & \text{for synonymous transversions} \\ \kappa\pi_j & \text{for synonymous transitions} \\ \omega\pi_j & \text{for nonsynonymous transversions} \\ \omega\kappa\pi_j & \text{for nonsynonymous transitions} \end{cases} \quad (1)$$

where κ is the transition bias and π_i is the stationary frequency of the i^{th} codon, both assumed to be the same for all codon sites. The ratio $\omega = dN/dS$ of the nonsynonymous substitution rate dN to the synonymous substitution rate dS (both adjusted for "opportunity" [2]) quantifies the stringency of selection at the site, with values closer to zero corresponding to sites that are more strongly conserved. We follow standard notation and use $\hat{\omega}$ to represent the maximum likelihood estimate (MLE) of ω obtained by fitting equation (1) to an alignment.

Equation (1) provides the building block for most CSMs, yet it is unsuitable as a means to think about the substitution process at a site. For instance, the rate ratio in (1) is assumed to be the same for all nonsynonymous pairs of codons. If interpreted mechanistically, this is tantamount to the assumption that the amino acid occupying a site has fitness f and all other amino acids have fitness $f + df$, and that, with each substitution, the newly fixed amino acid changes its fitness to f and the previous occupant changes it fitness to $f + df$. Such a narrow view of the substitution process, akin to frequency dependent selection (dos Reis, 2013; Jones *et al.*, 2017), is conceptually misleading for the majority of proteins. To be clear, CSMs are undoubtedly a valuable tool to make inferences about the evolution of a protein (e.g., Yang and Bielawski, 2000; Yang, 2005; Field *et al.*, 2006; Sawyer *et al.*, 2007); our point is that they do not necessarily provide the best way to *think* about the process.

[2]Single nucleotide (SN) mutations that are nonsynonymous occur more frequently that those that are synonymous due to idiosyncrasies in the genetic code. This is accounted for in the formulation of dN and dS, so that dN can be interpreted as the proportion of nonsynonymous SN mutations that are fixed. Likewise, dS is the proportion of synonymous SN mutations that are fixed. See Jones *et al.* (2017) for a discussion of various interpretations of dN/dS.

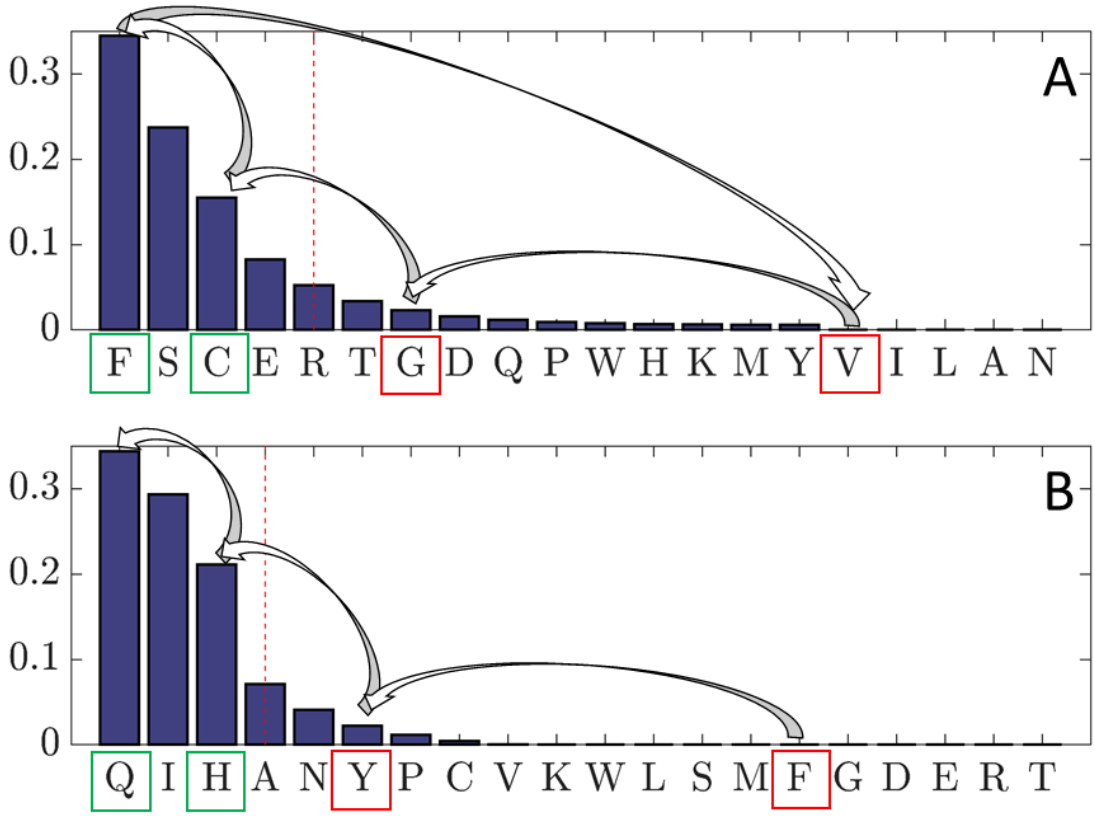

Figure 1: It can be useful to think of the substitution process at a site as movement on a site-specific fitness landscape. The horizontal axis in each figure shows the amino acids at a hypothetical site in order of their stationary frequencies indicated by the height of the bars. Frequency is a function of mutation and selection, but can be construed as a proxy for fitness. The site-specific dN/dS ratio (Jones *et al.*, 2017) is a function of the amino acid that occupies the site, and can be < 1 (left of the red dashed line) or > 1 (right of the dashed red line). **(A)** Suppose phenylalanine (F, TTT) is the fittest amino acid. The site-specific dN/dS ratio is much less than one when occupied by F because any nonsynonymous mutation will always be to an amino acid that is less fit. Nevertheless, it is possible for an amino acid such as valine (V, GTT) to be fixed on occasion, provided that selection is not too stringent. When this happens, dN/dS at the site is temporarily elevated to a value greater than one as positive selection moves the site back to F by a series of replacement substitutions e.g., V (GTT) $\rightarrow$ G (GGT) $\rightarrow$ C (TGT) $\rightarrow$ F (TTT). We call the episodic recurrence of this process **shifting balance** on a static fitness landscape. Shifting balance on a landscape for which all frequencies are approximately equal corresponds to **nearly-neutral** evolution (not depicted), when dN/dS is always ≈ 1. **(B)** Now consider what happens following a change in one or more external factors that impact the functional significance of the site. The relative fitnesses of the amino acids might change from that depicted in **A** to that in **B** for instance, where glutamine (G) is fittest. If at the time of the change the site is occupied by F (as is most likely), then dN/dS would be temporarily elevated as positive selection moves the site toward its new peak at Q e.g., F (TTT) $\rightarrow$ Y (TAT) $\rightarrow$ H (CAT) $\rightarrow$ Q (CAA). This process of **adaptive evolution** is followed by a return to shifting balance once the site is occupied by F.

The way we think about the substitution process should not be limited to unrealistic assump-

tions used to formulate a tractable CSM. It is more informative to conceptualize evolution at a codon site using the traditional metaphor of a fitness landscape upon which greater height represents greater fitness as depicted in Figure 1. If sites are assumed to evolve independently, a **site-specific fitness landscape** can be defined for the h^{th} site by a vector of fitness coefficients $\mathbf{f}^h$ and its implied vector of equilibrium codon frequencies $\boldsymbol{\pi}^h$. Combined with a model for the mutation process, $\boldsymbol{\pi}^h$ determines the evolutionary dynamics at the site, or the way it "moves" over its landscape (more formally, the way mutation and fixation events occur at a codon site in a population over time). This provides a way to think about evolution at a codon site in terms of three possible dynamic regimes: **shifting balance**, under which the site moves episodically away from the peak of its fitness landscape (i.e., the fittest amino acid) via drift and back again by positive selection (Figure 1A); **adaptive evolution**, under which a change in the landscape is followed by movement of the site toward its new fitness peak (Figure 1B); and **neutral** or **nearly-neutral evolution**, under which drift dominates and the site is free to move over a relatively flat landscape limited primarily by biases in the mutation process. This way of thinking about the alignment-generating process is encapsulated by the MutSel framework (dos Reis, 2013, 2015; Jones *et al.*, 2017). The precise relationship between the MutSel framework and the three dynamic regimes will be presented in Case Study C.

What is the objective of model-building?

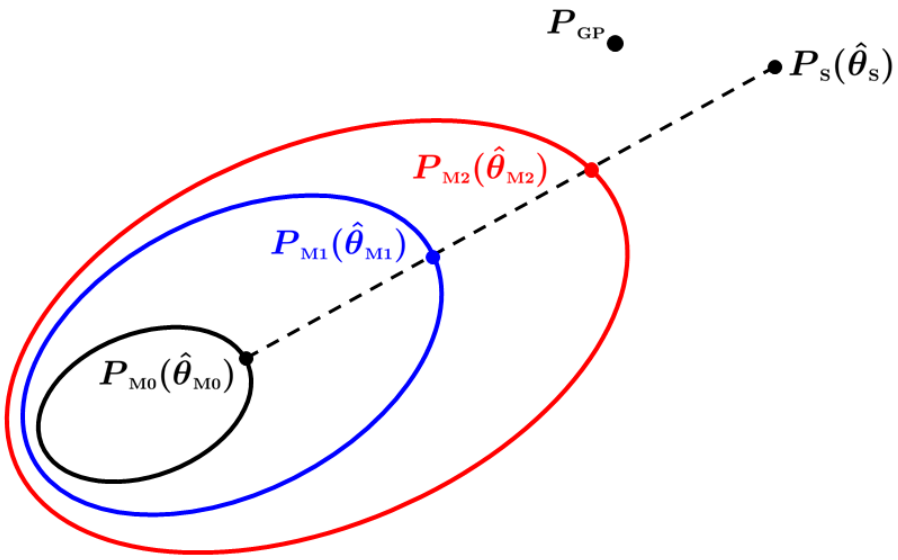

Figure 2: The $(61^N - 1)$-dimensional simplex containing all possible site-pattern distributions for an N-taxon alignment is depicted. The inner-most ellipse represents the subspace $\{P_{M0}(\theta_{M0}) \mid \theta_{M0} \in \Omega_{M0}\}$ that is the family of distributions that can be specified using M0, the simplest of CSMs. This is nested in the family of distributions that can be specified using M1 (blue ellipse), a hypothetical model that has the same parameters as M0 plus some extra parameters. Similarly, M1 is nested in M2 (red ellipse). Whereas models are represented by subspaces of distributions, the true generating process is represented by a single point P_{GP}, the location of which is unknown. The empirical site-pattern distribution $P_S(\hat{\theta}_S)$ correponds to the saturated model fitted to the alignment; with large samples $P_S(\hat{\theta}_S) \approx P_{GP}$. For any other model M, the member $P_M(\hat{\theta}_M) \in \{P_M(\theta_M) \mid \theta_M \in \Omega_M\}$ most consistent with X is the one that minimizes deviance, which is twice the difference between the maximum log-likelihood of the data under the saturated model and the maximum log-likelihood of the data under M.

CSMs have become increasingly complex with the addition of more free parameters since the introduction of the M-series models in Yang *et al.* (2000). The *prima facie* objective of this trend is to produce models that provide better mechanistic explanations of the data. The assumption is that this will lead to more accurate inferences about evolutionary processes, particularly as the volume of genetic data increases (Liberles *et al.*, 2013). However, the significance of a new model parameter is assessed by a comparison of site-pattern distributions without reference to mechanism. Combined with the possibility of confounding, this feature of the ML framework means that the objective of improving model fit does not necessarily coincide with the objective of providing a better representation of the mechanisms of the true generating process.

Given any CSM with parameters θ_M, it is possible to compute a vector $\boldsymbol{P}$ that assigns a

probability to each of the 61^{N} possible site patterns for an N-taxon alignment (i.e., a multinomial distribution for 61^{N} categories). We refer to $\boldsymbol{P} = P_{\text{M}}(\theta_{\text{M}})$ as the site-pattern distribution for that model. Figure 2 depicts the space of all possible site-pattern distributions for an N-taxon alignment. Each ellipse represents the family of distributions $\{P_{\text{M}}(\theta_{\text{M}}) \mid \theta_{\text{M}} \in \Omega_{\text{M}}\}$, where Ω_{M} is the vector-space of all possible values of θ_{M}. For example, $\{P_{\text{M0}}(\theta_{\text{M0}}) \mid \theta_{\text{M0}} \in \Omega_{\text{M0}}\}$ is the family of distributions that can be specified using M0, the simplest CSM that assumes a common substitution rate matrix Q for all sites and branches. This is nested inside $\{P_{\text{M1}}(\theta_{\text{M1}}) \mid \theta_{\text{M1}} \in \Omega_{\text{M1}}\}$, where M1 is a hypothetical model that is the same as M0 but for a few extra parameters. Likewise, M1 is nested in M2. The location of the site-pattern distribution for the true generating process is represented by P_{PG}. Its location is fixed but unknown. It is therefore not possible to assess the distance between it and any other distribution. Instead, comparisons are made using the site-pattern distribution inferred under the saturated model.

Whereas a CSM $\{P_{\text{M}}(\theta_{\text{M}}) \mid \theta_{\text{M}} \in \Omega_{\text{M}}\}$ can be thought of as a family of multinomial distributions for the 61^{N} possible site patterns, the fitted saturated model $P_{\text{S}}(\hat{\theta}_{\text{S}})$ is the unique distribution defined by the MLE $\hat{\theta}_{\text{S}} = (y_1/n, ..., y_m/n)^T$, where $y_i > 0$ is the observed frequency of the i^{th} site pattern, m is the number of unique site patterns, and n is the number of codon sites. In other words, the fitted saturated model is the empirical site-pattern distribution for a given alignment. Because it takes none of the mechanisms of mutation or selection into account, ignores the phylogenetic relationships between sequences, and excludes the possibility of site patterns that were not actually observed (i.e., $y_i/n = 0$ for site patterns i not observed in X), $P_{\text{S}}(\hat{\theta}_{\text{S}})$ can be construed as the maximally phenomenological explanation of the observed alignment. An alignment is always more likely under the saturated model than it is under any other CSM. $P_{\text{S}}(\hat{\theta}_{\text{S}})$ therefore provides a natural benchmark for model improvement.

For any alignment, the MLE over the family of distributions $\{P_{\text{M}}(\theta_{\text{M}}) \mid \theta_{\text{M}} \in \Omega_{\text{M}}\}$ is represented by a fixed point $P_{\text{M}}(\hat{\theta}_{\text{M}})$ in Figure 2. $P_{\text{M}}(\hat{\theta}_{\text{M}})$ is the distribution that minimizes the statistical deviance between $P_{\text{M}}(\theta_{\text{M}})$ and $P_{\text{S}}(\hat{\theta}_{\text{S}})$. Deviance is defined as twice the difference between the maximum log-likelihood (LL) of the data under the saturated model and the maximum log-likelihood of the data under M:

$$D(\hat{\theta}_{\text{M}}, \hat{\theta}_{\text{S}}) = 2\left\{\ell(\hat{\theta}_{\text{S}} \mid X) - \ell(\hat{\theta}_{\text{M}} \mid X)\right\} \tag{2}$$

A key feature of deviance is that it always decreases as more parameters are added to the model,

corresponding to an increase in the probability of the data under that model. For example, suppose $\{P_{\text{M2}}(\theta_{\text{M2}}) \mid \theta_{\text{M2}} \in \Omega_{\text{M2}}\}$ is the same family of distributions as $\{P_{\text{M1}}(\theta_{\text{M1}}) \mid \theta_{\text{M1}} \in \Omega_{\text{M1}}\}$ but for the inclusion of one additional parameter ψ, so that $\theta_{\text{M2}} = (\theta_{\text{M1}}, \psi)$. The improvement in the probability of the data under $P_{\text{M2}}(\hat{\theta}_{\text{M2}})$ over its probability under $P_{\text{M1}}(\hat{\theta}_{\text{M1}})$ is assessed by the size of the reduction in deviance induced by ψ:

$$\Delta D(\hat{\theta}_{\text{M1}}, \hat{\theta}_{\text{M2}}) = D(\hat{\theta}_{\text{M1}}, \hat{\theta}_{\text{S}}) - D(\hat{\theta}_{\text{M2}}, \hat{\theta}_{\text{S}}) = 2\left\{\ell(\hat{\theta}_{\text{M2}} \mid X) - \ell(\hat{\theta}_{\text{M1}} \mid X)\right\} \tag{3}$$

Equation (3) is just the familiar log-likelihood ratio (LLR) used to compare nested models under the maximum likelihood framework.

Given this measure of model improvement, the *de facto* objective of model building is not to provide a mechanistic explanation of the data that more accurately represents the true generating process, but only to move closer to the site-pattern distribution of the fitted saturated model. Real alignments are limited in size, so there will always be some distance between $P_{\text{S}}(\hat{\theta}_{\text{S}})$ and P_{GP} due to sampling error (as represented in Figure 2). But even with an infinite number of codon sites, when $P_{\text{S}}(\hat{\theta}_{\text{S}})$ converges to P_{GP}, the criterion of minimizing deviance does not inevitably lead to a better explanation of the data because of the possibility of confounding. Two process are said to be confounded if they can produce similar patterns in the data. Hence, if ψ represents a process E that did not actually occur when the data was generated, and if E is confounded with another process that did occur, the LLR in equation (3) can still be significant. Under this scenario, the addition of ψ to M1 would engender movement toward $P_{\text{S}}(\hat{\theta}_{\text{S}})$ and P_{GP}, but the new model M2 would also provide a worse mechanistic explanation of the data because it would falsely indicate that E occurred. The possibility of confounding and its impact on inference is demonstrated in Case Study D.

Phase I: Pioneering CSMs

The first effort to detect positive selection at the molecular level (Hughes and Nei, 1988) relied on heuristic counting methods (Nei and Gojobori, 1986). Phase I of CSM development followed with the introduction of formal statistical approaches based on ML (Muse and Gaut, 1994; Goldman and Yang, 1994). The first CSMs were used to infer whether the estimate $\hat{\omega}$ of a single nonsynonymous to synonymous substitution rate ratio averaged over all sites and branches was significantly greater

than one. Such CSMs were found to have low power due to the pervasiveness of synonymous substitutions at most sites within a typical gene (Yang and Bielawski, 2000). An early attempt to increase the statistical power to infer postive selection was the CSM designed to detect $\hat{\omega} > 1$ on specific branches (Yang and Nielsen, 1998). Models accounting for variations in ω across sites were subsequently developed, the most prominent of which are the M-series models (Yang and Nielsen, 1998; Yang *et al.*, 2000). These were accompanied by methods to identify individual sites under positive selection. The quest for power culminated in the development of models that account for variations in the rate ratio across both sites and branches. The appearance of various branch-site models (e.g., Yang and Nielsen, 2002; Forsberg and Christiansen, 2003; Bielawski and Yang, 2004; Zhang *et al.*, 2005) marks the end of Phase I of CSM development.

Two case studies are employed in this section to illustrate some of the inferential challenges associated with Phase I models. We use Case Study A to examine the impact of low information content on the inference of positive selection at individual codon sites. The subject of this study is the M1a vs M2a model contrast applied to the *tax* gene of the human T-cell lymphotropic virus type I (HTLV-I Suzuki and Nei, 2004; Yang *et al.*, 2005). We use Case Study B to illustrate how model misspecification (i.e., differences between the fitted model and the generating process) can lead to false inferences. The subject of this study is the Yang-Nielsen Branch-Site Model (YN-BSM, Yang and Nielsen, 2002) applied to simulated data.

Case Study A: Low Information Content

To study the impact of low information content on inference, we use a pair of nested M-series models known as M1a and M2a (Wong *et al.*, 2004; Yang *et al.*, 2005). Under M1a, sites are partitioned into two rate-ratio categories, $0 < \omega_0 < 1$ and $\omega_1 = 1$ in proportions p_0 and $p_1 = 1 - p_0$. M2a includes an additional category for the proportion of sites $p_2 = 1 - p_0 - p_2$ that evolved under positive selection with $\omega_2 > 1$. The use of multiple categories permits two levels of inference. The first is an omnibus likelihood ratio test (LRT) for evidence of positive selection somewhere in the gene, which is conducted by contrasting a pair of nested models. For example, the contrast of M1a vs M2a is made by computing the distance $\text{LLR} = \Delta D(\hat{\theta}_{\text{M1a}}, \hat{\theta}_{\text{M2a}})$ between the two models and comparing the result to the limiting distribution of the LLR under the null model. In this case, the limiting distribution of LLR is often taken to be χ^2_2 (Yang, 2017), which would be correct under regular likelihood theory because the models differ by two parameters. The second level of

inference is used to identify individual sites that underwent positive selection. This is conducted only if positive selection is inferred by the omnibus test (e.g., if LLR > 5.99 for the M1a vs M2a contrast at the 5% level of significance). Let c_0, c_1, and c_2 represent the event that a given site pattern x falls into the stringent ($0 < \hat{\omega}_0 < 1$), neutral ($\hat{\omega}_1 = 1$), or positive ($\hat{\omega}_2 > 1$) selection category, respectively. Applying Bayes' rule:

$$\Pr(c_2 \mid x, \hat{\theta}_{\text{M2a}}) = \frac{\Pr(x \mid c_2, \hat{\theta}_{\text{M2a}})\hat{p}_2}{\Pr(x \mid c_0, \hat{\theta}_{\text{M2a}})\hat{p}_0 + \Pr(x \mid c_1, \hat{\theta}_{\text{M2a}})\hat{p}_1 + \Pr(x \mid c_2, \hat{\theta}_{\text{M2a}})\hat{p}_2} \tag{4}$$

Sites with a sufficiently high posterior probability (e.g., $\Pr(c_2 \mid x, \hat{\theta}_{\text{M2a}}) > 0.95$) are inferred to have undergone positive selection. Equation (4) is representative of the naive empirical Bayes (NEB) approach under which MLEs ($\hat{\theta}_{\text{M2a}}$) are used to compute posterior probabilities.

The NEB approach ignores potential errors in parameter estimates that can lead to false inference of positive selection at a site (i.e., a false positive). The resulting false positive rate can be especially high for alignments with low information content. An example setting with low information content arises when there are a substantial number of invariant sites, since these provide little information about the substitution process. The issue of low information content is well illustrated by the extreme case of the *tax* gene, HTLV-I (Suzuki and Nei, 2004). The alignment consists of 20 sequences with 181 codon sites, 158 of which are invariant. The 23 variable sites have only one substitution each: 2 are synonymous and 21 are non-synonymous. The high ratio of nonsynonymous-to-synonymous substitutions suggests that the gene underwent positive selection. This hypothesis was supported by analytic results: the LLR for the M1a vs M2a contrast was 6.96 corresponding to a p-value of approximately 0.03 (Yang *et al.*, 2005). The omnibus test therefore supported the conclusion that the gene underwent positive selection. However, the MLE for p_2 under M2a was $\hat{p}_2 = 1$. Using this value in equation (4) gives $\Pr(c_2 \mid x, \hat{\theta}_{\text{M2a}}) = 1$ for all sites, including the 158 invariable sites. Such an unreasonable result can occur under NEB because, despite the possibility of large sampling errors in MLEs due to low information, $\hat{\theta}_{\text{M2a}}$ is treated as a known value in equation (4).

Bayes empirical Bayes (BEB Yang *et al.*, 2005), a partial Bayesian approach under which rate ratios and their corresponding proportions are assigned discrete prior distributions (cf. Huelsenbeck and Dyer, 2004), was proposed as an alternative to NEB. Numerical integration over the assumed priors tends to provide better estimates of posterior probabilities, particularly in cases where

information content is low. Using BEB in the analysis of the *tax* gene, for example, the posterior probability was $0.91 < \Pr(c_2 \mid x, \hat{\theta}_{\text{M2a}}) < 0.93$ for the 21 sites with a single nonsynonymous change and $0.55 < \Pr(c_2 \mid x, \hat{\theta}_{\text{M2a}}) < 0.61$ for the remaining sites in the *tax* gene (Yang *et al.*, 2005). Hence, the BEB approach mitigated the problem of low information content, as the posterior probability of positive selection at invariant sites was reduced. An alternative to BEB is called smoothed bootstrap aggregation (SBA) (Mingrone *et al.*, 2016). SBA entails drawing site patterns from X with replacement (i.e., bootstrap) to generate a set of alignments $\{X_1, ..., X_m\}$ with similar information content as X. The MLEs $\{\hat{\theta}_i\}_{i=1}^m$ for the vector of model parameters θ is then estimated by fitting the CSM to each $X_i \in \{X_1, ..., X_m\}$. A kernel smoother is applied to these values to reduce sampling errors. The mean value of the resulting smoothed $\{\hat{\theta}_i\}_{i=1}^m$ is then used in equation (4) in place of the MLE for θ obtained from the original alignment to estimate posterior probabilities. This approach was shown to balance power and accuracy at least as well as BEB. But SBA has the advantage that it can accommodate the uncertainty of all parameter estimates (not just those of the ω distribution, as in BEB) and is much easier to implement. When SBA was applied to the *tax* gene, the posterior probabilities for positive selection were further reduced: $0.87 < \Pr(c_2 \mid x, \hat{\theta}_{\text{M2a}}) < 0.89$ for the 21 sites with a single nonsynonymous change, and $0.55 < \Pr(c_2 \mid x, \hat{\theta}_{\text{M2a}}) < 0.60$ for the remaining sites (Mingrone *et al.*, 2016).

The problem of low information content was fairly obvious in the case of the *tax* gene, as 158 of the 181 codon sites within that dataset were invariant. However, it can sometimes be unclear whether there is enough variation in an alignment to ensure reliable inferences. It would be useful to have a method to determine whether a given data set might be problematic. An MLE $\hat{\theta}$ will always converge to a normal distribution centered at the true parameter value θ with variance proportional to $1/n$ as the sample size n (a proxy for information content) gets larger, provided the CSM satisfies certain "regularity" conditions (a set of technical conditions that must hold to guarentee that MLEs will converge in distribution to a normal, and that the LLR for any pair of nested models will converge to its expected chi-squared distribution). This expectation makes it possible to assess whether an alignment is sufficiently informative to obtain the benefits of regularity. The first step is to generate a set of bootstrap alignments $\{X_1, ..., X_m\}$. The CSM can then be fitted to these to produce a sample distribution $\{\hat{\theta}_i\}_{i=1}^m$ for the MLE of any model parameter θ. If the alignment is sufficiently informative with respect θ then a histogram of $\{\hat{\theta}_i\}_{i=1}^m$ should be approximately normal in distribution. Serious departures from normality (e.g.,

a bimodal distribution) indicate unstable MLEs, which are a sign of insufficient information or an irregular modelling scenario. Mingrone *et al.* (2016) recommend using this technique with real data as a means of gaining insight into potential difficulties of parameter estimation using a given CSM.

Irregularity and Penalized Likelihood

Issues associated with low information content can be made worse by violations of certain regularity conditions. For example, M2a is the same as M1a but for two extra parameters, p_2 and ω_2. Usual likelihood theory would therefore predict that the limiting distribution of the LLR is χ^2_2. However, this result is valid only if the regularity conditions hold. Among these conditions is that the null model is not obtained by placing parameters of the alternate on the boundary of parameter space. Since M1a is the same as M2a but with $p_2 = 0$, this condition is violated. The same can be said for many nested pairs of Phase I CSMs, such as M7 vs M8 (Yang *et al.*, 2000) or M1 vs branch-site Model A (Yang and Nielsen, 2002). Although the theoretical limiting distribution of the LLR under some irregular conditions has been determined by Self and Liang (1987), those results do not include cases where one of the model parameters is unidentifiable under the null (Anisimova *et al.*, 2001). Since M1a is M2a with $p_2 = 0$, the likelihood under M1a is the same for any value of ω_2. This makes ω_2 unidentifiable under the null. The limiting distribution for the M1a vs M2a contrast is therefore unknown (Yang, 2014).

A penalized likelihood ratio test (PLRT, Mingrone *et al.*, 2018) has been proposed to mitigate problems associated with unidentifiable parameters. Under this method, the likelihood function for the alternate model (e.g., M2a) is modified so that values of p_2 closer to zero are penalized. This has the effect of drawing the MLE for p_2 away from the boundary, and can interpreted as a way to "regularize" the model. PLRT seems to be more useful in cases where the analysis of a real alignment produces a small value of $\hat{p}_2$ accompanied by an unrealistically large value of $\hat{\omega}_2$. This can happen because $\hat{\omega}_2$ is influenced by fewer and fewer site patterns as $\hat{p}_2$ approaches zero, and is therefore subject to larger and larger sampling errors. In addition, $\hat{\omega}_2$ and $\hat{p}_2$ tend to be negatively correlated, which further contributes to the large sampling errors $\hat{\omega}_2$. For example, Mingrone *et al.* (2018) found that M2a fitted to a 5-taxon alignment with 198 codon sites without penalization gave $(\hat{p}_2, \hat{\omega}_2) = (0.01, 34.70)$. These MLEs, if taken at face value, suggest that a small number of sites in the gene underwent positive selection. However, such a large rate ratio is difficult to believe

given that its estimate is consistent with only approximately 2 codon sites (e.g., an estimated 1% of the 198 sites or $\approx$ 2 sites). Using the PLRT, the MLEs were $(\hat{p}_2, \hat{\omega}_2) = (0.09, 1.00)$. These suggest that selection pressure was nearly neutral at a significant proportion of sites in the gene. In this case, the rate ratio is consistent with 9% of the 198 sites or $\approx$ 18 sites and is therefore less likely to be an artifact of sampling error. We expect this approach to be useful in a wide variety of evolutionary applications that rely on mixture models to make inferences (e.g., Pagel and Meade, 2004; Lartillot and Philippe, 2004; Wang *et al.*, 2008; Gaston *et al.*, 2011).

Other approaches for dealing with low information content in the data for an individual gene include the empirical Bayes approach of Kosiol *et al.* (2008) and the parametric bootstrapping methods of Gibbs (2007). Both methods exploit the additional information content available from other genes. Kosiol *et al.* (2008) adopted an empirical Bayes approach, where ω values varied over edges and genes according to a distribution. Because empirical posterior distributions are used, the approach is more akin to detecting sites under positive selection (e.g., using NEB) than formal testing. By contrast, Gibbs (2007) adopted a test-based approach and utilized parametric bootstrapping (Goldman, 1993) to approximate the distribution of the likelihood ratio statistic using data from other genes to obtain parameter sets to use in the bootstrap. Whereas this approach can attenuate issues associated with low information content, it can also be computationally expensive, especially when applied to large alignments.

Case Study B: Model Misspecification

The mechanisms that give rise to the diversity of site patterns in a set of homologous genes are highly complex and not fully understood. CSMs are therefore necessarily simplified representations of the true generating process, and are in this sense misspecified. The extent to which misspecification might cause an omnibus LRT to falsely detect positive selection was of primary concern during Phase I of model development. We use a particular form of the YN-BSM called Model A (Yang and Nielsen, 2002) to illustrate this issue. In its orginal form, the omnibus LRT assumes a null under which a proportion p_0 of sites evolved under stringent selection with $\omega_0 = 0$ and the remaining sites evolved under a neutral regime with $\omega_1 = 1$ on all branches of the tree (i.e., model M1 in Nielsen and Yang, 1998). This is contrasted with Model A, which is the same as M1 except that it assumes that some stringent sites and some neutral sites evolved under positive selection with $\omega_2 > 1$ on a pre-specified branch called the foreground branch. The omnibus test contrasting

M1 with Model A was therefore designed to detect a subset of sites that evolved adaptively on the same branch of the tree.

sites	1-20	21-40	41-60	61-80	81-100	101-120	121-140	141-160	161-180	181-200
ω regime X	1.00	1.00	0.80	0.80	0.50	0.50	0.20	0.20	0.00	0.00
ω regime Z	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 1: Rate ratios (ω) for regimes X and Z taken from Zhang (2004).

During this period of model development, the standard method to test the impact of misspecification on the reliability of an omnibus LRT was to generate alignments *in silico* using a more complex version of the CSM to be tested as the generating model. This usually involved adding more variability in ω across sites and/or branches than assumed by the fitted CSM while leaving all other aspects of the generating model the same. In Zhang (2004), for example, alignments were generated using site-specific rate matrices, as in Equation (1), with rate ratios ω specified by pre-determined selection regimes, two of which are shown in Table 1. In one simulation, 200 alignments were generated using regime Z on a single foreground branch and regime X on all of the remaining branches of a 10 or 16 taxon tree. The gene therefore underwent a mixture of stringent selection and neutral evolution over most of the tree (regime X), but with complete relaxation of selection pressure on the foreground branch (regime Z). Positive selection did not occur at any of the sites. Nevertheless, the M1 vs Model A contrast inferred positive selection in 20% to 55% of the alignments, depending on the location of the foreground branch. Such a high rate of false positives was attributed to the mismatch between the process used to generated the data compared to the process assumed by the null model M1 (Zhang, 2004).

The branch-site model was subsequently modified to allow $0 < \omega_0 < 1$ instead of $\omega_0 = 0$ (Modified Model A in Zhang *et al.*, 2005). Furthermore, the new null model is specified under the assumption that some proportion p_0 of sites (the stringent sites) evolved under stringent selection with $0 < \omega_0 < 1$ everywhere in the tree except on the foreground branch, where those same sites evolved neutrally with $\omega_2 = 1$. All other sites in the alignment (the neutral sites) are assumed to have evolved neutrally with $\omega_1 = 1$ everywhere in the tree. This is contrasted with the Modified Model A, which assumes that some of the stringent sites and some of the neutral sites evolved under positive selection with $\omega_2 > 1$ on the foreground. Hence, unlike the original omnibus test that contrasts M1 with Model A, the new test contrasts Modified Model A with $\omega_2 = 1$ against Modified Model A with $\omega_2 > 1$. These changes to the YN-BSM were shown to mitigate the

problem of false inference. For example, using the same generating model with regimes X and Z, the modifed omnibus test falsely inferred positive selection in only 1% to 7.5% of the alignments, consistent with the 5% level of significance of the test (Zhang *et al.*, 2005).

This case study demonstrates how problems associated with model misspecification were traditionally identified, and how they could be completely corrected through relatively minor changes to the model. However, the generating methods employed by studies such as Zhang (2004) and Zhang *et al.* (2005), although sophisticated for their time, produced alignments that were highly unrealistic compared to real data. For example, it was recently shown that a substantial proportion of variation in many real alignments might be due to selection effects associated with shifting balance over static site-specific fitness landscapes (Jones *et al.*, 2017, 2018). This process results in random changes in site-specific rate ratios, or heterotachy, that cannot be replicated using traditional CSMs as the generating model. While the mitigation of statistical pathologies due to low information content (e.g., using BEB or SBA) or model misspecification (e.g., by altering the null and alternative hypotheses or the use of penalized likelihood) were critical advancements during Phase I of CSM development, other statistical pathologies went unrecognized due to reliance on unrealistic simulation methods. This issue is taken up in the next section.

Phase II: Advanced CSMs

A typical protein-coding gene evolves adaptively only episodically (Struder and Robinson-Rechavi, 2009). The evidence of adaptive evolution of this type can be very difficult to detect. For example, it is assumed under the YN-BSM that a random subset of sites switched from a stringent or neutral selection regime to positive selection together on the same set of foreground branches. The power to detect a signal of this kind can be very low when the proportion of sites that switched together is small (Yang and dos Reis, 2011). Perhaps encouraged by the reliability of Phase I models demonstrated by extensive simulation studies (Anisimova *et al.*, 2001, 2002; Wong *et al.*, 2004; Zhang, 2004; Kosakovsky Pond and Frost, 2005; Yang *et al.*, 2005; Zhang *et al.*, 2005; Yang and dos Reis, 2011; Kosakovsky Pond *et al.*, 2011; Lu and Guindon, 2013), combined with experimental validation of results obtained from their application to real data (Yang and Bielawski, 2000; Yang, 2005; Anisimova and Kosiol, 2009), investigators began to formulate increasingly complex and parameter-rich CSMs (Rodrigue *et al.*, 2010; Kosakovsky Pond *et al.*, 2011; Tamuri *et al.*,

2012, 2014; Rodrigue and Lartillot, 2014; Murrell *et al.*, 2015; Smith *et al.*, 2015). The hope was that carefully selected increases in model complexity would yield greater power to detect subtle signatures of positive selection overlooked by Phase I models. The introduction of such CSMs marks the beginning of Phase II of their historical development.

Phase II models fall into three broad categories:

1. The first consists of Phase I CSMs modified to account for more variability in selection effects across sites and branches than previously assumed, with the aim of increasing the power to detect subtle signatures of positive selection (e.g., the Branch-Site Random Effects Likelihood model, BSREL Kosakovsky Pond *et al.*, 2011).

2. The second category includes Phase I CSMs modified to contain parameters for mechanistic processes not directly associated with selection effects. Many such models have been motivated by a particular interest in the added mechanism (e.g., the fixation of double and triple mutations Miyazawa, 2011; Zaheri *et al.*, 2014; Jones *et al.*, 2018), or by the notion that increasing the mechanistic content of a CSM can only improve inferences about selection effects (e.g., by accounting for variations in the synonymous substitution rate, Kosakovsky Pond and Muse, 2007; Rubinstein *et al.*, 2011).

3. The third category of models abandons the traditional formulation of equation (1) in favor of a substitution process expressed in terms of explicit population genetic parameters, such as population size and selection coefficients (Nielsen and Yang, 2003; Rodrigue *et al.*, 2010; Tamuri *et al.*, 2012, 2014; Rodrigue and Lartillot, 2014, 2016).

An example of the first category of models is BSREL, which accounts for variations in selection effects across sites and over branches by assuming a different rate-ratio distribution $\{(\omega_i^b, p_i^b) : i = 1, ..., k_b)\}$ for each branch b of a tree (Kosakovsky Pond *et al.*, 2011). BSREL was later found to be more complex than necessary, so an adaptive version was formulated to allow the number of components k_b on a given branch to adjust to the apparent complexity of selection effects on that branch (aBSREL Smith *et al.*, 2015). A further reduction in model complexity led to the formulation of the test known as BUSTED (for Branch-site Unrestricted Statistical Test for Episodic Diversification, Murrell *et al.*, 2015), which we use to illustrate the problem of confounding in Case Study C. An example of the second category of models is the addition of parameters for the rate of double and triple mutations to traditional CSMs, the most sophisticated

version of which is RaMoSSwDT (for Random Mixture of Static and Switching sites with fixation of Double and Triple mutations, Jones *et al.*, 2018). This model is used in Case Study D to illustrate the problem of phenomenolgical load.

Models in the third category are the most ambitious CSMs currently in use, and are far more challenging to fit to real alignments than traditional models. One of the most impressive examples of their application is the site-wise mutation-selection model (swMutSel: Tamuri *et al.*, 2012, 2014) fitted to a concatenated alignment of 12 mitochondrial genes (3598 codon sites) from 244 mammalian species. Based on the mutation-selection framework of Halpern and Bruno (1998), swMutSel estimates a vector of selection coefficients for each site in an alignment. This and similar models (e.g., Rodrigue *et al.*, 2010; Rodrigue and Lartillot, 2014, 2016) appear to be reliable (Spielman and Wilke, 2016), but require a very large number of taxa (e.g., hundreds). Phase II models of this category are therefore impractical for the majority of empirical datasets. Here we utilize MutSel as an effective means to generate realistic alignments with plausible levels of variation in selection effects across sites and over time rather than as a tool of inference.

Case Study C: Confounding

By expressing the codon substitution process in terms of explicit population genetic parameters, the MutSel framework facilitates the investigation of complex evolutionary dynamics, such as shifting balance on a fixed fitness landscape or adaptation to a change in selective constraints (i.e., a peak shift, dos Reis, 2013; Jones *et al.*, 2017), that are missing from alignments generated using traditional methods. Specifically, by assigning a different vector of fitness coefficients for the twenty amino acids to each site, MutSel can generate more variation in rate ratio across sites and over time than has been realized in past simulation studies (e.g., Table 1). In this way, MutSel provides the basis of a generating model that can be adjusted to produce alignments that closely mimic real data (Jones *et al.*, 2018). MutSel therefore serves to connect demonstrably plausible evolutionary dynamics to the pathology we refer to as confounding.

Under MutSel, the dynamic regime at the h^{th} codon site (e.g., shifting balance, neutral, nearly neutral, or adaptive evolution) is uniquely specified by a vector of fitness coefficients $\mathbf{f}^h = \langle f_1^h, ..., f_m^h \rangle$. It is generally assumed that mutation to any of the three stop codons is lethal, so $m = 61$ for nuclear genes and $m = 60$ for mitochondrial genes. And although it is not a requirement, it is typical to assume that the f_j^h are constant across synonymous codons (Spielman

and Wilke, 2015b; Jones *et al.*, 2017). Given $\mathbf{f}^h$, the elements of a site-specific instantaneous rate matrix A^h can be defined as follows for all $i \neq j$ (cf. equation 1):

$$A_{ij}^h \propto \begin{cases} \mu_{ij} & \text{if } s_{ij}^h = 0 \\ \mu_{ij} \frac{s_{ij}^h}{1-\exp\left(-s_{ij}^h\right)} & \text{otherwise} \end{cases} \tag{5}$$

where μ_{ij} is the rate at which codon i mutates to codon j and $s_{ij}^h = 2N_e(f_j^h - f_i^h)$ is the scaled selection coefficient for a population of haploids with effective population size N_e. The probability that the new mutant j is fixed is approximated by $s_{ij}^h / \{1 - \exp\left(-s_{ij}^h\right)\}$ (Fisher, 1930; Kimura, 1962).

The rate matrix A^h defines the dynamic regime for the site as illustrated in Figure 3. The bar plot shows codon frequencies $\boldsymbol{\pi}^h = \left\langle \pi_1^h, ..., \pi_m^h \right\rangle$ sorted in descending order. A site spends most of its time occupied by codons to the left or near the "peak" of its landscape. The codon-specific rate ratio for the site (dN_i^h/dS_i^h for codon i) is low near the peak (red line plot in Figure 3) since mutations away from the peak are seldom fixed. However, if selection is not too stringent, the site will occasionally drift to the right into the "tail" of its landscape. When this occurs, the codon-specific rate ratio will be elevated for a time until a combination of drift and positive selection moves the site back to its peak. This dynamic between selection and drift is reminiscent of Wright's shifting balance. It implies that, when a population is evolving on a fixed fitness landscape (i.e., with no adaptive evolution), its gene sequences can nevertheless contain signatures of temporal changes in site-specific rate ratios (heterotachy), and that these might include evidence of transient elevation to values greater than one (i.e., positive selection). Such signatures of positive selection due to shifting balance can be detected by Phase II CSMs (Jones *et al.*, 2017).

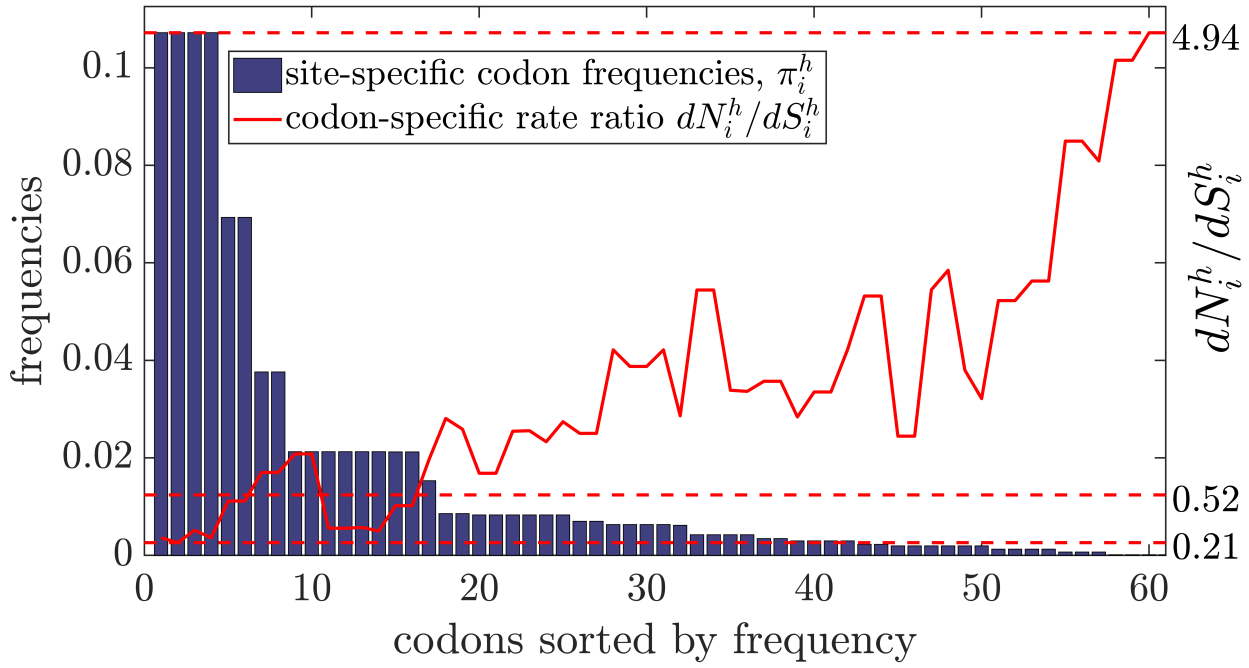

Figure 3: Fitness coefficients for the 20 amino acids were drawn from a normal distribution centered at zero and with standard deviation $\sigma = 0.001$. Bars show the resulting stationary frequencies (a proxy for fitness) sorted from largest to smallest. They compose a metaphorical site-specific landscape over which the site is imagined to move. The solid red line shows the codon-specific rate ratio dN_i^h/dS_i^h for the sorted codons. This varies depending on the codon currently occupying the site, and can be greater than one following a chance substitution into the tail (to the right) of the landscape. In this case, the codon-specific rate ratio for the site ranged from 0.21 to 4.94 with a temporally averaged site-specific rate ratio of $dN^h/dS^h = 0.52$.

For example, BUSTED (Murrell *et al.*, 2015) was developed as an omnibus test for episodic adaptive evolution. The underlying CSM was formulated to account for variations in the intensity of selection over both sites and time modeled as a random effect. This is in contrast to the YN-BSM, which treats temporal changes in rate ratio as a fixed effect that occurs on a pre-specified foreground branch (although the sites under positive selection are still a random effect). We therefore refer to the CSM underlying BUSTED as the Random Effects Branch-Site Model (RE-BSM) to serve as a reminder of this important distinction. Under RE-BSM, the rate ratio at each site and branch combination is assumed to be an independent draw from the distribution $\{(\omega_0, p_0), (\omega_1, p_1), (\omega_2, p_2)\}$. In this way, the model accounts for variations in selection effects both across sites and over time. BUSTED contrasts the null hypothesis that $\omega_0 \leq \omega_1 \leq \omega_2 = 1$ with the alternative that $\omega_0 \leq \omega_1 \leq 1 \leq \omega_2$. When applied to real data, rejection of the null is interpreted as evidence of episodic adaptive evolution.

Unlike the YN-BSM that aims to detect a subset of sites that underwent adaptive evolution together on the same foreground branches (i.e., coherently), BUSTED was designed to detect heterotachy similar to the type predicted by the mutation-selection framework: shifting balance on a static fitness landscape. Jones *et al.* (2017) recently demonstrated that plausible levels of shifting balance can produce signatures of episodic positive selection that can be detected. BUSTED inferred episodic positive selection in as many as 40% of alignments generated using the MutSel framework. Significantly, BUSTED was correct to identify episodic positive selection in these trials. Even though the generating process assumed fixed site-specific landscapes (so there was no episodic adaptive evolution), and the long-run average rate ratio at each site was necessarily less than one (Spielman and Wilke, 2015b), positive selection nevertheless did sometimes occur by shifting balance. This illustrates the general problem of confounding. Two processes are said to be confounded if they can produce the same or similar patterns in the data. In this case, episodic adaptive evolution (i.e., the evolutionary response to changes in site-specific landscapes) and shifting balance (i.e., evolution on a static fitness landscape) are confounded because they can both produce rate-ratio distributions that indicate episodic positive selection. The possibility of confounding underlines the fact that there are limitations in what can be inferred about evolutionary processes based on an alignment alone.

Case Study D: Phenomenological Load

Phenomenological load (PL) is a statistical pathology related to both model misspecification (Case Study B) and confounding (Case Study C) that was not recognized during Phase I of CSM development. When a model parameter that represents a process that played no role in the generation of an alignment (i.e., a misspecified process) nevertheless absorbs a significant amount of variation, its MLE is said to carry PL (Jones *et al.*, 2018). This is more likely to occur when the misspecified process is confounded with one or more other processes that did play a role in the generation of the data, and when a substantial proportion of the total variation in the data is unaccomodated by the null model (Jones *et al.*, 2018). PL increases the probability that a hypothesis test designed to detect the misspecified process will be statistically significant (as indicated by a large LLR) and can therefore lead to the incorrect conclusion that the misspecified process occurred. Critically, Jones *et al.* (2018) showed that PL was only detected when model contrasts were fitted to data generated with realistic evolutionary dynamics using the MutSel

model framework.

To illustrate the impact of PL, we consider the case of CSMs modified to detect the fixation of codons following simultaneous double and triple (DT) nucleotide mutations. The majority of CSMs currently in use assume that codons evolve by a series of single nucleotide substitutions, with the probability for DT changes set to zero. However, recent model-based analyses have uncovered evidence for DT mutations (Whelan and Goldman, 2004; Kosiol *et al.*, 2007; Zaheri *et al.*, 2014). Early estimates of the percentage of fixed mutations that are DT were perhaps unrealistically high. Kosiol *et al.* (2007), for example, estimated a value close to 25% in an analysis of over 7000 protein families from the Pandit database (Whelan *et al.*, 2006). Alternatively, when estimates were derived from a more realistic site-wise mutation-selection model, DT changes comprised less than one percent of all fixed mutations (Tamuri *et al.*, 2012). More recent studies suggest modest rates of between 1% to 3% (Keightley *et al.*, 2009; Schrider *et al.*, 2014; De Maio *et al.*, 2013; Harris and Nielsen, 2014). Whatever the true rate, several authors have argued that it would be beneficial to introduce a few extra parameters into a standard CSM to account for DT mutations (e.g., Miyazawa, 2011; Zaheri *et al.*, 2014). The problem with this suggestion is that episodic fixation of DT mutations can produce signatures of heterotachy consistent with shifting balance.

Recall the comparison of M1, a CSM containing parameters represented by the vector θ_1, and M2, the same model but for the inclusion of one additional parameter ψ, so that $\theta_2 = (\theta_1, \psi)$. The parameter ψ will reduce the deviance of M1 over M2 by some proportion of the baseline deviance between the simplest CSM (M0) and the saturated model $P_{\text{S}}(\hat{\theta}_{\text{S}})$. We call this the percent reduction in deviance (PRD) attributed to $\hat{\psi}$:

$$\text{PRD}(\hat{\psi}) = \frac{\Delta D(\hat{\theta}_{\text{M1}}, \hat{\theta}_{\text{M2}})}{\Delta D(\hat{\theta}_{\text{M0}}, \hat{\theta}_{\text{S}})} \tag{6}$$

Suppose M1 and M2 were fitted to an alignment and that the LLR $= \Delta D(\hat{\theta}_{\text{M1}}, \hat{\theta}_{\text{M2}})$ was found to be statistically significant. This would lead an analyst to attribute the PRD($\hat{\psi}$) to real signal for the process ψ was meant to represent, possibly combined with some PL and noise. Now consider the case in which the process represented by ψ did not actually occur (i.e., it was not a component of the true generating process). Under this scenario, PRD($\hat{\psi}$) would contain no signal, but would be entirely due to PL plus noise. When this is known to be the case, we set PRD($\hat{\psi}$) = PL($\hat{\psi}$). As illustrated below, PL($\hat{\psi}$) can be large enough to result in rejection of the null, and therefore

lead to a false conclusion about the data generating process.

We illustrate PL by contrasting the model RaMoSS with a companion model RaMoSSwDT that accounts for the fixation of DT mutations via two rate parameters, α (the double mutation rate) and β (the triple mutation rate) (Jones *et al.*, 2018). RaMoSS combines the standard M-series model M3 with the covarion-like model CLM3 (cf., Galtier, 2001; Guindon *et al.*, 2004). Specifically, RaMoSS mixes (with proportion p_{M3}) one model with two rate-ratio categories $\omega_0 < \omega_1$ that are constant over the entire tree with a second model (with proportion $p_{\text{CLM3}} = 1 - p_{\text{M3}}$) under which sites switch randomly in time between $\omega_0' < \omega_1'$ at an average rate of δ switches per unit branch length. Fifty alignments were simulated to mimic a real alignment of 12 concatenated H-strand mitochondrial DNA sequences (3331 codon sites) from 20 mammalian species as distributed in the PAML package (Yang, 2007). The generating model, MutSel-mmtDNA (Jones *et al.*, 2018), was based on the mutation-selection framework and produced alignments with single nucleotide mutations only. Since DT mutations are not fixed under MutSel-mmtDNA, the PRD carried by $(\hat{\alpha}, \hat{\beta})$ in each trial can be equated to PL (plus noise). The resulting distribution of $\text{PL}(\hat{\alpha}, \hat{\beta})$ is shown as a boxplot Figure 4.

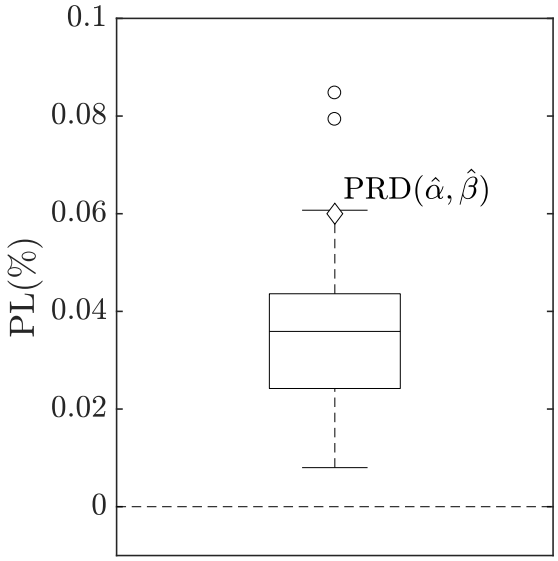

Figure 4: The box plot depicts the distribution of the phenomenological load (PL) carried by $(\hat{\alpha}, \hat{\beta})$ produced by fitting the RaMoSS vs RaMoSSwDT contrast to 50 alignments generated under MutSel-mmtDNA: the circles represent outliers of this distribution. The diamond is the percent reduction in deviance for the same parameters estimated by fitting RaMoSS vs RaMoSSwDT to the real mtDNA alignment.

Although DT mutations were not fixed when the data was generated, shifting balance on a static landscape can produce similar site patterns as a process that includes rare fixation of DT mutations (site patterns exhibiting both synonymous and nonsynonymous substitutions, Jones *et al.*, 2018)[3]. DT and shifting balance are therefore confounded. And since shifting balance tends to occur at a substantial proportion (approximately 20%) of sites when an alignment is generated under MutSel-mmtDNA, DT mutations were falsely inferred by the LRT in 48 of 50 trials at the 5% level of significance (assuming LLR $\approx \chi^2_2$ for the two extra parameters α and β in RaMoSSwDT compared to RaMoSS). The PRD$(\hat{\alpha}, \hat{\beta})$ when RaMoSS vs RaMoSSwDT was fitted to the real mmtDNA is shown as a diamond in the same plot. Although $(\hat{\alpha}, \hat{\beta})$ estimated from the real mmtDNA were found to be highly significant (LLR = 84, p-value $<<$ 0.001), the PRD$(\hat{\alpha}, \hat{\beta})$ was found to be just under the 95^{th} percentile of PL$(\hat{\alpha}, \hat{\beta})$ (PRD = 0.060% compared to the 95^{th} percentile of PL = 0.061). The evidence for DT mutations in the real data is therefore only marginal, and it is reasonable to suspect that its PRD$(\hat{\alpha}, \hat{\beta})$, if not entirely the result of PL, is at least partially caused by PL.

Discussion

CSMs have been subjected to a certain degree of censure, particularly during Phase I of their development. (Suzuki and Nei, 2001, 2002, 2004; Zhang, 2004; Hughes, 2007; Friedman and Hughes, 2007; Hughes and Friedman, 2008; Suzuki, 2008; Nozawa *et al.*, 2009). We maintain that it is not the model in and of itself, or the maximum likelihood framework it is based on, that gives rise to statistical pathologies, but the relationship between model and data. This principle was illustrated by our analysis of the history of CSM development, which we divided into two phases. Phase I was characterized by the formulation of models to account for differences in selection effects across sites and over time that comprise the major component of variation in an alignment. Starting with M0, such models represent large steps toward the fitted saturated model in Figure 2, and also provide a better representation of the true generating process. The main criticism of Phase I models was the possibility of falsely inferring positive selection in a gene or at an individual codon site (Suzuki and Nei, 2002, 2004; Zhang, 2004). But the most compelling empirical case of

[3]It has previously been noted that the rapid fixation of compensatory mutations following substitution to an unstable base pair (e.g., AT→GT→GC) can also produce site patterns that suggest fixation of DT mutations (Yang, 2014, page 46).

false positives was shown to be the result of inappropriate application of a complex model to a sparse alignment (Suzuki and Nei, 2004). Methods for identifying (bootstrap) and dealing with (BEB, SBA, PLRT) low information content were illustrated in Case Study A.

The other big concern that arose during Phase I development was the possibility of pathologies associated with model misspecification. The method used to identify such problems was to fit a model to alignments generated under a scenario contrived to be challenging, as illustrated in Case Study B. There, the omnibus test based on Model A of the YN-BSM was shown to result in an excess of false positives when fitted to alignments simulated using the implausible but difficult "XZ" generating scenario (e.g., with complete relaxation of selection pressure at all sites on one branch of the tree, Table 1). Subsequent modifications to the test reduced the false positive rate to acceptable values. Hence, Case Study B underlines the importance of the model-data relationship. However, it is not clear whether a model adjusted to suit an unrealistic data-generating process is necessarily more reliable when fitted to a real alignment. This difficulty highlights the need to find ways, for the purpose of model testing and adjustment, to generate alignments that mimic real data as closely as possible.

Confidence in the CSM approach, combined with the exponential increase in the volume of genetic data and the growth of computational power, spurred the formulation CSMs of ever-increasing complexity during Phase II. The main issue with these models, which has not been widely appreciated, is confounding. Two processes are confounded if they can produce the same or similar patterns in the data. It is not possible to identify such processes when viewed through the narrow lens of an alignment (i.e., site patterns) alone. This was illustrated by Case Study C, where shifting balance on a static landscape was shown to be confounded with episodic adaptive evolution (dos Reis, 2015; Jones *et al.*, 2017). Confounding can lead to what we call phenomenological load, as demonstrated in Case Study D. In that analysis, the parameters (α, β) were assigned a specific mechanistic interpretation, the rate at which double and triple mutations arise. It was shown that (α, β) can absorb variations in the data caused by shifting balance; hence, the MLEs $(\hat{\alpha}, \hat{\beta})$ resulted in a significant reduction in deviance in 48/50 trials (Figure 4), and therefore improved the fit of the model to the data. However, the absence of DT mutations in the generating process invalidated the intended interpretation of $(\hat{\alpha}, \hat{\beta})$. This result underlines that a better fit does not imply a better mechanistic representation of the true generating process.

It is natural to assume that a better mechanistic representation of the true generating process

can be achieved by adding parameters to our models to account for more of the processes believed to occur. The problem with this assumption is that the metric of model improvement under ML (reduction in deviance) is independent of mechanism. A parameter assigned a specific mechanist interpretation is consequently vulnerable to confounding with other processes that can produce the same distribution of site patterns. As CSMs become more complex, its seems likely that the opportunity for confounding will only increase. It would therefore be desirable to assess each new model parameter for this possibility using something like the method shown in Figure 4 whenever possible. The idea is to generate alignments using MutSel or some other plausible generating process in such as way as to mimic the real data as closely as possible, but with the new parameter set to its null value. To provide a second example, consider the test for changes in selection intensity in one clade compared to the remainder of the tree known as RELAX (Wertheim *et al.*, 2014). Under this model, it is assumed that each site evolved under a rate ratio randomly drawn from $\boldsymbol{\omega}_R = \{\omega_1, ..., \omega_k\}$ on a set of pre-specified reference branches, and from a modified set of rate ratios $\boldsymbol{\omega}_T = \{\omega_1^m, ..., \omega_k^m\}$ on test branches, where m is an exponent. A value $0 < m < 1$ moves the rate ratios in $\boldsymbol{\omega}_T$ closer to one compared to their corresponding values in $\boldsymbol{\omega}_R$, consistent with relaxation of selection pressure at all sites on the test branches. Relaxation is indicated when the contrast of the null hypothesis that $m = 1$ versus the alternative that $m < 1$ is statistically significant. The distribution of PL($\hat{m}$) can be estimated from alignments generated with $m = 1$. The PRD($\hat{m}$) estimated from the real data can then be compared to this to assess the impact of PL (cf. Figure 4). This approach is predicated on the existence of a generating model that could have plausibly produced the site patterns in the real data. Jones *et al.* (2018) present a variety of methods for assessing the realism of a simulated alignment, although further development of such methods is warranted. Software based on MutSel is currently available for generating data that mimic large alignments of 100-plus taxa (Pyvolve Spielman and Wilke, 2015a). Other methods have been developed to mimic smaller alignments of certain types of genes (e.g., MutSel-mmtDNA Jones *et al.*, 2017). It is only by the use of these or other realistic simulation methods that the relationship between a given model and an alignment can be properly understood.

References

Anisimova, M. and Kosiol, C. 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol. Biol. Evol.*, 26: 255–271.

Anisimova, M., Bielawski, J. P., and Yang, Z. H. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.*, 18: 1585–1592.

Anisimova, M., Bielawski, J. P., and Yang, Z. H. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol. Biol. Evol.*, 19: 950–958.

Bielawski, J. P. and Yang, Z. H. 2004. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *J. Mol. Evol.*, 59: 121–132.

De Maio, N., Holmes, I., Schlötterer, C., and Kosiol, C. 2013. Estimating empirical codon hidden Markov models. *Mol. Biol.Evol.*, 30: 725–736.

dos Reis, M. 2013. http://arxiv:1311.6682v1. last accessed November 26 2013.

dos Reis, M. 2015. How to calculate the non-synonymous to synonymous rate ratio protein-coding genes under the Fisher-Wright mutation-selection framework. *Biology Letters*, 11: 1–4.

Field, S. F., Bulina, M. Y., Kelmanson, I. V., Bielawski, J. P., and Matz, M. V. 2006. Adaptive evolution of multicolored fluorescent proteins in reef-building corals. *J. Mol. Evol.*, 62: 332–339.

Fisher, R. 1930. The distribution of gene ratios for rare mutations. *Proc. R. Soc. Edinb.*, 50: 205–220.

Forsberg, R. and Christiansen, F. B. 2003. A codon-based model of host-specific selection in parasites, with an application to the influenza a virus. *Mol. Biol. Evol.*, 20: 1252–1259.

Friedman, R. and Hughes, A. L. 2007. Likelihood-ratio tests for positive selection of human and mouse duplicate genes reveal nonconservative and anomalous properties of widely used methods. *Mol. Phylogenet. Evol.*, 542: 388–393.

Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.*, 18: 866–873.

727 Gaston, D., Susko, E., and Roger, A. J. 2011. A phylogenetic mixture model for the identification
728 of functionally divergent protein residues. *Bioinformatics*, 27: 2655–2663.

729 Gibbs, R. A. 2007. Evolutionary and biomedical insights from the Rhesus macaque genome.
730 *Science*, 316: 222–234.

731 Goldman, N. 1993. Statistical tests of models of dna substitution. *Journal of Molecular Evolution*,
732 36: 182–198.

733 Goldman, N. and Yang, Z. H. 1994. Codon-based model of nucleotide substitution for protein-
734 coding dna-sequences. *Mol. Biol. Evol.*, 11: 725–736.

735 Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. *Science*,
736 862–864.

737 Guindon, S., Rodrigo, A. G., Dyer, K. A., and Huelsenbeck, J. P. 2004. Modeling the site-specific
738 variation of selection patterns along lineages. *PNAS*, 101: 12957–12962.

739 Halpern, A. L. and Bruno, W. J. 1998. Evolutionary distances for protein-coding sequences:
740 modeling site-specific residue frequencies. *Mol. Biol. Evol.*, 15: 910–917.

741 Harris, K. and Nielsen, R. 2014. Error-prone polymerase activity causes multinucleotide mutations
742 in humans. *Genome Research*, 9: 1445–1554.

743 Huelsenbeck, J. P. and Dyer, K. A. 2004. Bayesian estimation of positively selected sites. *J. Mol.*
744 *Evol.*, 58: 661–672.

745 Hughes, A. L. 2007. Looking for Darwin in all the wrong places: the misguided quest for positive
746 selection at the nucleotide sequence level. *Heredity*, 99: 364–373.

747 Hughes, A. L. and Friedman, R. 2008. Codon-based tests of positive selection, branch lengths,
748 and the evolution of mammalian immune system genes. *Immunogenetics*, 60: 495–506.

749 Hughes, A. L. and Nei, M. 1988. Pattern of nucleotide substitution at major histocompatability
750 complex class-1 loci reveals overdominant selection. *Nature*, 335: 167–170.

751 Jones, C. T., Youssef, N., Susko, E., and Bielawski, J. P. 2017. Shifting balance on a static
752 mutation-selection landscape: a novel scenario of positive selection. *Mol. Biol. Evol.*, 34: 391–
753 407.

Jones, C. T., Youssef, N., Susko, E., and Bielawski, J. P. 2018. Phenomenological load on model parameters can lead to false biological conclusions. *Mol. Biol. Evol.*, In Press.

Keightley, P., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., and Blaxter, M. 2009. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genet. Res.*, 19: 1195–1201.

Kimura, M. 1962. On the probability of fixation of mutant genes in a population. *Genetics*, 47: 713–719.

Kosakovsky Pond, S. L. and Frost, S. D. W. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.*, 22: 1208–1222.

Kosakovsky Pond, S. L. and Muse, S. V. 2007. Site-to-site variations of synonymous substitution rates. *Mol. Biol. Evol.*, 22: 2375–2385.

Kosakovsky Pond, S. L., Murrell, B., Fourment, M., Frost, S. D. W., Delport, W., and Scheffler, K. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Mol. Biol. Evol.*, 28: 3033–3043.

Kosiol, C., Holmes, I., and Goldman, N. 2007. An empirical codon model for protein sequence evolution. *Mol. Biol. Evol.*, 24: 1464–1479.

Kosiol, C., ř, T., daFonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., and Siepel, A. 2008. Patterns of positive selection in six mammalian genomes. *PLoS Genetics*, 4: 1–17.

Lartillot, N. and Philippe, H. 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.*, 21: 1095–1109.

Liberles, D. A., Teufel, A. I., Liu, L., and Stadler, T. 2013. On the need for mechanistic models in computational genomics and metagenomics. *Genome Biol. Evol.*, 5: 2008–2018.

Lopez, P., Casane, D., and Phillipe, H. 2002. Heterotachy, and important process of protein evolution. *Mol. Biol. Evol.*, 19: 1–7.

Lu, A. and Guindon, S. 2013. Performance of standard and stochastic branch-site models for detecting positive selection among coding sequences. *Mol. Biol. Evol.*, 31: 484–495.

780 Mingrone, J., Susko, E., and Bielwaski, J. P. 2016. Smoothed bootstrap aggregation for assessing
781 selection pressure at amino acid sites. *Mol. Biol. Evol.*, 33: 2976–2989.

782 Mingrone, J., Susko, E., and Bielwaski, J. P. 2018. Modified likelihood ratio tests for positive
783 selection. *submitted*.

784 Miyazawa, S. 2011. Advantages of a mechanistic codon substitution model for evolutionary analysis
785 of protein-coding sequences. *PLoS ONE*, 6: 20pp.

786 Murrell, B., Weaver, S., Smith, M. D., Wertheim, J. O., Murrell, S., Aylward, A., Eren, K.,
787 Pollner, T., Martin, D. P., Smith, D. M., Scheffler, K., and Pond, S. L. K. 2015. Gene-wide
788 identification of episodic selection. *Mol. Biol. Evol.*, 32: 1365–1371.

789 Muse, S. V. and Gaut, B. S. 1994. A likelihood approach for comparing synonymous and nonsyn-
790 onymous nucleotide substitution rates, with applications to the chloroplast genome. *Mol. Biol.*
791 *Evol.*, 11: 715–724.

792 Nei, M. and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and
793 nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, 3: 418–426.

794 Nielsen, R. and Yang, Z. 2003. Estimating the distribution of selection coefficients from phyloge-
795 netic data with applications to mitochondrial and viral dna. *Mol. Biol. Evol.*, 20: 1231–1239.

796 Nielsen, R. and Yang, Z. H. 1998. Likelihood models for detecting positively selected amino acid
797 sites and applications to the HIV-1 envelope gene. *Genetics*, 148: 929–936.

798 Nozawa, M., Y.Suzuki, and Nei, M. 2009. Reliabilities of identifying positive selection by the
799 branch-site and the site-prediction methods. *PNAS*, 106: 6700–6705.

800 Pagel, M. and Meade, A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity
801 in gene sequence or character-state data. *Syst. Biol.*, 53: 571–581.

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

804 Rodrigue, N. and Lartillot, N. 2016. Detection of adaptation in protein-coding genes using a
805 bayesian site-heterogeneous mutation-selection codon substitution model. *Mol. Biol. Evol.*, 34:
806 204–214.

807 Rodrigue, N., Philippe, H., and Lartillot, N. 2010. Mutation-selection models of coding sequence
808 evolution with site-heterogeneous amino acid fitness profiles. *PNAS*, 107: 4629–4634.

809 Rubinstein, N. D., Doron-Faigenboim, A., Mayrose, I., and Pupko, T. 2011. Evolutionary model
810 accounting for layers of selection in protein-coding genes and their impact on the inference of
811 positive selection. *Mol. Biol. Evol.*, 28: 3297–3308.

812 Sawyer, S. L., Emerman, M., and Malik, H. S. 2007. Discordant evolution of the adjacent an-
813 tiretroviral genes trim22 and trim5 in mammals. *PLoS Pathog.*, 3: e197.

814 Schrider, D., Hourmozdi, J., and Hahn, M. 2014. Pervasive multinucleotide mutational events in
815 eukaryotes. *Curr. Biol.*, 21: 1051–1054.

816 Self, S. G. and Liang, K. Y. 1987. Asymptotic properties of maximum likelihood estimators and
817 likelihood ratio test under nonstandard conditions. *JASA*, 82: 605–610.

818 Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., and Pond, S. L. K. 2015.
819 Less is more: an adaptive branch-site random effects model for efficient detection of episodic
820 diversifying selection. *Mol. Biol. Evol.*, 32: 1342–1353.

821 Spielman, S. and Wilke, C. O. 2015a. Pyvolve: A flexible Python module for simulating sequences
822 along phylogenies. *PLoS ONE*, 10: 1–7.

823 Spielman, S. and Wilke, C. O. 2015b. The relationship between dN/dS and scaled selection
824 coefficients. *Mol. Biol. Evol.*, 34: 1097–1108.

825 Spielman, S. and Wilke, C. O. 2016. Extensively parameterized mutation-selection models reliably
826 capture site-specific selective constraints. *Mol. Biol. Evol.*, 33: 2990–3001.

827 Struder, R. A. and Robinson-Rechavi, M. 2009. Evidence for an episodic model of protein sequence
828 evolution. *Biochem. Soc. Trans.*, 37: 783–786.

829 Suzuki, Y. 2008. False-positive results obtained from the branch-site test of positive selection.
830 *Genes Genet. Syst.*, 83: 331–338.

831 Suzuki, Y. and Nei, M. 2001. Reliabilities of parsimony-based and likelihood-based methods for
832 detecting positive selection at single amino acid sites. *Mol. Biol. Evol.*, 18: 2179–2185.

Suzuki, Y. and Nei, M. 2002. Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.*, 19: 1865–1869.

Suzuki, Y. and Nei, M. 2004. False-positive selection identified by ML-based methods: examples from the Sig1 gene of the diatom Thalassiosira weissflogii and the tax gene of the human T-cell lymphotropic virus. *Mol. Biol. Evol.*, 21: 914–921.

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

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

Wang, H., Li, K., Susko, E., and Rodger, A. J. 2008. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC Evolutionary Biology*, 8: 1–13.

Wertheim, J. O., Murrell, B., Smith, M. D., Pond, S. L. K., and Scheffler, K. 2014. Relax: Detecting relaxed selection in a phylogenetic framework. *Mol. Biol. Evol.*, 32: 820–832.

Whelan, S. and Goldman, N. 2004. Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics*, 167: 2027–2043.

Whelan, S., de Bakker, P. I. W., Quevillon, E., Rodriguez, N., and Goldman, N. 2006. Pandit: an evolution-centric database of protein and associated nucleotide domains with inferred trees. *Nucleic Acids Res.*, 34(Database issue): D327–D331.

Wong, W. S. W., Yang, Z. H., Goldman, N., and Nielsen, R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics*, 168: 1041–1051.

Yang, Z. H. 2005. The power of phylogenetic comparison in revealing protein function. *PNAS*, 102: 3179–3180.

Yang, Z. H. 2006. On the varied pattern of evolution in 2 fungal genomes: a critique of Hughes and Friedman. *Mol. Biol. Evol.*, 23: 2279–2282.

Yang, Z. H. 2007. PAML4: phylogentic anallysis by maximum likelihood. *Mol. Biol. Evol.*, 24: 1586–1591.

Yang, Z. H. 2014. *Molecular Evolution: A Statistical Approach*. Oxford University Press, Oxford.

Yang, Z. H. 2017. *PAML: Phylogenetic Analysis by Maximum Likelihood*. http://abacus.gene.ucl.ac.uk/software/pamlDOC.pdf.

Yang, Z. H. and Bielawski, J. P. 2000. Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, 15: 496–503.

Yang, Z. H. and dos Reis, M. 2011. Statistical properties of the branch-site test of positive selection. *Mol. Biol. Evol.*, 28: 1217–1228.

Yang, Z. H. and Nielsen, R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.*, 46: 409–418.

Yang, Z. H. 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. and Nielsen, R. 2007. Mutation-selection models of codon subsitution and their use to estimate selective strengths on condon usage. *Mol. Biol. Evol.*, 25: 568–579.

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.

Yang, Z. H., Wong, S. W. S., and Nielsen, R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.*, 22: 1107–1118.

Zaheri, M., Dib, L., and Salamin, N. 2014. A generalized mechanistic codon model. *Mol. Biol. Evol.*, 31: 2528–2541.

Zhai, W., Nielsen, R., Goldman, N., and Yang, Z. H. 2012. Looking for Darwin in genomic sequences - validity and success of statistical methods. *Mol. Biol. Evol.*, 20: 2889–2893.

884 Zhang, J. 2004. Frequent false detection of positive selection by the likelihood method with
885 branch-site models. *Mol. Biol. Evol.*, 21: 1332–1339.

886 Zhang, J., Nielsen, R., and Yang, Z. H. 2005. Evaluation of an improved branch-site likelihood
887 method for detecting positive selection at the molecular level. *Mol. Biol. Evol.*, 22: 2472–2479.