genes

MDPI

Review

Using the Mutation-Selection Framework to Characterize Selection on Protein Sequences

Ashley I. Teufel [1,*], Andrew M. Ritchie [2], Claus O. Wilke [1] and David A. Liberles [2]

[1] Department of Integrative Biology, Institute for Cellular and Molecular Biology, and Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin, TX 78712, USA; wilke@austin.utexas.edu

[2] Department of Biology and Center for Computational Genetics and Genomics, Temple University, Philadelphia, PA 19122, USA; tuk03974@temple.edu (A.M.R.); daliberles@temple.edu (D.A.L.)

* Correspondence: ateufel@utexas.edu

Received: 19 June 2018; Accepted: 9 August 2018; Published: 13 August 2018

Abstract: When mutational pressure is weak, the generative process of protein evolution involves explicit probabilities of mutations of different types coupled to their conditional probabilities of fixation dependent on selection. Establishing this mechanistic modeling framework for the detection of selection has been a goal in the field of molecular evolution. Building on a mathematical framework proposed more than a decade ago, numerous methods have been introduced in an attempt to detect and measure selection on protein sequences. In this review, we discuss the structure of the original model, subsequent advances, and the series of assumptions that these models operate under.

Keywords: evolutionary modeling; protein evolution; mutation-selection models

1. Introduction

Alignments of protein sequences and their mapping onto a phylogenetic tree show the role of mutational and selective pressures in generating variation across taxa. Accurately estimating the evolutionary distances between sequences as well as mechanistic parameters associated with distance estimation is crucial to understanding evolutionary processes [1–4]. However, estimating these distances with parameterized models necessitates the non-trivial task of constructing a mapping from genotype to phenotype [5]. The value of methods to bridge from genotype to phenotype has long been appreciated, and inspired the development of early codon-based models of protein evolution that work at the DNA level [6,7]. These models recognized the importance of considering non-synonymous and synonymous evolutionary rates separately, because a non-synonymous mutation alters a protein's sequence and may have an effect on phenotype. The birth of mutation-selection models stems from applying these types of genotype to phenotype ideas to quantifying mutation-selection equilibrium [8].

2. Basic Structure of Model

Mutation-selection models rely on a central idea that changes in a DNA sequence can be thought of as the product of the probability of a mutation from one codon to another, the probability that the mutation fixes in the population, and an arbitrary scaling constant; such a model was first introduced by Halpern and Bruno [1] and parameterized directly from codon frequencies. For each codon in a sequence this allows for the definition of a 61 × 61 matrix of substitution rates. Thus,

$$\begin{aligned} r^i_{ab} &= k \times p_{ab} \times P^i_{\text{fix}}(a,b), \quad b \neq a, \\ r^i_{aa} &= -\sum_{b,b\neq a} r^i_{ab}, \end{aligned} \tag{1}$$

where k is a scaling constant, P^i_{fix} is the probability that the mutation fixes and p_{ab} is the probability that codon a mutates to codon b. The probability of a mutation between two codons can be expressed as

$$p_{ab} = \prod_{j=1}^{3} p_{a_j b_j}, \tag{2}$$

where $p_{a_j b_j}$ is the product of the mutational probability of each nucleotide position j within a codon. Notably, the model can be streamlined by setting p_{ab} equal to zero for double or triple mutations. To approximate the probability of fixation a weak mutation model [9] where mutations fix before the next one is introduced is used. This allows for the probability of fixation to be expressed as

$$P^i_{\text{fix}}(a,b) = \frac{\ln(\frac{\pi_b p_{ba}}{\pi_a p_{ab}})}{1 - \frac{\pi_a p_{ab}}{\pi_b p_{ba}}}, \tag{3}$$

where π_a and π_b are the equilibrium frequencies. This expression explicitly relies on detailed balance assumptions. Using this probability of fixation permits for the calculation of position-specific substitution rates (r^i) from the mutation rates and the position-specific equilibrium frequencies of each codon. It should be noted that in the original model this is done without parameters. The equilibrium frequency of a codon can be calculated from the frequency of the amino acid that it codes for and the nucleotide frequencies, and these values were estimated from empirical data.

Other variations of this framework [10] have also been proposed using different formulations of the probability of fixation based on a diffusion approximation to the evolution of allele frequencies under a Wright-Fisher process. The probability of fixation for a haploid population was originally given as [11]

$$P^i_{fix}(a,b) = \frac{1 - e^{-2s_{ab}}}{1 - e^{-2Ns_{ab}}}. \tag{4}$$

In this formulation, N is the population size and selective advantage is represented by the selection coefficient s_{ab}. Ẇhen $|s_{ab}|$ is small the equation can be reduced as

$$P^i_{fix^{YN}}(a,b) = \frac{2s_{ab}}{1 - e^{-2Ns_{ab}}}. \tag{5}$$

The selection coefficient is here in linear space, while some formulas use a log-transformed coefficient.

Other expressions for the fixation probability are also possible. For example the Sella-Hirsh approximation [12],

$$P^i_{fix^{SH}}(a,b) = \frac{1 - \left(\frac{f_a}{f_b}\right)^2}{1 - \left(\frac{f_a}{f_b}\right)^{2N}}, \tag{6}$$

where f_b is the fitness of a single mutant allele against a wild-type population with fitness f_a.

Each of the above expressions relies on the assumption that mutation is weak and selection strong. More complex dynamics arise if these assumptions are relaxed and more sophisticated methods are required for the calculation of fixation probabilities under these conditions [13,14].

Using this modeling framework, distances between sequences can be calculated by multiplying r^i by t and exponentiating the matrix. This allows for the use of maximum likelihood methods to estimate t, the evolutionary distance between sequences. While this framework was initially conceived to measure these evolutionary distances [1] the key idea that rate of codon change is a product of the mutation rate and the mutation fixation probability described in Equation (1) can also be used to detect and characterize selection in equilibrium and non-equilibrium frameworks.

3. Subsequent Implementations and Advances

While the Halpern and Bruno model [1] results in a set of site-specific matrices, this leads to an invalid assumption that sites behave independently. To help alleviate the reliance on the assumption of site independence, a technique to capture aspects of protein structure by employing a sequence-structure compatibility system [15] was introduced [16]. This model makes use of an empirical energy function that allows for non-synonymous substitution rates to be a function of how compatible a substitution is with a given structure. Additionally, this model was able to estimate selection to maintain sequence-structure compatibility with Markov chain Monte Carlo (MCMC) sampling. However, this model only considered pairs of coding nucleotide sequences, and was later expanded from considering only two taxa to n taxa [17]. Further, this expanded model incorporates information available from empirical amino acid replacement matrices and proposes a formulation at the amino acid level [17]. While these models use a single parameter to distinguish transitions and transversions, a further modification of the model moved to consider nucleotide exchangeabilities [18]. Other methods of capturing protein structure have also been applied. The use of a structurally constrained mean-field substitution model, which considers both unfolding and misfolding stability [19,20] was found to improve model fit over the model proposed at the amino acid level [21]. The speed of MCMC methods that augment substitution histories and assume site-independence was substantially improved by employing a partial sampling technique [22]. Additionally, it is also possible to map between selective coefficients estimated by mutation-selection frameworks and dN/dS values [23].

A further expansion of this modeling framework moved to consider site-specific fitness parameters as random effects by employing a Dirichlet Process (DP) Bayesian framework known as MG-MutSelDP [24]. This implementation of a site-specific mutation-selection model is widely used and currently available through the software package PhyloBayes [25]. A competing method, known as the swMutSel model, was introduced shortly after the MG-MutSelDP model. It employs a maximum penalized-likelihood (MPL) technique [26] and uses a one-model-per-datum approach to describe site-specific fitness. This approach results in a highly parameterized model. For a protein-coding gene of length L codons, it estimates $19 \times L$ site-specific fitnesses. The distinguishing difference between MG-MutSelDP and swMutSel is in how they consider site-specific fitness. The swMutSel model assumes that the site-specific fitness parameters are different at each site. The MG-MutSelDP model describes site-specific fitness parameters as random effects and assumes there is a set of site categories and each specific site is considered to have been generated by one component of a mixture of these categories. To the extent that there are general modes by which evolution interacts with biophysics to describe amino acid substitution, in principle random effects models should be able to identify these discrete modes of action. However, on the other side, the biophysics can be dependent on precise geometric details of interaction (such as interaction distances and side chain angles), so that in practice, this becomes a continuous space of context-dependent modes of interaction [27].

Which of these two models performs better has been a point of contention. The swMutSel model has been criticized for over-parameterization, while the MG-MutSelDP model avoids over-fitting and statistical inconsistencies associated with highly parameterized models [28]. In response the swMutSel model was then updated to use likelihood penalizing functions to partially address this issue [29]. Ultimately, the result of choosing to describe site-specific fitness in the highly parameterized way via the swMutSel model leads to estimates of a large proportion of highly deleterious scaled selection coefficients. It has been argued that these estimates are an erroneous artifact of model overparameterization [28]. The random-effects framework of the MG-MutSelDP model estimates that all scaled selection coefficients are either nearly neutral or weakly deleterious [30]. However, the MG-MutSelDP framework can produce misleading results for certain sites due to inappropriate prior distributions resulting in cases where the amino acid predicted to be the most highly abundant is in fact not abundant at all [28]. Despite the over-parameterization of the swMutSel model, this model is found to give slightly more reliably estimated distributions of selection coefficients, though both

methods perform similarly [30]. Notably, this comparison was done using simulations of 512 taxa on a balanced tree with equal edge lengths and the relative performance of these models depends on the underlying data. In information-poor settings, the over-parameterization of the swMutSel model may be more problematic. Considering that both methods for estimating site-specific selective coefficients from alignments suffer from pitfalls, a method to experimentally measure site-specific selection coefficients was introduced [31] and this method has been shown to significantly improve modeling in specific cases [32,33].

The descriptions of proteins present a characterization of protein evolution based on site-wise descriptions of amino acids. These are treated as site-independent and in equilibrium. Equilibrium assumptions characterize fixed evolutionary constraint and the consequences of this assumption are important. Similarly, while relaxing site-specific constancy of selective pressure is meant to accommodate compensatory processes associated with non-independence, this treatment ultimately breaks down and will also be discussed.

4. Equilibrium Assumptions and Likelihood

The original Halpern and Burno mutation-selection model [1] was formulated under the simplifying assumption that amino acid fitnesses at each site would remain constant over time and across the phylogeny, giving rise to a stationary process of evolution. Biologically, this implies a situation in which genes begin near their optimal state, balanced by occasional mildly deleterious mutations that tend to be removed over time, and do not undergo significant adaptation or compensatory covariation over the timescale of interest. It was further assumed that in this equilibrium state, the mutation-selection process was time-reversible. These assumptions allowed the formulation of fixation probabilities in terms of simpler and more biologically meaningful parameters through the relation $\frac{P_{\text{fix}}(a,b)}{P_{\text{fix}}(b,a)} = \frac{(\pi_b p_b a)}{(\pi_a p_a b)}$, and permitted efficient estimation of phylogenetic relationships via maximum likelihood.

Assuming that sequences were at equilibrium suited the original purpose of the models in reconstructing evolutionary distances from coding sequences. However, the assumptions are unrealistic, not only for genes that are likely to have been affected by directional selection, but also as a description of genes subject to compensatory processes (i.e., all protein-encoding genes where the folded structure is important). This renders the original models unsuitable for detecting adaptation or inferring site-wise changes in fitness. Nevertheless, the explicit modeling of the relationship between relative fitnesses and evolutionary rates makes an attractive foundation from which to address this problem, as it has roots in an underlying population genetic process. Applications aimed at inferring fitnesses from mutation-selection models build on older methods for detecting adaptation through estimating the mean ratio of the rates of non-synonymous and synonymous substitutions (dN/dS); site- and residue-specific fixation probabilities $P^i_{\text{fix}}(a,b)$ more closely model the processes giving rise to this quantity [34].

Especially when using mutation-selection models to detect adaptation, we must relax the assumptions of stationarity and reversibility. This requirement has several consequences for methods development. In the first instance, the detailed balance relation between fixation probabilities and equilibrium frequencies given above no longer holds in general. This formulation may still be used, but the interpretation of the parameters is altered. Following an event that modifies the relative fitnesses of amino acids at a set of similar sites, the distribution of amino acids at those sites will no longer equal the distribution at equilibrium. The parameters π_a no longer represent the distribution of observable quantities, but rather the stationary amino acid frequencies to which the new process will ultimately converge if not disturbed (the instantaneous equilibrium frequencies of the site). In practice, this means that an extra vector of parameters is required to represent amino acid frequencies at the root, since these cannot be identified with any parameters of the branch-specific models [35], assuming no part of the tree shares the selective process that generated the root sequence. If adaptive events are to be modeled along the tree, this will also require estimation of a new set of parameters following each event.

Since the rates of transition between amino acid pairs are described in terms of these parameters, these will also change over the tree, rendering the overall process inhomogeneous over time.

How changes over time in amino acid fitness profiles are modeled requires consideration. Time-inhomogeneous models of nucleotide evolution have a deep history in phylogenetics. These models have been extensively investigated in attempts to account for observed differences in nucleotide composition in the genomes of related species. The initial proposal allowed the application of an arbitrary transition matrix to each branch of the tree [36]. This model suffers from the theoretical issue that it is not identifiable in the most general case; that is, no amount of sequence data will allow a unique estimate of the nucleotide frequencies or transition probabilities unless certain conditions are placed on the model [37].

The identifiability problem that has received the most attention is that the fully general model cannot discriminate among different labellings of the states at internal nodes [38]. This appears to be a DNA-specific problem. For DNA, if the frequencies of nucleotide A at each node are swapped for the frequencies of C, and the transition probabilities P_{Aj} and P_{Cj} are swapped for each destination nucleotide j on each branch of the tree, the new model will give the same joint distribution of nucleotides at the leaves. This scenario is extremely unlikely when fitting models designed for protein-coding sequences because the genetic code dictates transition probabilities among codons that are then affected by the equilibrium fitnesses and corresponding fitnesses, meaning that the rows of the transition matrix cannot be swapped. Aside from this, additional factors which may raise identifiability concerns include trees that have nodes of degree two (breakpoints in the middle of branches for example) and cases in which more than one instantaneous rate matrix can generate the same branch transition probabilities [37,39].

Assuming that these theoretical concerns can be addressed, the large number of parameters involved in nonhomogeneous models means that estimation can be difficult in practice. An early application attempted to account for compositional differences among related species by retaining a global relative rate matrix (for example, dictated by the genetic code) and only allowing base frequencies to vary among branches [40]. In principle, this method is well justified, as the mutational process is generally conserved and is distinct from the selective process that gives rise to different frequencies. However, it remains computationally demanding and prone to over-fitting [41]. Accordingly, subsequent methods made efforts to reduce the computational burden through using a single composition parameter representing GC content (or HP, hydrophobic and polar amino acids, in a protein context if needed) or otherwise reducing the parameter space [35,42] or by selecting from a smaller number of models that are shared across multiple branches of the tree [43,44].

A major innovation was the development of the break point (BP) model [41]. This type of model allows shifts in selective pressures to occur at any point in the tree, not only at speciation, duplication, or lateral transfer event nodes [45]. Like their forebears, the breakpoint models are not designed to detect adaption or infer values related to fitness or selective pressure. They do not explicitly model fixation probabilities or variable transition rates, and effectively integrate over break point numbers and positions. However, they remain some of the most complete methods for modeling non-stationary evolution. While it may not be critical to change models over a branch, when averaging a process over a branch, an important consideration is if the model that is fit over the branch reflects the combination of processes and associated parameters accurately [46]. A comparison of each of these types of nonstationary non-time-homogenous models is given in Figure 1.

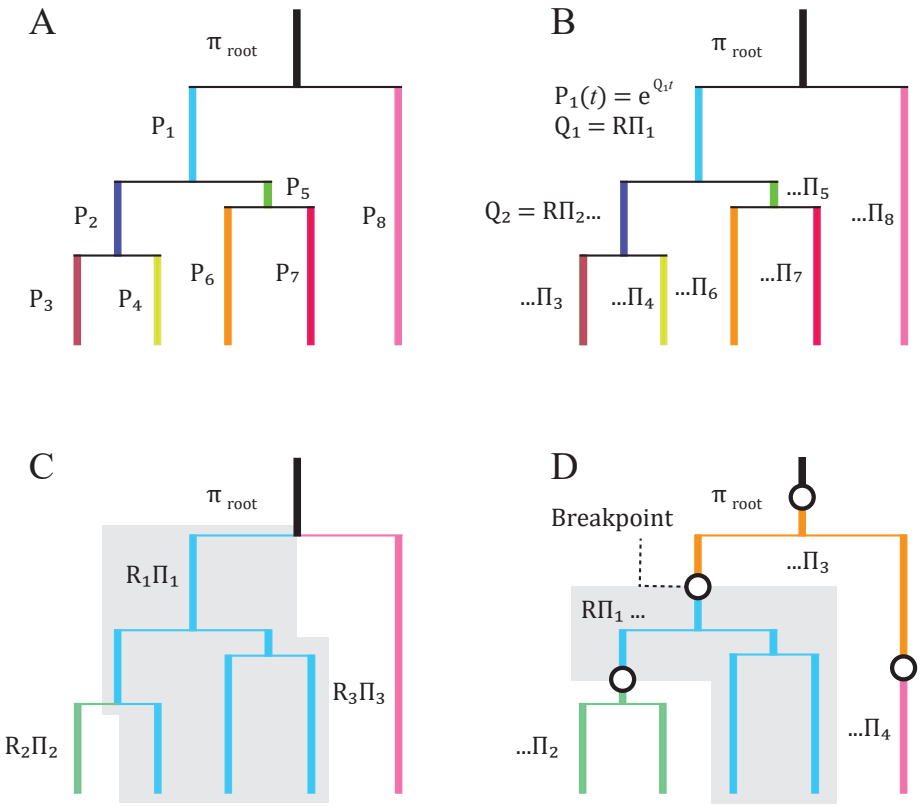

Figure 1. Illustration of the variety of nonstationary and non-time-homogeneous phylogenetic models. (**A**) The Barry-Hartigan model is the most general possible Markov substitution model. The substitution process is modeled by an arbitrary Markov transition matrix for each branch (P_1–P_8, indicated by branch colors) and a vector of initial frequencies at the root π_{root}. (**B**) A homogeneous but nonstationary model. Transition matrices are derived from a reversible continuous-time Markov model. All branches share the same rate matrix R, but the stationary frequencies Π_i are permitted to vary across the tree. (**C**) A non-homogeneous model reflecting more recent methods whereby a small number of reversible Markov models with different rates and frequencies are assigned to multiple branches within the tree (shaded region). (**D**) A nonstationary 'breakpoint' model in which state frequencies may differ within as well as among lineages.

Another consequence of relaxing the stationarity assumption is that standard phylogenetic optimization methods make strong use of time reversibility to allow efficient computation of model likelihoods over a tree. Given a time-reversible model of sequence evolution, the root may be considered as being anywhere on the tree without altering the likelihood. Using the standard phylogenetic likelihood algorithm [47], partial likelihoods for each state at the head of each subtree can be calculated recursively and stored. During optimization, the tree likelihood is recalculated after an update that alters the branch lengths or the tree topology, and the root of the tree can be placed so that only a minimum number of partial likelihoods need to be recomputed.

Under a non-reversible model, the position of the root cannot be altered, meaning that recalculation after an update will require more computation [40]. Solutions to this problem require caching an additional upper partial likelihood for each branch in the tree, which minimizes recomputation after branch length optimization and topology updates through nearest-neighbor interchange [48]. If the model includes break points, these can be considered extra nodes with a single descendant subtree and the likelihood calculated as normal, allowing for the new amino acid fitness values (Figure 2). Update mechanisms that add, delete or move break points or change frequencies within a lineage will also require recomputation of likelihoods, sometimes over several

nodes, although this is mitigated in BP models by ensuring that frequencies before and after a break point are independent. Models designed for reconstructing phylogenies typically include topology update mechanisms, including moves that the change the position of the root [41]; however, due to the highly parametric nature of time-variable mutation-selection models for amino acid data and the additional complexity created by non-stationarity, it is likely that practical methods in a maximum likelihood framework will need to rely on fixed topologies in the near future. This type of method is appropriate where traditional Markov substitution models can be trusted to estimate the proper topology with subsequent branch length optimization under the new model. Branch lengths estimated under a nonstationary model do not have the standard interpretation relating to the average substitutions per site but can be rescaled so as to give this interpretation [39,49].

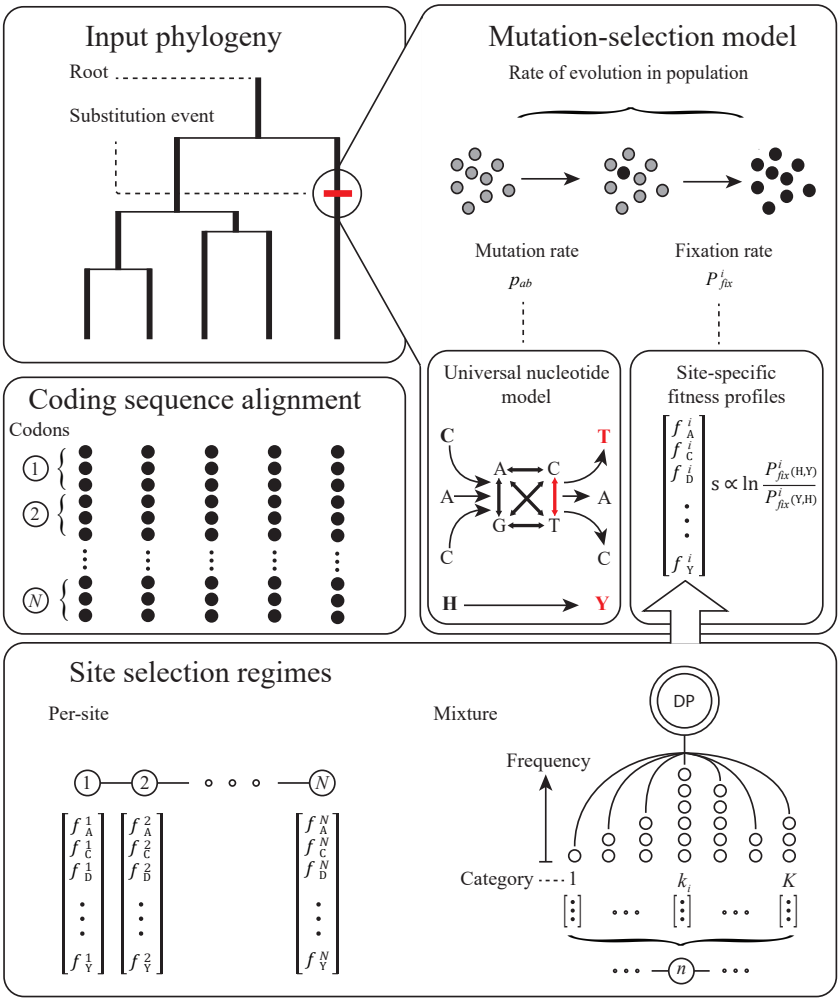

Figure 2. Schematic showing the construction of mutation-selection models in an equilibrium framework. Clockwise from top left: the model describes the DNA substitution process within a protein-coding sequence along a rooted phylogenetic tree. Each substitution is modeled by the product of a mutation rate and the rate of fixation within a population. The rate of mutation among codons is represented by the product of transition rates at each codon position. The fixation rate is represented by a population genetic model operating on a selection coefficient derived from a set of site-specific amino acid fitness values (AA values shown as single-letter codes). Fitness values may be treated as distinct parameters for each codon site, or probabilities may be calculated via a mixture over an assumed or unknown number of site categories (random effects). In a Bayesian framework the prior distribution of an unknown set of site categories may be given by a Dirichlet Process (DP).

5. Biochemical and Population Genetic Assumptions

While numerous advances in the complexity of mutation-selection models continue to be proposed, they all make a series of assumptions. The original Halpern and Bruno model [1] assumes that most positions are under purifying selection, implying that there are only a few amino acids have comparatively high fitness values. Further, it is assumed that this selective pressure is constant over time (reflecting diversifying rather than directional selection for positive s values). Therefore, these assumptions results in lack of consideration of site-interdependence. However, sites do not evolve in a vacuum and the fitness of sites are indeed interdependent for numerous reasons.

At the level of codons, codon usage can vary dramatically across a gene. Position-dependent codon usage bias has been observed in numerous taxa [50–52] and individual codon usage biases follow a position-dependent exponential decay model [53]. It has been theorized that codon that translated less efficiently are present in the first 90–150 nucleotides to create a "ramp" to slow elongation rates and prevent ribosomal traffic jams [54]. Whatever the underlying cause of site dependent codon usage, selection on synonymous codon usage may vary depending on the translational efficiency of nearby codons. Additionally, messenger RNA (mRNA) translation rates are coupled to proper protein folding and function [55]. As mentioned in the original paper [1], mutation-selection models with site-specific selection parameters are in principle capable of modeling these processes by decomposing fitness parameters into codon and amino acid components [26]. In practice, the framework has been adapted to study codon usage by applying fitness effects to synonymous as well as nonysynonymous changes [10,56], and these have been demonstrated to outperform simpler codon models on some data sets [57,58].

Accounting for protein structure leads to yet another level of site interdependence. In order for a protein to function it must be able to fold into a stable conformation. The stability of a folded protein (ΔG) depends on the difference in free energy between the native and unfolded forms of the protein. Further, proteins do not exist in isolation and their function often depends on the protein's ability to bind other proteins specifically. The stability of binding can be quantified in a similar manner to that of protein stability, by measuring the difference in free energy between bound and unbound and non-specifically bound proteins. While these measures can be combined into a single fitness metric [27,59] they are still considered independently. Further, it is unclear how to appropriately weight fitness contributions of binding and stability. Another issue arises when considering these sorts of metrics as measures of fitness over long evolutionary time periods, as significantly divergent sequences are not guaranteed to fold in a specific set structure. Hence, the use of ΔG as a fitness metric has limited ability to quantify the relationship between site-interdependence and selection [4]. The environment around a protein offers a further level of selective pressure on sites, including the chemical make up the surrounding solvent, protein concentration, temperature, and the presence of off-target interacting partners [4,60,61].

Beyond these physical aspects that result in site interdependence, there is a temporal aspect as well. Epistatic interactions result in changes to a site's fitness landscape as other interacting sites are modified. The fitness landscapes at these sites tend to change to stabilize the current state [62,63], a phenomenon referred to as entrenchment or an evolutionary Stokes shift. The constant shifting of site specific fitness landscapes imposed by interacting sites will tend to result in underestimations of dN/dS [34].

In addition to the biochemical constraints that result in site interdependence, there are also population level processes at play. Mutation-selection models have two components, a mutational probability and a fixation probability. The fixation probability can be boiled down to two key parameters, the effective population size (N_e) and the selection coefficient (s). The standard treatment from the model makes specific assumptions about the nature of both N_e and s. Phylogenetic implementations of mutation-selection models treat substitutions between amino acids at a site as being constant over the length of a branch. In some cases, this will lead to an averaging effect of selective strengths. This time averaging effect is equally acute for effective population sizes as it is for selective coefficients. Further, there are multiple effective population size parameters that matter, as the introduction of new mutations behaves differently from the fixation of mutations and occurs

on different time scales for the same proposed mutation [64]. Lastly, the modeling framework makes a strong assumption about weak mutational processes and the dynamic can be very different when mutational pressure is strong [14].

6. Conclusions

Though significant progress has been made towards using mutation-selection models to quantify selection on protein sequences, these models still suffer from a number of shortcomings. The basic framework of the model necessitates a set of biologically unrealistic assumptions. While many have introduced methods to account for certain aspects of violating these assumptions, further work towards constructing realistic models of evolution and using these models to quantify selection is necessary. However, constructing a modeling framework that accounts for a lack of evolutionary equilibrium as well as the biochemical and population level influences on protein evolution is a non-trivial task. Accounting for non-equilibrium processes is necessary for characterizing positive directional selection. Other assumptions about the nature of selection on protein structure and function, including compensatory amino acid substitution, and on the complexity of the underlying population genetics that are not being modeled await further modeling to determine the magnitude of incorrect inference. Lastly, branch averaging model mis-specification effects for heterogeneous processes appear both difficult to account for and to be small in affect, although averaging of negative and positive selection along a branch leads to reduced power (to detect positive selection) with increased branch length. Even with large data sets, complex models struggle with various parameter estimation problems and the ultimate solution may involve finding the right balance of biologically meaningful parameters and assumptions that is tractable [61,65].

Funding: A.I.T. and C.O.W. were supported by National Institutes of Health Grant R01 GM088344 and National Science Foundation Cooperative agreement no. DBI-0939454 (BEACON Center). A.M.R. and D.A.L. were supported by National Science Foundation Grant DBI-1515704.

Acknowledgments: The authors wish to thank A.J.H. for helpful discussions regarding codon bias.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Halpern, A.L.; Bruno, W.J. Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies. *Mol. Biol. Evol.* **1998**, *15*, 910–917. [CrossRef] [PubMed]
2. Yang, Z. *Computational Molecular Evolution*; Oxford University Press: Oxford, UK, 2006.
3. O'Brien, J.D.; Minin, V.N.; Suchard, M.A. Learning to count: Robust estimates for labeled distances between molecular sequences. *Mol. Biol. Evol.***2009**, *26*, 801–814. [CrossRef] [PubMed]
4. Chi, P.B.; Liberles, D.A. Selection on protein structure, interaction, and sequence. *Protein Sci.* **2016**, *25*, 1168–1178. [CrossRef] [PubMed]
5. Alberch, P. From genes to phenotype: dynamical systems and evolvability. *Genetica* **1991**, *84*, 5–11. [CrossRef] [PubMed]
6. Goldman, N.; Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **1994**, *11*, 725–736. [PubMed]
7. Muse, S.V.; Gaut, B.S. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* **1994**, *11*, 715–724. [PubMed]
8. Thorne, J.L.; Lartillot, N.; Rodrigue, N.; Choi, S.C. Codon models as a vehicle for reconciling population genetics with inter-specific sequence data. In *Codon Evolution: Mechanisms and Models*; Oxford University Press: Oxford, UK, 2012; pp. 97–110.
9. Golding, B.; Felsenstein, J. A maximum likelihood approach to the detection of selection from a phylogeny. *J. Mol. Evol.* **1990**, *31*, 511–523. [CrossRef] [PubMed]
10. Yang, Z.; Nielsen, R. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol. Biol. Evol.* **2008**, *25*, 568–579, doi:10.1093/molbev/msm284. [CrossRef] [PubMed]

11. Kimura, M. On the probability of fixation of mutant genes in a population. *Genetics* **1962**, *47*, 713–719. [PubMed]
12. Sella, G.; Hirsh, A. The application of statistical physics to evolutionary biology. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 9541–9546, doi:10.1073/pnas.0501865102. [CrossRef] [PubMed]
13. Krukov, I.; de Sanctis, B.; de Koning, A.P.J. Wright–Fisher exact solver (WFES): Scalable analysis of population genetic models without simulation or diffusion theory. *Bioinformatics* **2017**, *33*, 1416–1417, doi:10.1093/bioinformatics/btw802. [CrossRef] [PubMed]
14. De Koning, A.J.; De Sanctis, B.D. The rate of observable molecular evolution when mutation may not be weak. *bioRxiv* **2018**, 259507. [CrossRef]
15. Jones, D.T. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences1. *J. Mol. Biol.* **1999**, *287*, 797–815. [CrossRef] [PubMed]
16. Robinson, D.M.; Jones, D.T.; Kishino, H.; Goldman, N.; Thorne, J.L. Protein evolution with dependence among codons due to tertiary structure. *Mol. Biol. Evol.* **2003**, *20*, 1692–1704. [CrossRef] [PubMed]
17. Rodrigue, N.; Lartillot, N.; Bryant, D.; Philippe, H. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene* **2005**, *347*, 207–217. [CrossRef] [PubMed]
18. Rodrigue, N.; Kleinman, C.L.; Philippe, H.; Lartillot, N. Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. *Mol. Biol. Evol.* **2009**, *26*, 1663–1676. [CrossRef] [PubMed]
19. Arenas, M.; Dos Santos, H.G.; Posada, D.; Bastolla, U. Protein evolution along phylogenetic histories under structurally constrained substitution models. *Bioinformatics* **2013**, *29*, 3020–3028. [CrossRef] [PubMed]
20. Arenas, M.; Weber, C.C.; Liberles, D.A.; Bastolla, U. ProtASR: An evolutionary framework for ancestral protein reconstruction with selection on folding stability. *Syst. Biol.* **2017**, *66*, 1054–1064. [CrossRef] [PubMed]
21. Arenas, M.; Sánchez-Cobos, A.; Bastolla, U. Maximum-likelihood phylogenetic inference with selection on protein folding stability. *Mol. Biol. Evol.* **2015**, *32*, 2195–2207. [CrossRef] [PubMed]
22. De Koning, A.J.; Gu, W.; Pollock, D.D. Rapid likelihood analysis on large phylogenies using partial sampling of substitution histories. *Mol. Biol. Evol.* **2009**, *27*, 249–265. [CrossRef] [PubMed]
23. Spielman, S.J.; Wilke, C.O. The relationship between *dN/dS* and scaled selection coefficients. *Mol. Biol. Evol.* **2015**, *32*, 1097–1108. [CrossRef] [PubMed]
24. Rodrigue, N.; Philippe, H.; Lartillot, N. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 4629–4634. [CrossRef] [PubMed]
25. Rodrigue, N.; Lartillot, N. Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. *Bioinformatics* **2013**, *30*, 1020–1021. [CrossRef] [PubMed]
26. Tamuri, A.U.; dos Reis, M.; Goldstein, R.A. Using site-wise mutation-selection models to estimate the distribution of selection coefficients from phylogenetic data. *Genetics* **2011**, 111. [CrossRef]
27. Grahnen, J.A.; Nandakumar, P.; Kubelka, J.; Liberles, D.A. Biophysical and structural considerations for protein sequence evolution. *BMC Evol. Biol.* **2011**, *11*, 361. [CrossRef] [PubMed]
28. Rodrigue, N. On the statistical interpretation of site-specific variables in phylogeny-based substitution models. *Genetics* **2012**. [CrossRef] [PubMed]
29. Tamuri, A.U.; Goldman, N.; dos Reis, M. A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. *Genetics* **2014**, *197*, 257–271. [CrossRef] [PubMed]
30. Spielman, S.J.; Wilke, C.O. Extensively parameterized mutation–selection models reliably capture site-specific selective constraint. *Mol. Biol. Evol.* **2016**, *33*, 2990–3002. [CrossRef] [PubMed]
31. Bloom, J.D. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Mol. Biol. Evol.* **2014**, *31*, 1956–1978. [CrossRef] [PubMed]
32. Bloom, J.D. An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs. *Mol. Biol. Evol.* **2014**, *31*, 2753–2769. [CrossRef] [PubMed]
33. Bloom, J.D. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biol. Direct* **2017**, *12*. [CrossRef] [PubMed]
34. Rodrigue, N.; Lartillot, N. Detecting adaptation in protein-coding genes using a Bayesian site-heterogeneous mutation-selection codon substitution model. *Mol. Biol. Evol.* **2017**, *34*, 204–214, doi:10.1093/molbev/msw220. [CrossRef] [PubMed]

35. Galtier, N.; Gouy, M. Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* **1998**, *15*, 871–879, doi:10.1093/oxfordjournals.molbev.a025991. [CrossRef] [PubMed]
36. Barry, D.; Hartigan, J.A. Statistical analysis of hominoid molecular evolution. *Stat. Sci.* **1987**, *2*, 191–207, doi:10.1214/ss/1177013353. [CrossRef]
37. Chang, J.T. Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. *Math. Biosci.* **1996**, *137*, 51–73, doi:10.1016/s0025-5564(96)00075-2. [CrossRef]
38. Zou, L.; Susko, E.; Field, C.; Roger, A.J. The parameters of the Barry and Hartigan general Markov model are statistically nonIdentifiable. *Syst. Biol.* **2011**, *60*, 872–875, doi:10.1093/sysbio/syr034. [CrossRef] [PubMed]
39. Kaehler, B.D.; Yap, V.B.; Zhang, R.L.; Huttley, G.A. Genetic distance for a general non-stationary Markov substitution process. *Syst. Biol.* **2015**, *64*, 281–293, doi:10.1093/sysbio/syu106. [CrossRef] [PubMed]
40. Yang, Z.; Roberts, D. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol. Biol. Evol.* **1995**, *12*, 451–458, doi:10.1093/oxfordjournals.molbev.a040220. [CrossRef] [PubMed]
41. Blanquart, S.; Lartillot, N. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol. Biol. Evol.* **2006**, *23*, 2058–2071, doi:10.1093/molbev/msl091. [CrossRef] [PubMed]
42. Groussin, M.; Boussau, B.; Gouy, M. A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. *Syst. Biol.* **2013**, *62*, 523–538, doi:10.1093/sysbio/syt016. [CrossRef] [PubMed]
43. Foster, P. Modeling compositional heterogeneity. *Syst. Biol.* **2004**, *53*, 485–495, doi:10.1080/10635150490445779. [CrossRef] [PubMed]
44. Gowri-Shankar, V.; Rattray, M. A reversible jump method for Bayesian phylogenetic inference with a nonhomogeneous substitution model. *Mol. Biol. Evol.* **2007**, *24*, 1286–1299, doi:10.1093/molbev/msm046. [CrossRef] [PubMed]
45. Blanquart, S.; Lartillot, N. A site- and time-heterogeneous model of amino acid replacement. *Mol. Biol. Evol.* **2008**, *25*, 842–858, doi:10.1093/molbev/msn018. [CrossRef] [PubMed]
46. Shore, J.A.; Sumner, J.G.; Holland, B.R. Closed codon models: Just a hopeless dream? *arXiv* **2018**, arXiv:1804.11249.
47. Felsenstein, J. Evolutionary trees from DNA-sequences—A maximum-likelihood approach. *J. Mol. Evol.* **1981**, *17*, 368–376, doi:10.1007/Bf01734359. [CrossRef] [PubMed]
48. Boussau, B.; Gouy, M. Efficient likelihood computations with nonreversible models of evolution. *Syst. Biol.* **2006**, *55*, 756–768, doi:10.1080/10635150600975218. [CrossRef] [PubMed]
49. Zou, L.W.; Susko, E.; Field, C.; Roger, A.J. Fitting nonstationary general-time-reversible models to obtain edge-lengths and frequencies for the Barry-Hartigan model. *Syst. Biol.* **2012**, *61*, 927–940, doi:10.1093/sysbio/sys046. [CrossRef] [PubMed]
50. Goodman, D.B.; Church, G.M.; Kosuri, S. Causes and effects of N-terminal codon bias in bacterial genes. *Science* **2013**, 1241934. [CrossRef] [PubMed]
51. Bentele, K.; Saffert, P.; Rauscher, R.; Ignatova, Z.; Blüthgen, N. Efficient translation initiation dictates codon usage at gene start. *Mol. Syst. Biol.* **2013**, *9*, 675. [CrossRef] [PubMed]
52. Qin, H.; Wu, W.B.; Comeron, J.M.; Kreitman, M.; Li, W.H. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics* **2004**, *168*, 2245–2260. [CrossRef] [PubMed]
53. Hockenberry, A.J.; Sirer, M.I.; Amaral, L.A.N.; Jewett, M.C. Quantifying position-dependent codon usage bias. *Mol. Biol. Evol.* **2014**, *31*, 1880–1893. [CrossRef] [PubMed]
54. Tuller, T.; Carmi, A.; Vestsigian, K.; Navon, S.; Dorfan, Y.; Zaborske, J.; Pan, T.; Dahan, O.; Furman, I.; Pilpel, Y. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **2010**, *141*, 344–354. [CrossRef] [PubMed]
55. Spencer, P.S.; Barral, J.M. Genetic code redundancy and its influence on the encoded polypeptides. *Comput. Struct. Biotechnol. J.* **2012**, *1*, e201204006. [CrossRef] [PubMed]
56. Pouyet, F.; Bailly-Bechet, M.; Mouchiroud, D.; Guéguen, L. SENCA: A multilayered codon model to study the origins and dynamics of codon usage. *Gen. Biol. Evol.* **2016**, *8*, 2427–2441, doi:10.1093/gbe/evw165. [CrossRef] [PubMed]
57. Rodrigue, N.; Lartillot, N.; Philippe, H. Bayesian comparisons of codon substitution models. *Genetics* **2008**, *180*, 1579–1591, doi:10.1534/genetics.108.092254. [CrossRef] [PubMed]

58. Rodrigue, N.; Philippe, H. Mechanistic revisions of phenomenological modeling strategies in molecular evolution. *Trend. Genet.* **2010**, *26*, 248–252, doi:10.1016/j.tig.2010.04.001. [CrossRef] [PubMed]
59. Kachroo, A.H.; Laurent, J.M.; Yellman, C.M.; Meyer, A.G.; Wilke, C.O.; Marcotte, E.M. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science* **2015**, *348*, 921–925. [CrossRef] [PubMed]
60. Liberles, D.A.; Tisdell, M.D.; Grahnen, J.A. Binding constraints on the evolution of enzymes and signalling proteins: The important role of negative pleiotropy. *Proc. R. Soc. Lond. B Biol. Sci.* **2011**. [CrossRef] [PubMed]
61. Echave, J.; Wilke, C.O. Biophysical models of protein evolution: understanding the patterns of evolutionary sequence divergence. *Ann. Rev. Biophys.* **2017**, *46*, 85–103. [CrossRef] [PubMed]
62. Pollock, D.D.; Thiltgen, G.; Goldstein, R.A. Amino acid coevolution induces an evolutionary Stokes shift. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, E1352–E1359. [CrossRef] [PubMed]
63. Shah, P.; McCandlish, D.M.; Plotkin, J.B. Contingency and entrenchment in protein evolution under purifying selection. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E3226–E3235. [CrossRef] [PubMed]
64. Platt, A.; Weber, C.C.; Liberles, D.A. Protein evolution depends on multiple distinct population size parameters. *BMC Evol. Biol.* **2018**, *18*, 17. [CrossRef] [PubMed]
65. Liberles, D.A.; Teufel, A.I.; Liu, L.; Stadler, T. On the need for mechanistic models in computational genomics and metagenomics. *Gen. Biol. Evol.* **2013**, *5*, 2008–2018. [CrossRef] [PubMed]

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).