

A mutation-selection model of protein evolution under persistent positive selection

Asif U. Tamuri,^{1,2} and Mario dos Reis^{*,3}

¹Centre for Advanced Research Computing, University College London, Gower St, London WC1E 6BT, UK

²EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

³School of Biological and Behavioural Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, UK

*Corresponding author: E-mail: m.dosreisbarros@qmul.ac.uk

Associate Editor:

Abstract

We use first principles of population genetics to model the evolution of proteins under persistent positive selection (PPS). PPS may occur when organisms are subjected to persistent environmental change, during adaptive radiations, or in host-pathogen interactions. Our mutation-selection model indicates protein evolution under PPS is an irreversible Markov process, and thus proteins under PPS show a strongly asymmetrical distribution of selection coefficients among amino acid substitutions. Our model shows the criteria $\omega > 1$ (where ω is the ratio of non-synonymous over synonymous codon substitution rates) to detect positive selection is conservative and indeed arbitrary, because in real proteins many mutations are highly deleterious and are removed by selection even at positively-selected sites. We use a penalized-likelihood implementation of the PPS model to successfully detect PPS in plant RuBisCO and influenza HA proteins. By directly estimating selection coefficients at protein sites, our inference procedure bypasses the need for using ω as a surrogate measure of selection and improves our ability to detect molecular adaptation in proteins.

Key words: Positive selection, Distribution of fitness effects, Influenza, RuBisCO, Cytochrome B, Mutation-selection model

Introduction

Understanding how natural selection acts on molecular sequences has long been a pursuit of evolutionary biology. For example, Kimura (1983), using a model that assumes the genome has an infinite number of sites, showed the relative rate of molecular evolution is approximately given by

$$u = \frac{S}{1 - e^{-S}}, \quad (1)$$

where S is the selection coefficient acting on mutations. If new mutations in the genome are positively selected ($S > 0$) the relative rate of molecular evolution is accelerated ($u > 1$), while the rate is the neutral

mutation rate ($u=1$) if there is no selection ($S=0$), and the rate is decelerated ($u<1$) if mutations are negatively selected ($S<0$).

Equation 1, which is the relative probability of fixation of selected over neutral mutations (Fisher, 1930; Kimura, 1962; McCandlish *et al.*, 2015; Wright, 1931), has important implications for understanding molecular adaptation in proteins. For a sample of protein-coding sequences from various species, the ratio between the number of substitutions at non-synonymous sites (which are under selection) and at synonymous sites (which are under weak or no selection) should approximately follow the dynamics of Eq. 1 (Nielsen and Yang, 2003). This ratio, commonly known as $\omega=d_N/d_S$, is widely used as a test of molecular adaptation in proteins, with $\omega>1$, $\omega=1$, and $\omega<1$ interpreted as evidence of molecular adaptation (positive selection), neutral evolution, and purifying selection respectively.

However, Kimura's relative rate of molecular evolution (Eq. 1), based on the infinite-sites model (Kimura, 1969; Sawyer and Hartl, 1992), assumes all new mutations appear at new sites in the genome. This assumption appears unrealistic for proteins. Nielsen and Yang (2003) have argued that if amino acid fitnesses are re-assigned every time a new mutation appears at a site in a protein (so that the selection coefficient, S , is always the same at the site), then Eq. 1 gives the relationship of S and ω under a finite-sites model. However, it is not clear in which condition this fitness reassignment should apply: If an i to j mutation has selection coefficient S , then the reverse j to i mutation should have coefficient $-S$, but Nielsen and Yang's model assumes it reverts to S . Without this assumption it does not appear possible to equate $\omega=S/(1-e^{-S})$.

Spielmann and Wilke (2015), and dos Reis (2015), used the Fisher-Wright mutation-selection model (Fisher, 1930; Halpern and Bruno, 1998; Wright, 1931) to derive the relationship between ω and the selection coefficients acting on codon sites within a finite-sites model. They showed that $\omega\leq 1$ when selection coefficients are constant over time, i.e. they are not reassigned, (dos Reis, 2015; Spielman and Wilke, 2015); whereas $\omega>1$ can be achieved for a short period of time after selection coefficients undergo a single shift during an adaptive event, for example, when a virus adapts to a new host (dos Reis, 2015).

However, the relationship between ω and selection coefficients under the more general case of persistent changes in selection over time appears unclear. This case, which we term persistent positive selection (PPS), is important because selection coefficients acting at codon sites may change repeatedly during persistent environmental changes, during adaptive radiations, and in host-pathogen interactions (such as in a virus evading herd immunity in a host population). Thus, understanding how PPS affects ω in proteins can inform the development of methods to detect positive selection and give us insight onto the mechanisms of adaptive evolution in general.

Here we develop a mutation-selection model of codon substitution under PPS. The new model can be used to study the mechanistic relationship between the scaled selection coefficients and ω , providing insight into the evolutionary dynamics of proteins under PPS. Furthermore, we develop a penalised-likelihood implementation of the model and successfully use it to detect PPS directly in real proteins bypassing the need to use ω as a surrogate measure of selection. Analysis under the new model indicates codon substitution is an irreversible Markov process, leading to a highly asymmetrical distribution of selection coefficients among substitutions in proteins under PPS. More strikingly, the PPS model shows the criteria $\omega > 1$ to detect molecular adaptation in proteins is conservative and indeed arbitrary, as we find evidence of PPS at codon sites where $\omega < 1$.

New Approaches

The PPS codon substitution model

We develop the new model by integrating the non-homogeneous selection model of Kimura and Ohta (1970) with the mutation-selection codon substitution model of Halpern and Bruno (1998). Consider a population of organisms with haploid genome number N . That is, the number of copies of the genome in the population is N (i.e., the population size is N if the organism is haploid and $N/2$ if it is diploid). Suppose a site k in a protein-coding gene is fixed for codon i in the population, and the scaled Malthusian fitness of i is $F_{i,k}$. A new mutant codon j appears at the site and has initial selective advantage $S_{ij,k}^* = F_{j,k}^* - F_{i,k}$, $F_{j,k}^* > F_{i,k}$. The selective advantage then decays exponentially as a function of time (Kimura and Ohta, 1970), for example, due to gradual environmental change. Kimura and Ohta (1970) showed that the fixation probability of j is approximately $S_{ij,k}/(1 - e^{-S_{ij,k}}) \times N^{-1}$ where $S_{ij,k}$ is constant and $0 < S_{ij,k} < S_{ij,k}^*$. In other words, the fixation probability of j is the same as that of an allele with intermediate, but constant, selective advantage $S_{ij,k}$.

It appears that other types of decay function lead to the same fixation probability. For example, the same result is obtained in the case of frequency-dependent selection (FDS) when the fitness of j decays exponentially as a function of the frequency of j in the population (dos Reis, 2013). In the case of FDS, once j becomes fixed, any new mutant alleles may have high fitnesses because they would be rare. We expect this type of dynamics in, for example, viruses escaping the herd immunity of a host population. Similarly, if the environment gradually shifts between two states, then the selective advantage of j or i would be continuously reset depending on the particular environment. This would then lead to re-setting (or re-assignment) of the fitnesses of i and j . This persistent change in the selection coefficient is what we term persistent positive selection (PPS). We formalize codon substitution under the PPS model next.

Let the selection coefficient for the $i \rightarrow j$ mutation be $S_{ij,k} = F_{j,k} - F_{i,k} + Z_k$, where $F_{j,k}$, $F_{i,k}$ and $Z_k (\geq 0)$ are constant. Let the selection coefficient for the reverse mutant, $j \rightarrow i$, be $S_{ji,k} = F_{i,k} - F_{j,k} + Z_k$. In other words, we have partitioned the fitnesses of j and i into two components: A constant component, $F_{j,k}$ and $F_{i,k}$, representing structural constraints of the protein on the amino acid encoded by the codon; and Z_k , the PPS component. Thus, when $Z_k > 0$, the selection coefficient is persistently reset with new mutations.

The substitution rate from i to j at location k , $q_{ij,k}$, is equal to the neutral mutation rate, μ_{ij} , times the number of i alleles in the population, N , times the fixation probability of the j mutant (Halpern and Bruno, 1998; Kimura, 1983). Assuming synonymous substitutions are neutral, this gives

$$q_{ij,k} = \begin{cases} \mu_{ij} \frac{S_{ij,k}}{1 - e^{-S_{ij,k}}}, & \text{if the substitution is non-synonymous,} \\ \mu_{ij} & \text{otherwise.} \end{cases} \quad (2)$$

Irreversibility of codon substitution under PPS

Eq. 2 describes codon substitution as a continuous Markov process. Polymorphisms are ignored and the population is assumed to switch from i to j instantaneously. This assumption appears reasonable if $N\mu_{ij} \ll 1$, for all μ_{ij} (e.g. Bulmer 1991). The proportion of time location k remains fixed for j (i.e. the stationary frequency of j) is $\pi_{j,k}$. A Markov process is said to be reversible in equilibrium if it satisfies the detailed-balance condition $\pi_{i,k}q_{ij,k} = \pi_{j,k}q_{ji,k}$ (Grimmet and Stirzaker, 2004). When $Z_k = 0$, the model of Eq. 2 is reversible (Yang and Nielsen, 2008). However, when $Z_k > 0$ the process is, in general, irreversible because the detailed balance condition does not hold. When $Z_k > 0$, the stationary frequencies are found by solving the system of equations $\sum \pi_{j,k}q_{ji,k} - \sum \pi_{i,k}q_{ij,k} = 0$ with the constraint $\sum \pi_{i,k} = 1$. We calculate the irreversibility index for site k as $I_k = |\pi_{i,k}q_{ij,k} - \pi_{j,k}q_{ji,k}|$, where $I_k > 0$ indicates evolution at site k is irreversible, and $I_k = 0$ otherwise (Huelsenbeck *et al.*, 2002).

Identifying protein locations under PPS

Given an alignment of protein-coding genes with corresponding phylogeny, the model of Eq. 2 can be used to estimate the $F_{i,k}$ and Z_k using maximum penalised likelihood. To estimate the $F_{i,k}$, we use the Dirichlet-based penalty of Tamuri *et al.* (2014) and for Z_k we use an exponential penalty with parameter λ (see Methods). For each site in the alignment, we compare the null model $Z_k = 0$ (no PPS) against $Z_k > 0$ (PPS) using a likelihood-ratio test. Because of the boundary condition ($Z_k > 0$) in the test and the use of penalised likelihood, the distribution of the likelihood-ratio statistic does not follow the typical χ^2 distribution. Thus, we use Cox (1962) simulation approach as used in phylogenetics (Goldman, 1993) to obtain the appropriate null distribution (see Methods).

The relationship between selection coefficients and ω

The average substitution rate of codon site k , averaged over time is

$$\rho_k = \sum_{i \neq j} \pi_{i,k} q_{i,j,k}.$$

This rate can be separated into its non-synonymous and synonymous components, $\rho_k = \rho_{N,k} + \rho_{S,k}$, where

$$\rho_{N,k} = \sum_{i \neq j} \pi_{i,k} q_{i,j,k} \mathbf{I}_N \text{ and } \rho_{S,k} = \sum_{i \neq j} \pi_{i,k} q_{i,j,k} (1 - \mathbf{I}_N),$$

and where the indicator function $\mathbf{I}_N = 1$ if the i to j substitution is non-synonymous, and $= 0$ otherwise.

For a neutrally evolving sequence (e.g. a pseudogene) the corresponding rates are

$$\rho_N^{(0)} = \sum_{i \neq j} \pi_i^{(0)} \mu_{ij} \mathbf{I}_N \text{ and } \rho_S^{(0)} = \sum_{i \neq j} \pi_i^{(0)} \mu_{ij} (1 - \mathbf{I}_N),$$

where $\pi_i^{(0)}$ is the stationary frequency of i without selection, which is the same for all sites. Thus, the relative non-synonymous rate is

$$\omega_k = \frac{\rho_{N,k}}{\rho_N^{(0)}}. \quad (3)$$

See dos Reis (2015) for the full derivation. Spielman and Wilke (2015) give a slightly different definition of ω_k (see also Jones *et al.* 2016; Youssef *et al.* 2020).

We note the PPS model is general and has other models as special cases. For example, when $Z_k \neq 0$ and $F_{i,j} = F_{j,i}$ for all i, j , we have

$$\omega_k = \frac{Z_k}{1 - e^{-Z_k}},$$

and the model of Eq. 2 can be written as $q_{i,j} = \mu_{ij} \omega_k$ if the substitution is non-synonymous and $q_{i,j} = \mu_{ij}$ otherwise. In other words, the classic codon models (Muse and Gaut, 1994; Yang and Nielsen, 1998) are a special case of Eq. 2 when all codons are assumed to have the same fitness. On the other hand, when $Z_k = 0$ and $F_{i,j} \neq F_{j,i}$, the model of Eq. 2 reduces to the mutation-selection model of Halpern and Bruno (1998).

Results

Detection of PPS in simulated data

Extensive simulations on the estimation of $F_{i,k}$ are available in Tamuri *et al.* (2012; 2014). Here our focus is on using simulations to assert whether sites under PPS ($Z_k > 0$) can be identified using Cox's method. We simulate codon alignments (1,000 codons in length) on a 512-taxa phylogeny, under various strengths of PPS, with $Z_k = 0, 2, 5$, and 10. The values of $F_{i,j,k}$ are drawn from random distributions to produce sharp amino acid profiles as in real proteins (see Methods). These $F_{i,k}$ and Z_k values result in ω_k values roughly between 0.05 and 6 (Eq. 3). When $Z_k = 0$, 6.6% of sites are incorrectly detected

Table 1. Performance of the LRT for detecting PPS sites in simulated data after FDR correction (5%)

True model	$\lambda=0.01$	$\lambda=0.5$	$\lambda=1.0$
FPR at 0.05 significance			
swMutSel ($Z=0$)	0.066	0.066	0.066
TPR at 0.05 significance			
swMutSel+PPS ($Z=2$)	0.441	0.452	0.449
($Z=5$)	0.952	0.952	0.947
($Z=10$)	0.965	0.963	0.960

FPR: False positive rate, TPR: True positive rate.

to be under PPS, which is slightly higher than the 5% error I threshold (Table 1). When the selective advantage is slight ($Z_k=2$), the method roughly identifies 44% of sites under PPS (Table 1). The power of the method is excellent and roughly over 95% when the selective advantage is strong ($Z_k \geq 5$). We note the exponential penalty on Z_k has a noticeable, albeit slight, effect on the power of the test. When the penalty parameter, λ is small, the resulting penalty is diffuse and the penalty is weak. However, as λ increases, the penalty becomes stronger with probability density in the exponential moving towards zero. In this case, estimates of $Z_k \gg 0$ are more strongly penalised and this translates in a small reduction in the power of the test (Table 1). We note the penalised likelihood method used here is essentially the same as posterior mode finding giving our penalties are proper probability densities (Cox and O’Sullivan, 1990; Tamuri *et al.*, 2014), and thus the penalties on Z_k and $F_{i,k}$ act as prior densities which regularise the parameter estimates (Cox and O’Sullivan, 1990).

Detection of PPS in real proteins

We tested for PPS sites in three real sequence datasets: the haemagglutinin protein (HA) from human influenza H1N1 virus, the rbcL protein subunit from flowering plants, and the mitochondrial cytochrome b (CYTB) protein from mammals (Table 2). Given the multiple sequence alignment, phylogeny, and mutational parameters, we estimated the $F_{i,k}$ and Z_k using two penalty strengths, $\lambda=0.001$, and 0.05. We then performed the LRT of PPS vs no PPS and used false discovery rate (FDR) at the 5% level to identify sites under PPS. Using the weak penalty, $\lambda=0.001$, we detected PPS ($Z_k > 0$) at 65 sites in the plant rbcL and 18 sites in the influenza HA, but we found no PPS sites in mammal CYTB (Table 2). Interestingly, only 55 out 65 of PPS sites in rbcL have $\omega > 1$. For HA, all 18 PPS sites also have $\omega > 1$. The

Table 2. Number of sites estimated to be under PPS in three real datasets.

Dataset	# taxa	# sites	# $Z > 0$ ($\omega > 1$)	
			$\lambda = 0.001$	$\lambda = 0.05$
Plant rbcL	478	466	65 (55)	50 (40)
Influenza HA	466	589	18 (18)	17 (14)
Mammal CYTB	418	407	0 (0)	- (-)

location of PPS sites and estimated ω_k values are shown in Fig. 1A-A". When using the stronger penalty, $\lambda = 0.05$, the number of sites detected in rbcL and HA are reduced to 50 and 17 sites respectively (Table 2). This is not unexpected because, as noted above, stronger penalties push estimates of Z_k towards zero affecting the likelihood ratio test.

The distribution of selection coefficients at sites under PPS is asymmetrical

We estimated the distribution of selection coefficients among non-synonymous substitutions (Tamuri *et al.*, 2014) in the three protein-coding genes analysed (Fig. 1B-B"). For non-PPS sites (i.e. sites where $Z_k = 0$), the distribution of selection coefficients is symmetrical, with a mode at $S = 0$, because in this case codon substitution is reversible and the detailed balance condition guarantees the proportions of slightly advantageous and deleterious mutations fixed in the population will be equal over time (Yang and Nielsen, 2008). However, among PPS sites in plant rbcL and influenza HA, the distribution is highly skewed with a mode at $S > 10$ because irreversibility of the substitution process means the detailed balance condition does not apply, and hence there is a persistent excess of advantageous mutations being substituted into the population. For example, for sites with $Z_k \geq 10$, the irreversibility index indicates a deviation of up to 12% of substitutions from detailed balance, which is a strong deviation (Fig. 2a). Larger values of Z_k are also associated with faster substitution rates (Fig. 2b) and larger ω_k values (Fig. 2c). For example, for sites with $Z_k \geq 10$, the corresponding ω_k values range from about 1 to 4 (Fig. 2c).

PPS sites are under strong purifying constraints

At equilibrium, the average selection coefficient of new mutations at site k is

$$\bar{S}_k = \sum_{i \neq j} \pi_{i,k} P_{ij} S_{ij,k},$$

where $P_{ij} = \mu_{ij} / \sum_j \mu_{ij}$ is the probability that the next mutation is j given the site is currently fixed for i (dos Reis, 2015). If most new mutations are very deleterious, then the site is under purifying selection and $\bar{S}_k < 0$, while if most new mutations are advantageous the site is under diversifying selection and $\bar{S}_k > 0$. Historically, ω_k has been used as a proxy for \bar{S}_k , based on the approximation of Eq. 1 (Yang and

Nielsen, 2002). Thus calculating \bar{S}_k should provide insight into the relationship between the strength of selection at a site and ω_k .

Fig. 1C-C" shows the estimated \bar{S}_k for the three datasets plotted against ω_k . For 43 PPS sites in rbcL and one PPS site in HA, we find that $\bar{S}_k < 0$. This shows PPS sites are effectively under a mixture of purifying selection against deleterious amino acid substitutions, and diversifying selection in favor of a few amino acids that substitute rapidly among each other. This trend is evidenced when studying the pattern of PPS substitution in the influenza HA protein. The H1N1 influenza virus entered the human population sometime prior to the 1918 influenza pandemic (dos Reis *et al.*, 2009; Taubenberger *et al.*, 2005) and has remained largely as a single lineage since then, except from the introduction of a separate lineage of reassortant H1N1 swine virus in the 2009 pandemic (Smith *et al.*, 2009). Fig. 3 shows the pattern of amino acid substitution for the 18 PPS sites in influenza HA between 1918 and 2009. For example, site 3 remained virtually fixed for alanine between 1918 and the late 1990's, and then suffered several back and forth substitutions between alanine and valine between the late 1990's and 2009, while site 142 has been characterised by shifts between lysine and asparagine between 1918 and 2009. It's clear from Fig. 3 that the majority of PPS sites in the HA protein are characterised by back-and-forth substitutions among a fairly reduced set of amino acids.

Discussion

Mutation-selection models of codon substitution have been successfully used to study the distribution of selection coefficients in proteins (Rodrigue *et al.*, 2010; Tamuri *et al.*, 2014), to detect selection shifts during adaptation (Parto and Lartillot, 2017), shifting balance (Jones *et al.*, 2016), and to understand protein evolution given structural constraints (Youssef *et al.*, 2020). Previous works have also accommodated a ω parameter within the mutation-selection model to detect adaptation at amino acid sites (Rodrigue and Lartillot, 2017; Rodrigue *et al.*, 2021; Yang and Nielsen, 2008). However in these works ω is a separate parameter and not a function of the selection coefficients and thus its population genetics interpretation is not clear (Rodrigue and Lartillot, 2017). Here we extended the mutation-selection framework to the case of PPS without the need for the additional ω parameter. Instead, in the new model, ω is a function of the selection coefficients and we believe this modelling approach can help gain insight on the nature of protein adaptation.

The new PPS model is flexible as it appears to have performed well for the different modes of selection studied here. For example, rbcL is the major subunit of the RuBisCo enzyme responsible for the fixation of carbon during photosynthesis. The efficiency of RuBisCo is affected by environmental factors and rbcL

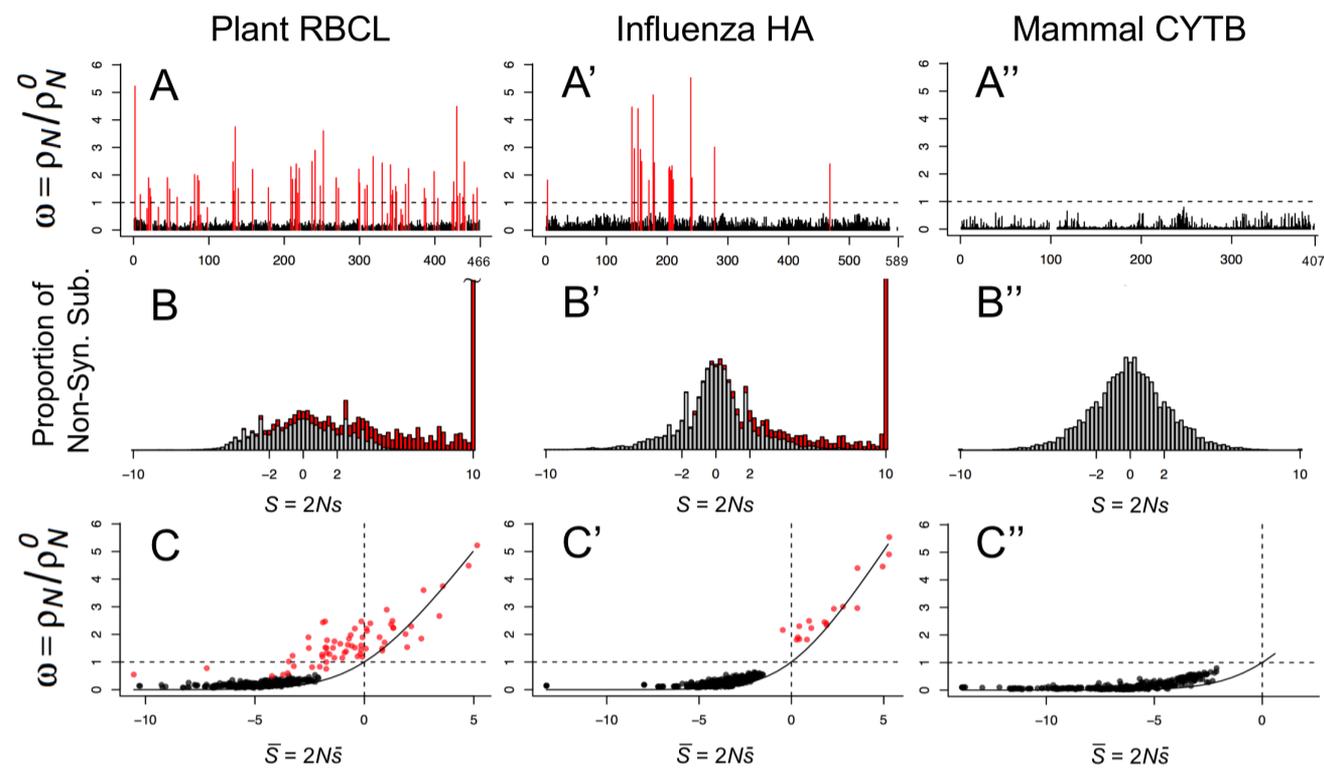


FIG. 1. Analysis of proteins under the PPS mutation-selection model. (A-A'') Estimates of ω at protein sites. (B-B'') Distribution of selection coefficients among non-synonymous substitutions. (C-C'') Relationship between ω and average selection (\bar{S}) at protein sites. Sites under PPS ($Z_k > 0$) are indicated in red in A-A'' and C-C'', and their contribution to the distribution of selection coefficients indicated in red in B-B''. In C-C'' the solid line is Eq. 1. The penalty on Z_k is $\lambda = 0.001$.

has been under persistent adaptive pressures during the successful adaptive radiation of angiosperms around the ecoregions of the world (Kapralov and Filatov, 2007; Parto and Lartillot, 2018). This is akin to the environment change model envisaged by Kimura and Ohta (1970). On the other hand, the influenza HA protein is the classic example of positive selection on a pathogen evading its hosts' herd immunity (Fitch *et al.*, 1997), and we showed here the PPS model performed well in detecting this mode of adaptation. We believe the new PPS model, together with previous mutation-selection models that relaxed the assumption of constant fitnesses (Parto and Lartillot, 2017; Tamuri *et al.*, 2012), now encompass the major modes of selection in proteins.

We would like to note here two features of coding-sequence evolution that are ignored in our formulation of the PPS mutation selection model. First, the model assumes amino acid sites within the protein evolve independently. This is unrealistic because amino acids are linked and their substitution pattern is affected by interactions with other amino acids within the folded protein (Pollock *et al.*, 2012; Youssef *et al.*, 2020). In particular, substitutions toward suboptimal amino acids can be compensated by rapid substitution in another interacting amino acid, so as to reduce contact energies in the folded protein (Pollock *et al.*, 2012). How these rapid substitutions affect evolutionary dynamics within PPS and how they should be

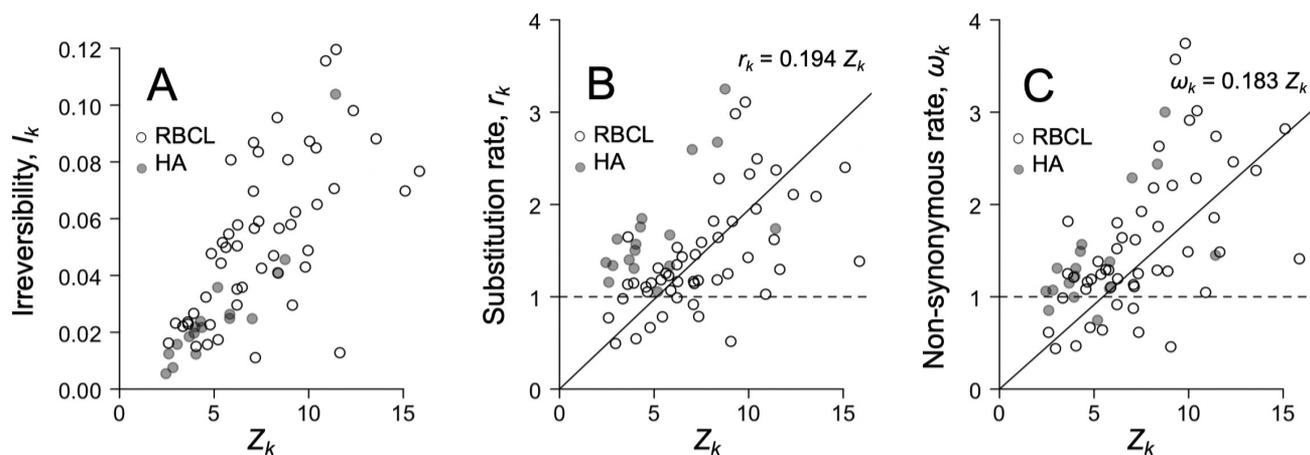


FIG. 2. Relationship between Z_k and evolutionary parameters for PPS sites in HA and rbcL. A: Irreversibility index, I_k , v Z_k . The index is normalised to give the expected excess number of substitutions from detailed balance. B: Site substitution rate, $r_k = -\sum \pi_k q_{ii,k}$, v Z_k . Note the $q_{ij,k}$ are scaled so that they give the relative rate with respect to a neutral sequence (Tamuri *et al.*, 2014). Thus, if $r_k = 1$, then the site evolves at the same rate as, say, a pseudogene. C: Non-synonymous rate, ω_k v Z_k . The penalty on Z_k is $\lambda = 0.001$ in all cases.

accommodated within the inference model will require further research (Youssef *et al.*, 2020). Second, the model assumes polymorphism is absent and new mutations either become fixed or lost instantaneously. This assumption appears reasonable for most populations of plants and animals because, in these, the scaled mutation rates, $N\mu$, are much less than one (Lynch and Conery, 2003). Even for influenza, a fast-evolving RNA virus, estimates of $N\mu$ are in the order of 10^{-3} (Zhao and Illingworth, 2019). However, levels of standing polymorphism can be substantial in many microorganisms (Lynch and Conery, 2003) or for some loci under certain forms of selection (e.g. selection in favour heterozygotes, Hughes and Nei 1988). Incorporating polymorphism within the mutation-selection inference machinery will be challenging, but recent polymorphism-aware phylogenetic approaches may provide a way forward (De Maio *et al.*, 2015).

Perhaps the most important insight from the application of the PPS model to real data is that the criteria $\omega > 1$ to detect positive selection in proteins is conservative. As we show here, sites under PPS are also under strong purifying constraints, and, at equilibrium, produce many deleterious mutations that are removed by selection. Because ω_k is the weighted average over the rate of all possible synonymous substitutions at a site, it follows that ω_k will be reduced if there are many deleterious mutations at the site even if the site is shifting rapidly among a few positively selection amino acids. We believe this insight should be incorporated into the much faster codon substitution models used in phylogenomic analyses, such as the branch-site model (Yang and Nielsen, 2002), to improve power in detecting adaptation in proteins.

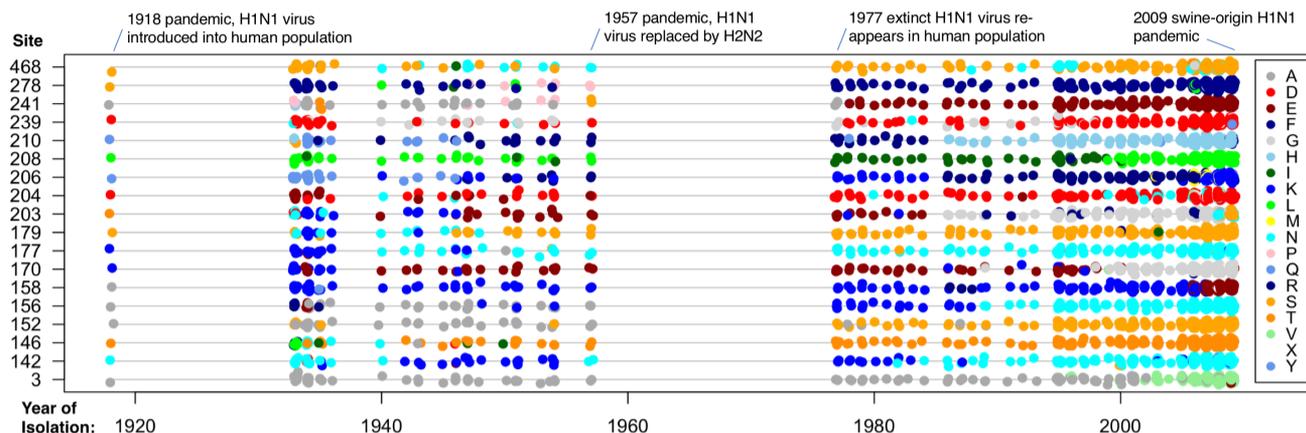


FIG. 3. Pattern of amino acid substitution in PPS sites of human influenza (H1N1) HA protein between 1918 and 2009. The penalty on Z_k is $\lambda=0.001$. Each colored dot represents a particular amino acid as indicated in the legend.

Methods

Maximum penalised likelihood estimation and likelihood ratio test of PPS

The swMutSel model (Tamuri *et al.*, 2012, 2014) is the special case of swMutSel-PPS when $Z_k=0$. We use swMutSel as a null model ($H_0:Z_k=0$) and swMutSel-PPS as the alternative model ($H_1:Z_k>0$) in a likelihood-ratio test. The vector of fitnesses at site k , $\mathbf{F}_k=(F_{i,k})$ and the PPS component, Z_k are estimated by maximising a penalised likelihood. The penalty on \mathbf{F}_k is the Dirichlet-based penalty of (Tamuri *et al.*, 2014), while for Z_k we use an exponential penalty $P(Z_k)=e^{-\lambda Z_k}$, where the regularization parameter, λ , controls the strength of the penalty. When $\lambda=0$ there is no penalty while $\lambda>0$ leads to increasingly stronger penalties on the estimation of Z_k . During inference, the $q_{ij,k}$ (Eq. 2) are scaled in terms of the expected number of neutral substitutions per site (see Tamuri *et al.* 2012). This guarantees all sites are normalised to the same timescale. To speed up computation, the mutational parameters, required to construct μ_{ij} , and the branch lengths on the phylogeny are estimated under the FMutSel0 model (Yang and Nielsen, 2008) as explained in Tamuri *et al.* (2014). We note only the differences of the $F_{i,k}$ enter Eq. 2, thus, the fitness for the most common amino acid at site k is set to zero. Large negative $F_{i,k}$ values are capped to -10 during numerical optimisation. We recommend the optimisation routine is repeated three times using different parameter start values to ensure convergence to the correct estimates.

Let the maximum penalised log-likelihood for site k be $\ell_{0,k}$ and $\ell_{1,k}$, under the H_0 and H_1 hypotheses respectively. The test statistic is the difference in log-likelihoods $\delta_k=\ell_{1,k}-\ell_{0,k}$. If the test statistic is significantly different from zero, this is evidence site k is evolving under PPS. The distribution of the $2\delta_k$ statistic, when the null hypothesis is true, does not follow a χ^2 distribution. There are two reasons for this. First, because $Z_k=0$ is at the boundary of parameter space, the test statistic would be, asymptotically, distributed as a 50:50 mixture of a χ^2 distribution and a 0.5 point probability mass at 0 (Goldman

and Whelan, 2000; Self and Liang, 1987). The second reason is that the penalty on Z_k affects the 50:50 proportion because the penalty forces the estimates of Z_k towards zero.

Because we do not know what the asymptotic distribution of δ_k should be, we use Cox's (1962) Monte Carlo simulation to obtain the null distribution of δ_k . For a given site k in the alignment, we simulate N replicate sites on the phylogeny using the maximum penalised likelihood estimates (MPLEs) of \mathbf{F}_k under H_0 . The distribution, Δ_k , is determined by the difference in log-likelihood between the two models for each simulated site: $\Delta_k = (\delta_k^{(1)}, \delta_k^{(2)}, \dots, \delta_k^{(N)})$ where $\delta_k^i = \ell_{1,k}^{(i)} - \ell_{0,k}^{(i)}$ is the log-likelihood difference for the i -th simulation. If the test statistic from the real data (δ_k) is larger than, say, 95% of Δ_k , we reject the null hypothesis H_0 (no PPS) and accept the alternative hypothesis H_1 (PPS) at the $\alpha=0.05$ significance level. Cox's approach has been shown to work well in phylogenetic datasets (Goldman, 1993). When analysing an ensemble of sites in a multiple sequence alignment, we correct for multiple testing using a false discovery rate procedure to select candidate PPS sites (Benjamini and Hochberg, 1995).

Padé approximation to calculate the matrix exponential

Calculation of the likelihood along a branch of length t in the phylogeny requires calculation of $\mathbf{P}(t) = \exp t\mathbf{Q}_k$, where $\mathbf{Q}_k = (q_{ij,k})$ is the substitution matrix (Eq. 2). However, because the PPS model is irreversible, the usual Eigen decomposition algorithm used to calculate $\mathbf{P}(t)$ is not stable (Yang, 2014). Here we use the Padé approximation (Moler and Van Loan, 2003)

$$\exp \mathbf{A} \approx R_{qq}(\mathbf{A}) = [D_{qq}(\mathbf{A})]^{-1} N_{qq}(\mathbf{A}),$$

where $N_{qq}(\mathbf{A}) = \sum_{i=0}^q c_q(i) \mathbf{A}^i$, $D_{qq}(\mathbf{A}) = \sum_{i=0}^q c_q(i) (-1)^i \mathbf{A}^i$, and $c_q(i) = (2q-i)!q!/(2q)!i!(q-i)!$. Note $\exp \mathbf{A} = (\exp \mathbf{A}/m)^m$, with $m=2^j$ for some integer j . Accuracy is improved considerably by choosing a suitable j such that the Padé approximation works well for $\exp \mathbf{A}/m$. Then

$$\exp \mathbf{A} \approx [R_{qq}(\mathbf{A}/2^j)]^{2^j}.$$

Appropriate values for q and j are chosen according to the size of \mathbf{A} and the desired accuracy in the calculation of $\exp \mathbf{A}$ (Moler and Van Loan 2003: Table 1).

In our model, the calculation of the likelihood at a site involves multiple computations of $\exp t\mathbf{Q}$ for every branch in the phylogeny. We choose q and j according to the largest branch length t . Because $(t\mathbf{Q}/m)^i = (t/m)^i \mathbf{Q}^i$, we calculate all necessary $c_q(i)$ and \mathbf{Q}^i once and cache these in memory throughout the likelihood calculation. Calculating \mathbf{Q}^i once is more efficient than setting $\mathbf{A} = \mathbf{Q}t$ and applying the Padé approximation directly. Instead, we compute $\mathbf{B}^i = (t/m)^i \mathbf{Q}^i$, $R_{qq}(\mathbf{B})$, and finally $[R_{qq}(\mathbf{B})]^{2^j}$ for each value of t . We found this matrix exponentiation algorithm is approximately 1.5 times faster than the

Taylor series approximation suggested in phylogenetics (Yang, 2014), albeit using more memory to store the precalculated matrix powers.

Simulated data

To test the specificity and sensitivity of the LRT for PPS, we simulated sites on a balanced 512-taxa tree with branch lengths equal to 0.0125 neutral substitutions per site (Tamuri *et al.*, 2014). We simulated sites under a null model with no PPS ($H_0: Z_k=0$), and under the alternative model with PPS ($H_1: Z_k>0$) with three strengths of selection $Z_k=\{2,5,10\}$, and with 1,000 sites simulated under each model setup. Following Tamuri *et al.* (2014), amino acid fitnesses for each site were sampled from a bimodal normal distribution with ten randomly-selected amino acids chosen to have $F\sim N(0,1)$ and the remaining amino acids to have $F\sim N(-10,1)$. This simulation setup was chosen because it leads to simulated data that captures two important features seen in real data: (1) A bimodal distribution of selection coefficients among mutations, and (2) a sharp distribution of amino acid preferences among sites (Tamuri *et al.*, 2012, 2014).

Simulated data were then analysed with the swMutSel software to estimate model parameters. The branch lengths and mutation parameters were fixed to their true values ($k=2$, $\pi^*=0.25$) throughout the analysis and only the sitewise fitnesses (\mathbf{F}_k) and diversifying selection (Z_k) parameters were estimated. For each simulation setup, we calculated the MPLE and the LRT as described above, using $N=100$ replicates in Cox's procedure. In all analyses the Dirichlet penalty on \mathbf{F}_k has $\alpha=0.01$, and three strengths of penalty on Z_k were tested, $\lambda=\{0.01, 0.5, 1.0\}$.

Using the LRT results, we determined the false positive and false negative rates. The false positive rate is calculated by determining the number of tests that incorrectly rejected the null hypothesis ($Z_k=0$). The true positive rate is calculated from the number of tests that correctly rejected the null hypothesis ($Z_k>0$).

Real sequence data

We downloaded 3,120 HA protein-coding sequences of human influenza H1 viruses (excluding 2009 pandemic-H1N1 and partial sequences) from the NIAID Influenza Research Database (Squires *et al.*, 2012); we downloaded 3,490 RuBisCO eudicotyledon sequences from a previous study (Stamatakis *et al.*, 2010); and we downloaded CYTB genes of placental mammals from NCBI RefSeq (O'Leary *et al.*, 2016) mitochondria genomes. We reduced the HA and RuBisCO datasets to 466 and 478 sequences respectively by using CD-HIT (Fu *et al.*, 2012) with clustering thresholds of 99.3% and 96% of amino acid sequence identity. The CYTB data was reduced to 418 sequences by keeping one sequence per mammal genus. Sequences were aligned using PRANK (Loytynoja and Goldman, 2005), and the alignments used to

estimate tree topologies with RAxML under the GTRCAT model (Stamatakis, 2014). Because the swMutSel-PPS model is irreversible, trees must be rooted. Thus, outgroups were used to root the trees: Avian influenza (HA), monocotyledons (RuBisCO) and monotremes (CYTB). Outgroups were removed and analyses carried out on the rooted ingroup tree (for the PPS model), and the unrooted ingroup tree (for the no PPS model). Sites with residues in fewer than 50 taxa were not analysed. This corresponds to 31, 23 and 27 sites in the rbcL, HA and CYTB alignments respectively. We note Z_k is not identifiable if a site is conserved for a single amino acid. Such conserved sites have the same likelihood under the H_0 and H_1 hypotheses. MPLE and LRT were carried out as described above using $\alpha=0.01$ in the Dirichlet penalty. We note estimates of $F_{i,k}$ are different between the two models ($Z_k=0$ v $Z_k \neq 0$, Fig. S1). Before carrying out the FDR correction to select candidate sites under PPS, we verify the distribution of p -values is uniform (Fig. S2).

Supplementary data and software

Alignments, phylogenies and analysis output are available in FigShare, doi: 10.6084/m9.figshare.14637765.v1. The mutation-selection PPS codon substitution model is implemented in the swMutSel computer program available at github.com/tamuri. Supplementary Information is available at Molecular Biology and Evolution online.

Acknowledgments

We thank Nick Goldman for valuable comments in the design of the LRT test. MdR is supported by Biotechnology and Biological Sciences Research Council (BBSRC, UK) award BB/T01282X/1.

References

- Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.
- Bulmer, M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129(3): 897–907.
- Cox, D. D. and O’Sullivan, F. 1990. Asymptotic analysis of penalized likelihood and related estimators. *The Annals of Statistics*, 18(4): 1676–1695.
- Cox, D. R. 1962. Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 406–424.
- De Maio, N., Schrepf, D., and Kosiol, C. 2015. Pomo: an allele frequency-based approach for species tree estimation. *Systematic biology*, 64(6): 1018–1031.
- dos Reis, M. 2013. Population genetics and substitution models of adaptive evolution. *arXiv preprint arXiv:1311.6682*.
- dos Reis, M. 2015. How to calculate the non-synonymous to synonymous rate ratio of protein-coding genes under the fisher-wright mutation-selection framework. *Biology Letters*, 11(4): 20141031.
- dos Reis, M., Hay, A. J., and Goldstein, R. A. 2009. Using non-homogeneous models of nucleotide substitution to identify host shift events: application to the origin of the 1918 ‘Spanish’ influenza pandemic virus. *J Mol Evol*, 69(4): 333–45.

- Fisher, R. 1930. The distribution of gene ratios for rare mutations. *Proc R Soc Edinb*, 50: 205–220.
- Fitch, W. M., Bush, R. M., Bender, C. A., and Cox, N. J. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc Natl Acad Sci U S A*, 94(15): 7712–8.
- Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)*, 28(23): 3150–3152.
- Goldman, N. 1993. Statistical tests of models of dna substitution. *J Mol Evol*, 36(2): 182–98.
- Goldman, N. and Whelan, S. 2000. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Molecular Biology and Evolution*, 17(6): 975–978.
- Grimmet, G. and Stirzaker, D. 2004. *Probability and random processes*. Oxford University Press, Oxford, third edition edition.
- Halpern, A. L. and Bruno, W. J. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol*, 15(7): 910–7.
- Huelsenbeck, J. P., Bollback, J. P., and Levine, A. M. 2002. Inferring the root of a phylogenetic tree. *Syst Biol*, 51(1): 32–43.
- Hughes, A. L. and Nei, M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, 335(6186): 167–70.
- Jones, C. T., Youssef, N., Susko, E., and Bielawski, J. P. 2016. Shifting balance on a static mutation–selection landscape: A novel scenario of positive selection. *Molecular biology and evolution*, 34(2): 391–407.
- Kapralov, M. V. and Filatov, D. A. 2007. Widespread positive selection in the photosynthetic rubisco enzyme. *BMC Evolutionary Biology*, 7(1): 73.
- Kimura, M. 1962. On the probability of fixation of mutant genes in a population. *Genetics*, 47: 713–9.
- Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4): 893–903.
- Kimura, M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press.
- Kimura, M. and Ohta, T. 1970. Probability of fixation of a mutant gene in a finite population when selective advantage decreases with time. *Genetics*, 65(3): 525–534.
- Loytynoja, A. and Goldman, N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A*, 102(30): 10557–62.
- Lynch, M. and Conery, J. 2003. The origins of genome complexity. *Science*, 302(5649): 1401–4.
- McCandlish, D. M., Epstein, C. L., and Plotkin, J. B. 2015. Formal properties of the probability of fixation: identities, inequalities and approximations. *Theoretical population biology*, 99: 98–113.
- Moler, C. and Van Loan, C. 2003. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM review*, 45(1): 3–49.
- Muse, S. V. and Gaut, B. S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*, 11(5): 715–24.
- Nielsen, R. and Yang, Z. 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol Biol Evol*, 20(8): 1231–9.
- O’Leary, N. A., Wright, M. W., Brister, J. R., Ciufu, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E.,

- Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. 2016. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, 44(D1): D733–45.
- Parto, S. and Lartillot, N. 2017. Detecting consistent patterns of directional adaptation using differential selection codon models. *BMC evolutionary biology*, 17(1): 147.
- Parto, S. and Lartillot, N. 2018. Molecular adaptation in rubisco: Discriminating between convergent evolution and positive selection using mechanistic and classical codon models. *PLoS one*, 13(2): e0192697.
- Pollock, D. D., Thiltgen, G., and Goldstein, R. A. 2012. Amino acid coevolution induces an evolutionary stokes shift. *Proceedings of the National Academy of Sciences*, 109(21): E1352–E1359.
- Rodrigue, N. and Lartillot, N. 2017. Detecting adaptation in protein-coding genes using a bayesian site-heterogeneous mutation-selection codon substitution model. *Molecular biology and evolution*, 34(1): 204–214.
- Rodrigue, N., Philippe, H., and Lartillot, N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc Natl Acad Sci U S A*, 107(10): 4629–34.
- Rodrigue, N., Latrille, T., and Lartillot, N. 2021. A bayesian mutation–selection framework for detecting site-specific adaptive evolution in protein-coding genes. *Molecular biology and evolution*, 38(3): 1199–1208.
- Sawyer, S. A. and Hartl, D. L. 1992. Population genetics of polymorphism and divergence. *Genetics*, 132(4): 1161–76.
- Self, S. G. and Liang, K.-Y. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398): 605–610.
- Smith, G. J., Vijaykrishna, D., Bahl, J., Lycett, S. J., Worobey, M., Pybus, O. G., Ma, S. K., Cheung, C. L., Raghvani, J., Bhatt, S., Peiris, J. S., Guan, Y., and Rambaut, A. 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, 459(7250): 1122–5.
- Spielman, S. J. and Wilke, C. O. 2015. The relationship between dn/ds and scaled selection coefficients. *Molecular biology and evolution*, page msv003.
- Squires, R. B., Noronha, J., Hunt, V., García-Sastre, A., Macken, C., Baumgarth, N., Suarez, D., Pickett, B. E., Zhang, Y., Larsen, C. N., Ramsey, A., Zhou, L., Zaremba, S., Kumar, S., Deitrich, J., Klem, E., and Scheuermann, R. H. 2012. Influenza Research Database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza and Other Respiratory Viruses*, 6(6): 404–416.
- Stamatakis, A. 2014. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics*, 30(9): 1312–1313.
- Stamatakis, A., Göker, M., and Grimm, G. W. 2010. Maximum Likelihood Analyses of 3,490 rbcL Sequences: Scalability of Comprehensive Inference versus Group-Specific Taxon Sampling. *Evolutionary bioinformatics online*, 6: 73–90.
- 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(3): 1101–15.
- Tamuri, A. U., Goldman, N., and dos Reis, M. 2014. A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. *Genetics*, 197(1): 257–71.

- Taubenberger, J. K., Reid, A. H., Lourens, R. M., Wang, R., Jin, G., and Fanning, T. G. 2005. Characterization of the 1918 influenza virus polymerase genes. *Nature*, 437(7060): 889–93.
- Wright, S. 1931. Evolution in mendelian populations. *Genetics*, 16: 97–159.
- Yang, Z. 2014. *Molecular evolution: a statistical approach*. Oxford University Press, Oxford.
- Yang, Z. and Nielsen, R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol*, 46(4): 409–18.
- Yang, Z. and Nielsen, R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*, 19(6): 908–17.
- Yang, Z. and Nielsen, R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol*, 25(3): 568–79.
- Youssef, N., Susko, E., and Bielawski, J. P. 2020. Consequences of stability-induced epistasis for substitution rates. *Molecular Biology and Evolution*, 37(11): 3131–3148.
- Zhao, L. and Illingworth, C. J. 2019. Measurements of intrahost viral diversity require an unbiased diversity metric. *Virus Evolution*, 5(1): vey041.