SI Theory of measurement for site-specific evolutionary rates in amino-acid sequences

Dariya K. Sydykova and Claus O. Wilke

SI Appendix

Maximum Likelihood approach. Our goal is to develop a framework that allows us to analytically determine the effects of different measurement matrices Q_{M} on inferred site-wise rates $\hat{r}^{(k)}$. We employ a maximum likelihood (ML) approach, and we use the simplest log-likelihood function to be able to derive an analytical solution. This log-likelihood function describes a pair of sequences that have diverged from a common ancestor. To infer divergence time $\hat{t}$ from these sequences, we can use the log-likelihood function

$$L(\hat{t}) = \sum_{i,j} n_{ij} \log[\hat{\pi}_i p_{\mathrm{M},ij}(\hat{t})]. \qquad [1]$$

Here, n_{ij} is the number of times amino acid i has been substituted by an amino acid j across sites in the two sequences, $\hat{\pi}_i$ is the estimated equilibrium frequency of amino acid i, and $p_{\mathrm{M},ij}(\hat{t})$ is the probability of an amino acid i replacing amino acid j after time t under the inference model. The probabilities are defined by the model $P_{\mathrm{M}}(\hat{t}) = e^{\hat{t}Q_{\mathrm{M}}}$, where $p_{\mathrm{M},ij}(\hat{t})$ are the elements of $P_{\mathrm{M}}(\hat{t})$, and Q_{M} is an amino acid substitution matrix under the inference model. The product $\hat{\pi}_i p_{\mathrm{M},ij}(\hat{t})$, therefore, is the probability that amino acid i is substituted by amino acid j, assuming the inference model is correct. To infer $\hat{t}$, we find the value of $\hat{t}$ for which the log-likelihood function is maximized.

Equation 1 is not site-specific, since the n_{ij} aggregate information over all sites. However, we seek to infer a site-specific rate $\hat{r}^{(k)}$. We cannot directly use equation 1 for this purpose, because the n_{ij} will either be 0 or 1. (Either a given substitution did happen at a site or it did not.) In a multiple sequence alignment, rates can be inferred at individual sites because we have more than two sequences and hence can observe more than one substitution at one site. However, working with more than two sequences adds unnecessary complications to the analytic calculations we seek to perform. Here, we instead employ a mathematical trick where we envision that each site is duplicated many times, and each duplicate of a site evolves independently according to the same true model that governs that site. This trick makes our analytical derivation feasible. We will show later that the number of assumed duplicates cancels from the calculation and can hence be set to 1. The site-specific log-likelihood equation for a pair of sequences with duplicates can be written as

$$L(\hat{r}^{(k)}) = \sum_{i,j} n_{ij}^{(k)} \log[\hat{\pi}_i^{(k)} p_{\mathrm{M},ij}^{(k)}(\hat{t}, \hat{r}^{(k)})]. \qquad [2]$$

Here, $n_{ij}^{(k)}$ is the number of times the amino acid i has been substituted by an amino acid j among all the duplicates of site k, and $\hat{\pi}_i^{(k)}$ and $p_{\mathrm{M},ij}^{(k)}(\hat{t}, \hat{r}^{(k)})$ are specific to site k. The inference model here has an additional parameter to the model used previously in the simplest ML. We define the inference model as $P_{\mathrm{M}}(\hat{t}, \hat{r}^{(k)}) = e^{\hat{t}\hat{r}^{(k)}Q_{\mathrm{M}^{(k)}}}$ with the parameter $\hat{r}^{(k)}$ for the rate of evolution at site k. Using this model, we seek to infer divergence time $\hat{t}$ from all sites and their duplicates and the rate of evolution $\hat{r}^{(k)}$ from individual sites' duplicates.

We can write $n_{ij}^{(k)}$ in terms of the true model as

$$n_{ij}^{(k)} = n\pi_i^{(k)} p_{\mathrm{T},ij}^{(k)}(t), \qquad [3]$$

where n is the number of duplicates of a site (we assume n is the same for all sites), and $\pi_i^{(k)}$ is the true equilibrium frequency of an amino acid i at site k. Strictly speaking, this equation is only correct in the limit of infinitely large n, because the left-hand side represents the observed numbers of specific substitutions and the right-hand side represents the expected numbers of specific substitutions under the true model. For smaller n, equation 3 is correct only on average.

To solve for the rate that maximizes the log-likelihood function, we insert the expression for $n_{ij}^{(k)}$ into equation 2, take the derivative with respect to $\hat{r}^{(k)}$, set it to zero, and then solve for $\hat{r}^{(k)}$:

$$\begin{aligned} 0 &= \frac{d}{d\hat{r}^{(k)}} L(\hat{r}^{(k)}) \\ &= \sum_{i,j} \frac{d}{d\hat{r}^{(k)}} \left(n\pi_i^{(k)} p_{\mathrm{T},ij}^{(k)}(t) \log[\hat{\pi}_i^{(k)} p_{\mathrm{M},ij}^{(k)}(\hat{t}, \hat{r}^{(k)})] \right) \\ &= n \sum_{i,j} \frac{\pi_i^{(k)} p_{\mathrm{T},ij}^{(k)}(t)}{p_{\mathrm{M},ij}^{(k)}(\hat{t}, \hat{r}^{(k)})} \frac{d}{d\hat{r}^{(k)}} p_{\mathrm{M},ij}^{(k)}(\hat{t}, \hat{r}^{(k)}). \end{aligned} \qquad [4]$$

Here, $\hat{\pi}_i^{(k)}$ cancels out when we take the derivative because $\frac{d}{dx}\log(af(x)) = \frac{1}{ax} a \frac{d}{dx} f(x) = \frac{1}{x}\frac{d}{dx} f(x)$. We see that n also cancels, since it is a positive constant. Thus, the ML solution is independent of the number of assumed site duplicates. Because $\frac{d}{d\hat{r}^{(k)}} p_{\mathrm{M},ij}^{(k)}(\hat{t}, \hat{r}^{(k)}) = \left(P_{\mathrm{M}}^{(k)}(\hat{t}, \hat{r}^{(k)}) \hat{t} Q_{\mathrm{M}}^{(k)} \right)_{ij}$, equation 4 becomes

$$0 = \sum_{i,j} \frac{\pi_i^{(k)} p_{\mathrm{T},ij}^{(k)}(t)}{p_{\mathrm{M},ij}^{(k)}(\hat{t}, \hat{r}^{(k)})} \left(P_{\mathrm{M}}^{(k)}(\hat{t}, \hat{r}^{(k)}) \hat{t} Q_{\mathrm{M}}^{(k)} \right)_{ij}. \qquad [5]$$

At this point in our calculations, we can start assessing the effects of different $Q_{\mathrm{M}}^{(k)}$ on the inference of $\hat{r}^{(k)}$. In the following subsections, we will solve the equation for $\hat{r}^{(k)}$ using a variety of different choices for $Q_{\mathrm{M}}^{(k)}$, including $Q_{\mathrm{M}}^{(k)} = Q_{\mathrm{T}}^{(k)}$ (the measurement matrix equals the true substitution matrix at site k) and $Q_{\mathrm{M}}^{(k)} = Q_{\mathrm{JC}}$ (the measurement matrix equals the Jukes-Cantor-like matrix at all sites). Note that we also need to infer $\hat{t}$, so the two parameters will be derived as a product $\hat{t}\hat{r}^{(k)}$ in the following sections.

Site-wise rate for an arbitrary measurement matrix and small divergence. If divergence is limited, we can solve equation 5 for an arbitrary measurement matrix. We assume that $t \to 0$, so that we can expand the ML equation to first order. In this limit, without loss of generality we can assume that $\hat{t}$ is proportional to t, and for simplicity we write $\hat{t} = t$. Under these assumptions, the true model becomes $P_{\mathrm{T}}^{(k)}(t) = I + tQ_{\mathrm{T}}^{(k)}$ and the inference model becomes $P_{\mathrm{M}}^{(k)}(\hat{t}, \hat{r}^{(k)}) = I + t\hat{r}^{(k)} Q_{\mathrm{M}}^{(k)}$, where I is the identity matrix. We insert these expressions into equation 5 and obtain

$$0 = \sum_{i,j} \frac{\pi_i^{(k)} \left(I + tQ_{\mathrm{T}}^{(k)}\right)_{ij}}{\left(I + t\hat{r}^{(k)} Q_{\mathrm{M}}^{(k)}\right)_{ij}} \left(tQ_{\mathrm{M}}^{(k)}\right)_{ij}. \qquad [6]$$

We separate the diagonal and off-diagonal elements,

$$0 = \sum_{i} \frac{\pi_i^{(k)} t q_{\mathrm{M},ii}^{(k)}}{1 + \hat{r}^{(k)} t q_{\mathrm{M},ii}^{(k)}} + \sum_{i,j\neq i} \frac{\pi_i^{(k)} t q_{\mathrm{T},ii}^{(k)}}{\hat{r}^{(k)}}, \qquad [7]$$

expand each sum to first order, and solve for $\hat{r}^{(k)}$. We find

$$\hat{r}^{(k)} = \frac{\sum_{i,j\neq i} \pi_i^{(k)} q_{\mathrm{T},ij}^{(k)}}{\sum_{i,j\neq i} \pi_i^{(k)} q_{\mathrm{M},ij}^{(k)}}. \qquad [8]$$

True matrix as measurement matrix. When the measurement matrix equals the true matrix, $Q_{\mathrm{M}}^{(k)} = Q_{\mathrm{T}}^{(k)}$, then $\hat{t} = t$ and $\hat{r}^{(k)} = 1$. We can show this by noting that equation 5 becomes

$$0 = \sum_{i,j} \frac{\pi_i^{(k)} \left(e^{tQ_{\mathrm{T}}^{(k)}}\right)_{ij}}{\left(e^{\hat{t}\hat{r}^{(k)} Q_{\mathrm{T}}^{(k)}}\right)_{ij}} \left(e^{\hat{t}\hat{r}^{(k)} Q_{\mathrm{T}}^{(k)}} \hat{t} Q_{\mathrm{T}}^{(k)}\right)_{ij}, \qquad [9]$$

and the right-hand side of this equation simplifies to 0 under these assumptions. First, for $\hat{t} = t$ and $\hat{r}^{(k)} = 1$, equation 9 simplifies to

$$0 = \sum_{i,j} \pi_i^{(k)} \left(e^{tQ_{\mathrm{T}}^{(k)}} tQ_{\mathrm{T}}^{(k)}\right)_{ij}. \qquad [10]$$

We note that if a is a vector and B is a matrix, then $\sum_{i,j} a_i B_{ij} = \sum_j (aB)_j$. Using this fact and the identity $e^{tQ} tQ = tQe^{tQ}$, we can rewrite equation 10 as

$$0 = \sum_{j} \left(\pi^{(k)} tQ_{\mathrm{T}}^{(k)} e^{tQ_{\mathrm{T}}^{(k)}}\right)_{j}, \qquad [11]$$

where $\pi^{(k)}$ is the vector of equilibrium frequencies at site k, $\pi^{(k)} = (\pi_1^{(k)}, \pi_2^{(k)}, \pi_3^{(k)}, \dots)$. Because $Q_{\mathrm{T}}^{(k)}$ is an infinitesimal generator of a continuous-time Markov process, $\pi^{(k)}$ is a left-eigenvector to $Q_{\mathrm{T}}^{(k)}$ with eigenvalue 0, $\pi^{(k)} Q_{\mathrm{T}}^{(k)} = 0$. Thus, the right-hand-side of equation 11 vanishes and $\hat{t} = t$ and $\hat{r}^{(k)} = 1$ are the solution to equation 5.

True matrix as a multiple of the measurement matrix. We now consider the case where $Q_{\mathrm{T}}^{(k)} = r^{(k)} Q$ for an arbitrary transition matrix Q. Here, $r^{(k)}$ is the true rate parameter at site k. In this case, if we use Q as measurement matrix, $Q_{\mathrm{M}} = Q$, then $\hat{t} = t$ and $\hat{r}^{(k)} = r^{(k)}$. The argument is similar to the preceding subsection. We insert the expressions for $Q_{\mathrm{T}}^{(k)}$ and Q_{M} into equation 5 and then show that the right-hand side vanishes if $\hat{t} = t$ and $\hat{r}^{(k)} = r^{(k)}$. Following the same steps as before, we find

$$0 = \sum_{j} \left(\pi tQe^{tr^{(k)} Q}\right)_{j}, \qquad [12]$$

where π is the site-independent vector of equilibrium frequencies for Q. Again, because π is a left-eigenvector to Q with eigenvalue 0, so that $\pi Q = 0$, the right-hand side of equation 12 vanishes.

Naïve substitution matrix as measurement matrix. We now assume that $Q_{\mathrm{M}}^{(k)} = Q_{\mathrm{JC}}$. The Q_{JC} matrix is not site-specific, and it assumes that each amino acid is equally likely to be replaced by any other amino acid. For this reason, we also refer to this matrix as the naïve substitution matrix. The matrix elements of Q_{JC} are given by

$$q_{\mathrm{JC},ij} = \begin{cases} \dfrac{1}{19} & \text{if } i \neq j, \\ -1 & \text{if } i = j. \end{cases} \qquad [13]$$

If $Q_{\mathrm{M}}^{(k)} = Q_{\mathrm{JC}}$, equation 5 can be written as

$$0 = \sum_{i,j} \pi_i^{(k)} p_{\mathrm{T,ij}}^{(k)}(t) A_{ij}(\hat{r}^{(k)}), \qquad [14]$$

where

$$A_{ij}(\hat{t}, \hat{r}^{(k)}) = \begin{cases} \dfrac{20\hat{t}}{19(-1 + \exp(20\hat{t}\hat{r}^{(k)}/19))} & \text{if } i \neq j, \\ \dfrac{-20\hat{t}}{19 + \exp(20\hat{t}\hat{r}^{(k)}/19)} & \text{if } i = j. \end{cases} \qquad [15]$$

This expression is the result of a computer-algebra calculation of the matrix exponential of Q_{JC}, performed in Mathematica and validated numerically for correctness. We solve equation 14 for $\hat{t}\hat{r}^{(k)}$ and find

$$\hat{t}\hat{r}^{(k)} = \frac{19}{20} \log \frac{19}{19 - 20\sum_{i,j\neq i} \pi_i^{(k)} p_{\mathrm{T,ij}}^{(k)}(t)}. \qquad [16]$$

Since we normalize site-wise rates by their mean, we divide this equation by $\hat{t}\langle\hat{r}(t)\rangle = \frac{\hat{t}}{m}\sum_l \hat{r}^{(l)}$, where m is the total number of sites in the protein sequence, and find

$$\hat{r}^{(k)}(t) = \frac{\log\left[1 - \frac{20}{19}\sum_{i,j\neq i} \pi_i^{(k)} p_{\mathrm{T,ij}}^{(k)}(t)\right]}{\frac{1}{m}\sum_l \log\left[1 - \frac{20}{19}\sum_{i,j\neq i} \pi_i^{(l)} p_{\mathrm{T,ij}}^{(l)}(t)\right]}. \qquad [17]$$

Here, for simplicity, we have assumed without loss of generality that $\hat{r}^{(k)}(t) = \hat{r}^{(k)}(t)/\langle\hat{r}(t)\rangle$. We can further simplify equation 17 by making assumptions about the true time t.

Limiting cases for small and large t. To find $\hat{r}^{(k)}(t)$ when t is 0, we take the limit of equation 17 as $t \to 0$. By applying l'Hospital's rule, we find

$$\hat{r}^{(k)}(0) = \lim_{t\to 0} \hat{r}^{(k)}(t)$$
$$= \frac{\lim\limits_{t\to 0} \dfrac{\sum_{i,j\neq i} \pi_i^{(k)} \frac{d}{dt} p_{\mathrm{T,ij}}^{(k)}(t)}{1 - \frac{20}{19}\sum_{i,j\neq i} \pi_i^{(k)} p_{\mathrm{T,ij}}^{(k)}(t)}}{\lim\limits_{t\to 0} \frac{1}{m}\sum_l \dfrac{\sum_{i,j\neq i} \pi_i^{(l)} \frac{d}{dt} p_{\mathrm{T,ij}}^{(l)}(t)}{1 - \frac{20}{19}\sum_{i,j\neq i} \pi_i^{(l)} p_{\mathrm{T,ij}}^{(l)}(t)}}. \qquad [18]$$

Because $\frac{d}{dt} p_{\mathrm{T,ij}}^{(k)}(0) = q_{\mathrm{T,ij}}^{(k)}$ and $p_{\mathrm{T,ij}}(0) = 0$ when $i \neq j$, the equation simplifies to

$$\hat{r}^{(k)}(0) = \frac{\sum_{i,j\neq i} \pi_i^{(k)} q_{\mathrm{T,ij}}^{(k)}}{\frac{1}{m}\sum_l \sum_{i,j\neq i} \pi_i^{(l)} q_{\mathrm{T,ij}}^{(l)}}. \qquad [19]$$

When $t \to \infty$, off-diagonal elements of $P^{(k)}(t)$ become $p_{\mathrm{T,ij}}^{(k)}(\infty) = \pi_j^{(k)}$. Therefore, equation 17 becomes

$$\hat{r}^{(k)}(\infty) = \frac{\log\left[1 - \frac{20}{19}\sum_{i,j\neq i} \pi_i^{(k)} \pi_j^{(k)}\right]}{\frac{1}{m}\sum_l \log\left[1 - \frac{20}{19}\sum_{i,j\neq i} \pi_i^{(l)} \pi_j^{(l)}\right]}. \qquad [20]$$

True model is codon. The derivations above assumed that the true model operates in an amino-acid space. Here, we derive the inferred rate assuming the true model is a codon model. We define a true site-wise codon model as $P_{\mathrm{T}}^{(k)}(t) = e^{tQ_{\mathrm{T}}^{(k)}}$, where $Q_{\mathrm{T}}^{(k)}$ is a site-wise matrix that captures the true rates of substitution between all codons except the stop codons. Here, t is the true time measured relative to codon substitutions (as opposed to amino-acid substitutions as was done above).

Similarly to the ML derivations for amino-acid models, we infer site-wise rate from two sequences with m sites that have evolved from a common ancestor. We also assume that the inference model is in an amino acid model. The ML function, therefore, is

$$L(\hat{r}^{(k)}) = \sum_{i,j\neq i}\sum_{a\in\mathcal{C}_i}\sum_{b\in\mathcal{C}_j} n\pi_a^{(k)} p_{\mathrm{T},ab}^{(k)}(t)\log[\hat{\pi}_i^{(k)} p_{\mathrm{M},ij}^{(k)}(\hat{t},\hat{r}^{(k)})). \qquad [21]$$

Here, $\hat{\pi}_i^{(k)}$ is the estimated equilibrium frequency of an amino acid i at site k, and $p_{\mathrm{M},ij}^{(k)}(\hat{t},\hat{r}^{(k)})$ is the probability that at site k an amino acid i changes into an amino acid j under the inference model. The subset of codons that translate to amino acid i is indicated as $\mathcal{C}_i$. The true equilibrium frequency of codon a at site k is $\pi_a^{(k)}$, and the true probability of a codon a changing into codon b after time t is $p_{\mathrm{T},ab}^{(k)}(t)$.

We take the derivative of equation 21 with respect to $\hat{r}^{(k)}$ and set it equal to zero. We assume that the amino acid inference model uses the Jukes-Cantor like matrix, so that $Q_{\mathrm{M}}^{(k)} = Q_{\mathrm{JC}}$, and then solve for $\hat{r}^{(k)}$. The derivations follow the same procedure as equations 13–17.

The final result for the inferred site-wise rate at arbitrary t is

$$\hat{r}^{(k)}(t) = \frac{\log\left[1-\frac{20}{19}\sum_{i,j\neq i}\sum_{a\in\mathcal{C}_i}\sum_{b\in\mathcal{C}_j}\pi_a^{(k)}p_{\mathrm{T},ab}^{(k)}(t)\right]}{\frac{1}{m}\sum_l\log\left[1-\frac{20}{19}\sum_{i,j\neq i}\sum_{a\in\mathcal{C}_i}\sum_{b\in\mathcal{C}_j}\pi_a^{(l)}p_{\mathrm{T},ab}^{(l)}(t)\right]}. \qquad [22]$$

In the limiting case of $t \to 0$, this expression can be simplified to

$$\hat{r}^{(k)}(0) = \frac{\sum_{i,j\neq i}\sum_{a\in\mathcal{C}_i}\sum_{b\in\mathcal{C}_j}\pi_a^{(k)}q_{\mathrm{T},ab}^{(k)}}{\frac{1}{m}\sum_l\sum_{i,j\neq i}\sum_{a\in\mathcal{C}_i}\sum_{b\in\mathcal{C}_j}\pi_a^{(l)}q_{\mathrm{T},ab}^{(l)}}. \qquad [23]$$

Similarly, in the limiting case of $t \to \infty$, the expression can be simplified to

$$\hat{r}^{(k)} = \frac{\log\left[1-\frac{20}{19}\sum_{i,j\neq i}\sum_{a\in\mathcal{C}_i}\sum_{b\in\mathcal{C}_j}\pi_a^{(k)}\pi_b^{(k)}\right]}{\frac{1}{m}\sum_l\log\left[1-\frac{20}{19}\sum_{i,j\neq i}\sum_{a\in\mathcal{C}_i}\sum_{b\in\mathcal{C}_j}\pi_a^{(l)}\pi_b^{(l)}\right]}. \qquad [24]$$

The Halpern–Bruno mutation–selection model. We used the Halpern–Bruno mutation–selection model (1) to parameterize our simulations of sequence evolution. For reference, we here recapitulate the model's derivation.

The model assumes that sites evolve independently of each other and experience the same selection pressure across all branches in the phylogenetic tree. The model also assumes that the mutation process is the same at all sites and that time is reversible. For simplicity, we describe the model at a single site, and refer to the site-specific rate of substitution from amino acid i to amino acid j at a site as q_{ij}. The rate of substitution is the product of the probability of a mutation and the probability that the mutation will go to fixation:

$$q_{ij} = N_e m_{ij} u_{ij}, \qquad [25]$$

where m_{ij} is the probability of an amino acid i mutating into an amino acid j, u_{ij} is the probability of fixation, and N_e is the effective population size. The probability of fixation is given by (2)

$$u_{ij} = \frac{2s_{ij}}{1-e^{-2N_e s_{ij}}} = \frac{1}{N_e}\frac{2N_e s_{ij}}{1-e^{-2N_e s_{ij}}}, \qquad [26]$$

where s_{ij} is the difference in fitness between amino acids i and j. We introduce scaled selection coefficients $S_{ij} = 2N_e s_{ij}$ and rewrite the fixation probability as

$$u_{ij} = \frac{1}{N_e}\frac{S_{ij}}{1-e^{-S_{ij}}}. \qquad [27]$$

After inserting equation 27 into equation 25, we arrive at

$$q_{ij} = m_{ij}\frac{S_{ij}}{1-e^{-S_{ij}}}. \qquad [28]$$

Because we are using a time-reversible model, we can estimate S_{ij} using the detailed-balance condition:

$$q_{ij}\pi_i = q_{ji}\pi_j \quad \text{for all } i,\ j, \qquad [29]$$

where π_i is the equilibrium frequency of amino acid i. We substitute equations 25 and 27 into equation 29 and find

$$\frac{S_{ji}}{S_{ij}}\frac{1-e^{-S_{ij}}}{1-e^{-S_{ji}}} = \frac{m_{ji}\pi_j}{m_{ij}\pi_i} \qquad [30]$$

Because $S_{ij} = -S_{ji}$ by definition, we can simplify this equation to

$$S_{ij} = \ln\left(\frac{m_{ji}\pi_j}{m_{ij}\pi_i}\right). \qquad [31]$$

After inserting equation 31 into equation 25, we arrive at a site-specific substitution matrix q_{ij} defined entirely in terms of the mutation process and the equilibrium amino-acid frequencies:

$$q_{ij} = m_{ij}\frac{\ln\left(\frac{m_{ji}\pi_j}{m_{ij}\pi_i}\right)}{1-\frac{m_{ij}\pi_i}{m_{ji}\pi_j}}. \qquad [32]$$

Using the mutation–selection model to populate a protein substitution matrix. For our simulations of sequence evolution, we needed an appropriate model of site-wise substitution. We chose the Halpern–Bruno mutation–selection model (1) parameterized by amino-acid equilibrium frequencies predicted from a structural model of protein evolution. As shown in equation 32, the site-specific substitution matrix $q_{ij}^{(k)}$ can be written in terms of the mutation matrix m_{ij} and the amino-acid equilibrium frequencies $\pi_i^{(k)}$. We calculate the equilibrium frequencies $\pi_i^{(k)}$ using the theory and data from Echave et al. (3), who provide an expression for $\pi_i^{(k)}$ in terms of the stability effects of mutations,

$$\pi_i^{(k)} = \frac{e^{-\alpha\Delta\Delta G_{oi}^{(k)}}}{\sum_j e^{-\alpha\Delta\Delta G_{oj}^{(k)}}}, \qquad [33]$$

where α is a positive free parameter and $\Delta\Delta G_{oi}^{(k)}$ is the change in stability of the protein structure when an arbitrarily chosen reference amino acid at site k is substituted with amino acid i.

Here, we assume all mutations are equally likely, so that $m_{ij} = 1$ for $i \neq j$. Further, following Ref. (3), we set $\alpha = 1$, so that the scaled selection coefficients directly correspond to the difference in $\Delta\Delta G$ reference values,

$$S_{ij} = \Delta\Delta G_{oi} - \Delta\Delta G_{oj}. \qquad [34]$$

We calculated corresponding substitution matrices q_{ij} for all 124 sites for which Ref. (3) provides $\Delta\Delta G$ values for the protein egg white lysozyme (PDB ID: 132L) .

1. Halpern AL, Bruno WJ (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.* 15:910–917.
2. Kimura M (1962) On the probability of fixation of mutant genes in a population. *Genetics* 4:713–719.
3. Echave J, Jackson EL, Wilke CO (2015) Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. *Phys. Biol.* 12:025002.

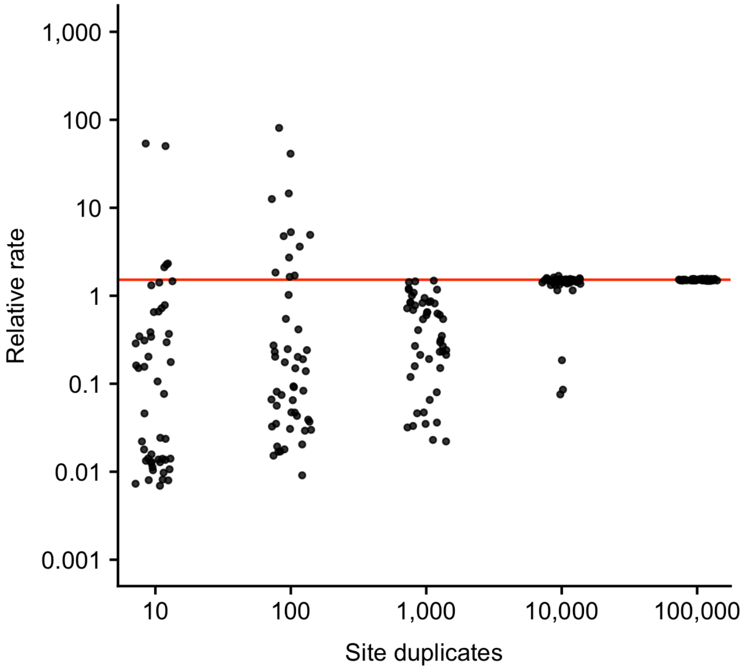

Fig. S1. Inferred rates converge to the analytically predicted value as the number of site duplicates is increased. Each dot represents one rate inference for the same site in a simulated alignment. A moderate amount of random jitter has been applied to the x position of each point to visualize multiple points with similar relative rates. Rate was inferred for the same site in 250 replicate simulations (50 simulations per number of site duplicates). The red line represents the site's analytically derived rate.

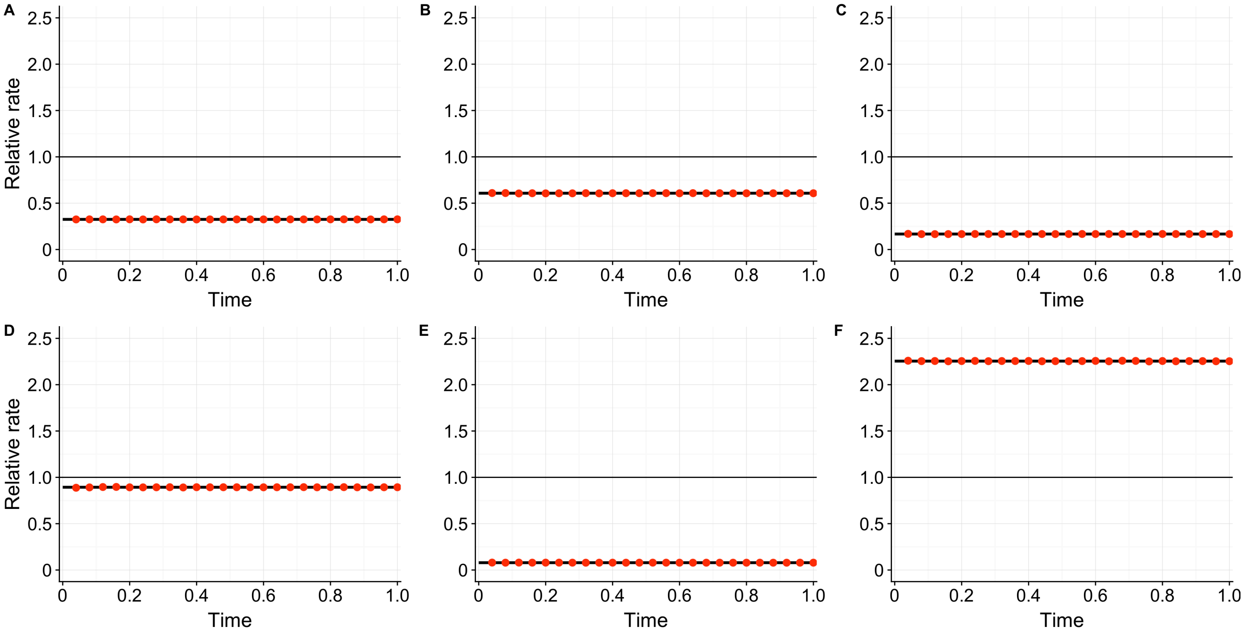

Fig. S2. Comparison of the true and inferred rate when the true model matrix is a scalar multiple of the inference model matrix. In the shown example, the true model matrix is $r^{(k)}Q_{\text{JC}}$ and the inference matrix is Q_{JC}. The black line represents the true rate, and the red dots represent the mean inferred rate over 30 simulations. The error bars represent the standard error. For all points, the bars are smaller than the symbol size. The thin horizontal line at 1 represents the average rate in the sequence. (A-F) Rate over time for sites 1-6, respectively.

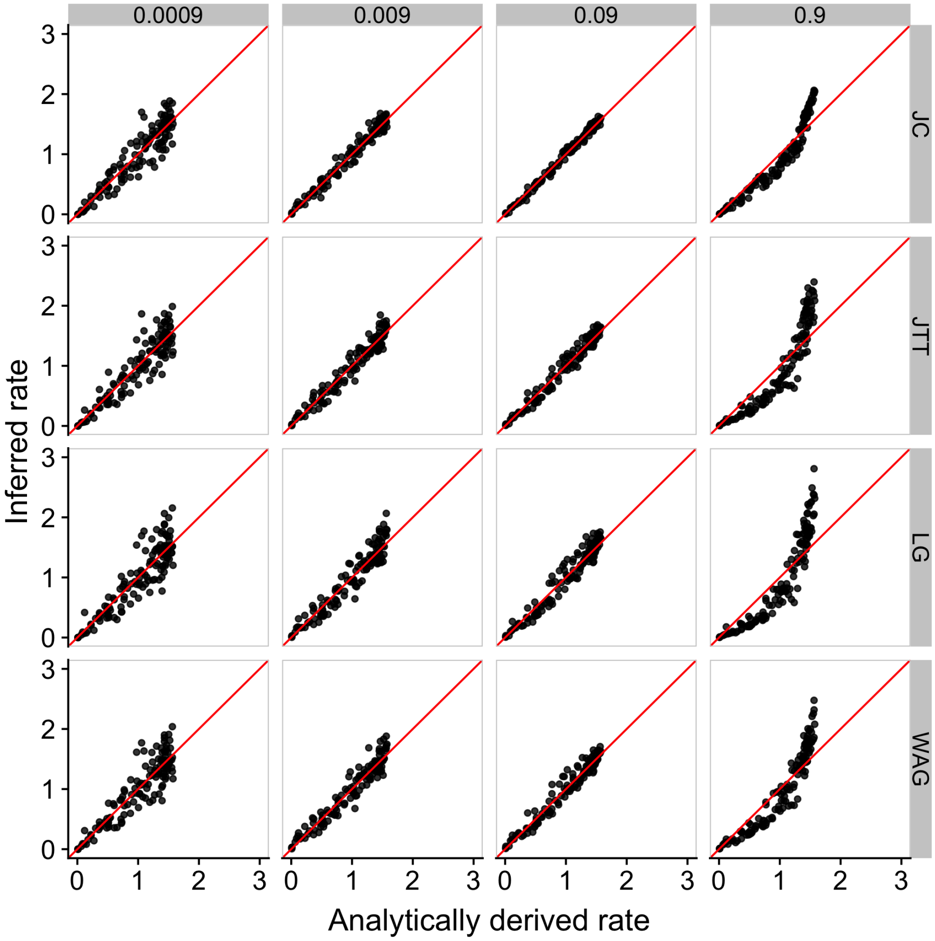

Fig. S3. Relationship between analytically derived rates and rates inferred with Jukes-Cantor-like (JC), JTT, LG, and WAG matrices. Sequence alignments were simulated for binary trees with 512 taxa and 124 sites, parameterized using data from egg white lysozyme (see Methods). No site duplicates were used in these simulations. The inferences with the JTT, LG, and WAG matrices assumed that each amino acid's equilibrium frequency is equal to the frequency of that amino acid in the entire alignment. The inference with JC matrix assumed that each amino acid's equilibrium frequency is $1/20$. The inferred rates plotted above are mean inferred rates over 50 simulations for each time point and site. The analytically derived rate was calculated with equation 19. The numbers on top of the plot panels indicate the time t used for each simulation. The labels on the right indicate the substitution matrix used for inference. Each point represents one site, and the diagonal line represents $x = y$.

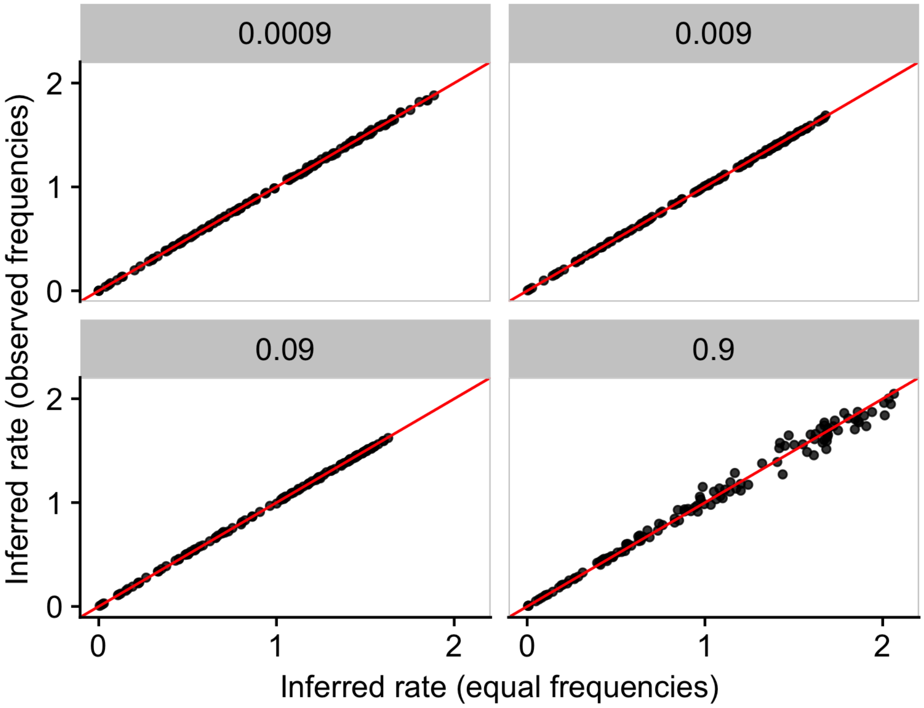

Fig. S4. Relationship between rates inferred with the Jukes-Cantor-like matrix with different assumptions about amino-acid equilibrium frequencies. One inference approach assumes that all equilibrium frequencies are $1/20$ for each amino acid (denoted as "equal frequencies"). The other approach assumes that each amino acid's equilibrium frequency is equal to the frequency of that amino acid in the alignment (denoted as "observed frequencies"). The numbers at the top of the plot indicate the time t used for each simulation. The alignments were simulated over trees with 512 number of taxa. Each point represents the mean inferred rate per site over 50 simulations, and the diagonal line represents $x = y$.

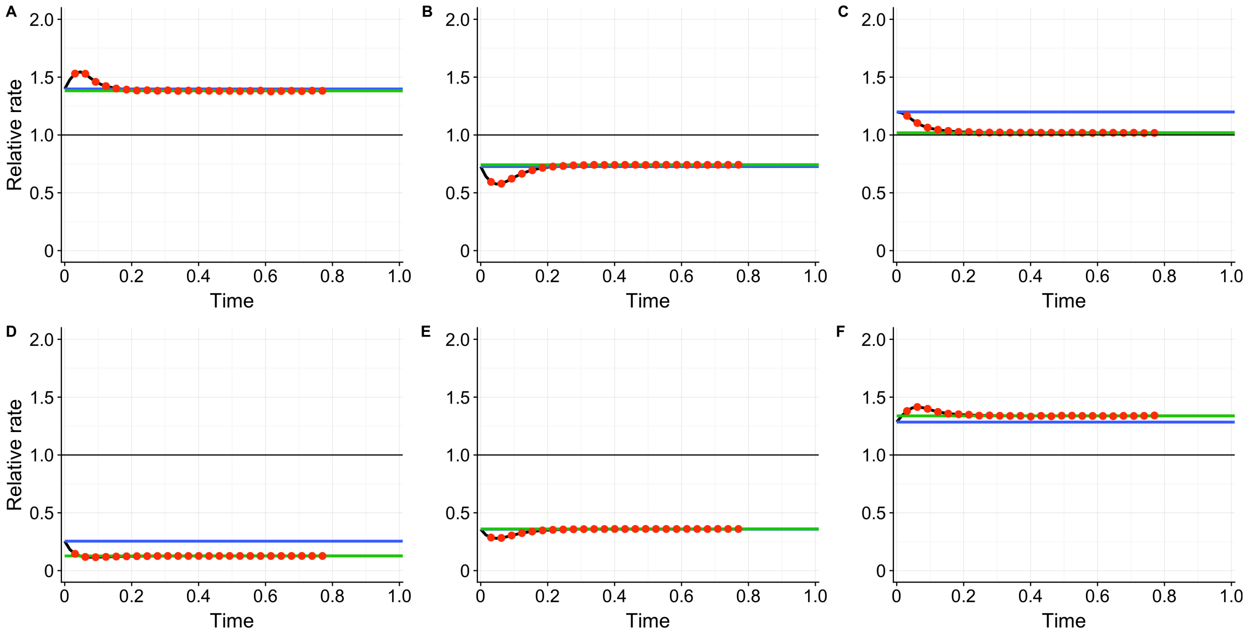

Fig. S5. Comparison of the inferred rates and the analytically derived rates when the true model is a codon model. The inference model used an amino acid Jukes-Cantor-like matrix. The black line represents the analytically derived rate (equation 22). The blue line represents analytically derived rate when time t is small (equation 23). The green line represent the analytically derived rate when time t is large (equation 24). The red dots represent the mean inferred rate over 30 simulations. The error bars represent the standard error. For all points, the bars are smaller than the symbol size. The horizontal line at 1 represents the average rate in the sequence. The time plotted is measured relative to amino acid substitutions. (A-F) Rate over time for sites 1, 2, 4, 5, 7, and 9, respectively.