SUPPLEMENTARY MATERIAL

Estimation of C/D ratio for Figure 1
To better demonstrate the downward curve of the C/D ratio versus distance graph in Figure 1 and our confidence in the curve, we determined the credible intervals for the C/D ratios for the Figure 1 data split into 20 windows using Markov Chain Monte Carlo (MCMC) simulations (one for each window). The 99% credible intervals for the C/D ratio for each window are presented in Table S1 below and are graphed in Figure S2. The high effective sample sizes and tight intervals show that we are quite confident in our estimations. There is a clear downward trend in the first five windows, as their intervals did not intersect at all. The C/D ratios continue to decrease steadily after the first five windows. The last few windows contained fewer data points and so the C/D ratio estimates were less constrained.

The common ancestor convergence ratio decreases
We also wanted to test our confidence that the C/D ratios in Figure 2A were decreasing. We used a linear model and estimated the slope using an MCMC. After burnin, the slope never came close to zero as the 99% credible region was between -0.109 and -0.094. We sampled the posterior distribution well with a high effective sample size and so we are confident that the slope is indeed negative. We reject the null hypothesis that the slope is zero in favor of the alternative that the slope is negative with a p-value less than 0.01. The 99% credible linear fits are graphed in Figure S3.

Estimate of effective number of accessible amino acids
We can calculate an effective number of accessible residues by considering the size of the alphabet of states $m$ that would result in a particular value of C/D if all substitutions were equally likely (i.e., a Jukes Cantor model (Jukes and Cantor 1969)). If there is only a short evolutionary distance between the branches, and the current amino acids in the two branches are likely to be the same, each amino acid can change to $m-1$ other amino acids. The probability that a substitution occurring at each branch would result in the same amino acid would be $\frac{1}{m-1}$, resulting in C/D

$$\frac{1}{1-\frac{1}{m-1}} = \frac{1}{m-2}.$$ When the amino acids are different, the probability that two substitutions would result in the same amino acid is slightly reduced, as neither substitution can result in either of the original amino acids. The probability that a substitution of one amino acid does not result in the other amino acid is $\frac{m-2}{m-1}$. The probability that the second amino acid substitution results in the same new amino
acid is \( \frac{1}{m-1} \), resulting in the probability of two substitutions yielding a convergent change of \( \frac{m-2}{(m-1)^2} \), giving an expected C/D of \( \frac{m-2}{m^2-3m+3} \). For large divergence times, the two amino acids have probability \( \frac{1}{m} \) of being the same, and probability \( \frac{m-1}{m} \) of being different, so the probability of a double substitution resulting in a convergent change is 
\[
\left( \frac{1}{m} \right)^2 \left( \frac{m-1}{m} \right) = \frac{m^2-3m+3}{m(m-1)(m-2)}.
\]
This corresponds to \( C/D = \frac{m^2-3m+3}{m^3-3m^2+5m-3} \).

**Exact averages: Phenomenological substitution models**

The various phenomenological substitution models (JC, Jukes-Cantor; general time reversible (GTR); Z, Ziheng Yang’s codon model, and CAT, the Lartillot group’s CAT_{60} model) specify a substitution matrix \( \mathbf{Q} \) (where \( Q_{ij} \) is the substitution rate from state \( i \) and state \( j \)) as well as \( \pi_i \), the equilibrium frequencies of state \( i \) (Goldman and Yang 1994; Quang et al. 2008). These states are amino acids for the JC, GTR, and CAT models, and codons for the Z model. Consider two points on a phylogenetic tree separated by an evolutionary distance \( \tau \), as shown in Figure 1, where the two states of the sequence at these points are \( i \) and \( j \) respectively. The probability that these two states would be found at these sites, given a simple reversible Markov process described by \( \mathbf{Q} \) and \( \pi \), is given by \( \pi_i \left[ e^{\mathbf{Q} \tau} \right]_{ij} \) (Yang 2014). Given these two states, the probability that substitutions occur from \( i \) to \( m \) and \( j \) to \( n \) \( (m \neq i, n \neq j) \) during short intervals of length \( \delta \) is \( Q_{im} Q_{jn} \delta^2 \). This would be convergent if \( m = n \) and divergent if \( m \neq n \). The probability of a set of convergent substitutions is therefore
\[
C(\tau) = \delta^2 \sum_{i,j} \pi_i \left[ e^{\mathbf{Q} \tau} \right]_{ij} \sum_{m \neq i,j} Q_{im} Q_{jm}
\]
(S.1)

while the corresponding probability of a set of divergent substitutions is

![Figure 1: Two points on a phylogenetic tree separated by evolutionary distance \( \tau \), where the sequence is in state \( i \) and \( j \) respectively.](image)
\[ D(\tau) = \delta^2 \sum_{i,j} \tau_{ij} \left[ e^{Q_\tau} \right]_{ij} \sum_{m \neq i} Q_{im} \sum_{n \neq j} Q_{jn} \]

\[ = \delta^2 \sum_{i,j} \tau_{ij} \left( \sum_{m \neq i} Q_{im} \left( \sum_{n \neq j} Q_{jn} \right) - C(\tau) \right) \]

or by substitution process \( k \), then we can write

\[ \frac{C}{D}(\tau) = \frac{\sum_{i,j} \tau_{ij} \left[ e^{Q_\tau} \right]_{ij} \sum_{m \neq i} Q_{im} Q_{jm}}{\sum_{i,j} \tau_{ij} \left( \sum_{m \neq i} Q_{im} \left( \sum_{n \neq j} Q_{jn} \right) - \sum_{m \neq i,j} Q_{im} Q_{jm} \right)} \]

As can be seen, \( \frac{C}{D}(\tau) \) is independent of \( \delta \) as long as \( \delta \) is sufficiently short. For real data, \( \delta \) is must be sufficiently large so that sufficient substitutions can occur. For the theoretical models we can take the limit as \( \delta \to 0 \), in which case it is the ratio of substitution rates rather than the ratio of substitution probabilities.

Some of the phenomenological models incorporate spatial variation, in that some sites evolve faster or slower (when a Gamma distribution of rates are used in the GTR or Z models) or when there are completely different mechanisms at different sites (as in the CAT model). This is incorporated into the calculations by considering that there is a set of substitution models where substitution process \( k \) is characterized by substitution rate \( Q^k \) and equilibrium frequencies \( \pi^k \). If \( P_k \) is the proportion of sites characterized by substitution process \( k \), then we can write

\[ \frac{C}{D}(\tau) = \frac{\sum_k P_k \sum_{i,j} \tau_{ij} \left[ e^{Q_\tau} \right]_{ij} \sum_{m \neq i} Q_{im}^k Q_{jm}^k}{\sum_k P_k \sum_{i,j} \tau_{ij} \left( \sum_{m \neq i} Q_{im}^k \left( \sum_{n \neq j} Q_{jn}^k \right) - \sum_{m \neq i,j} Q_{im}^k Q_{jm}^k \right)} \]

Note that, as we are counting the total number of convergent versus divergent changes at each branch over all sites, \( C(\tau) \) and \( D(\tau) \) are independently summed over substitution models.

**Averages: Stokes Fisher model**

For calculating the expected value of \( \frac{C}{D}(\tau) \) for the Stokes Fisher model, we performed one hundred sets of evolutionary simulations, where each set consisted of ten different lineages arranged as a star phylogeny, each lineage in the set starting
from the same initial sequence. This provided us with \(100 \frac{10!}{8!2!}\) pairs of lineages.

Consider lineages \(k\) and \(l\), both having evolved for evolutionary time \(\frac{\tau}{2}\) from the same initial sequence, i.e., they are separated by evolutionary distance \(\tau\), as in Figure 1. At these instances in evolutionary time, we know the identity of the amino acid at each site \(s\) in the protein \((A(k, s, \frac{\tau}{2})\) and \(A(l, s, \frac{\tau}{2})\) in lineages \(k\) and \(l\), respectively) and the Stokes Fisher model provides us with the instantaneous substitution rate matrices at this site in these lineages at that time, \(Q^{k, s, \frac{\tau}{2}}\) and \(Q^{l, s, \frac{\tau}{2}}\). We can then compute average values of \(C(\tau)\) and \(D(\tau)\) by summing over all sites, with the resulting \(C/D(\tau)\) averaged over all pairs of lineages \(\langle k, l \rangle\) with a common ancestral sequence:

\[
\frac{C}{D}(\tau) = \left( \frac{\sum_{s} \sum_{m=0}^{\tau} Q^{k, s, \frac{\tau}{2}}_{A(k, s, \frac{\tau}{2})m} Q^{l, s, \frac{\tau}{2}}_{A(l, s, \frac{\tau}{2})m} - \sum_{m=0}^{\tau} Q^{k, s, \frac{\tau}{2}}_{A(k, s, \frac{\tau}{2})m} Q^{l, s, \frac{\tau}{2}}_{A(l, s, \frac{\tau}{2})m}}{\sum_{s} \sum_{m=0}^{\tau} Q^{k, s, \frac{\tau}{2}}_{A(k, s, \frac{\tau}{2})s} Q^{l, s, \frac{\tau}{2}}_{A(l, s, \frac{\tau}{2})s}} \right)_{\langle k, l \rangle}
\]

\[(S.5)\]

**Estimating the C/D ratio with Distance**

We modeled each data point as the result of a Bernoulli process with C+D trials and C events. The probability that any double substitution is convergent can be derived from the overall C/D ratio for the window \((R_w)\) using Equation S.6. Therefore the likelihood of the data point with C convergences and C+D total double substitutions in window \(w\) with a probability above can be calculated using the binomial distribution. The MCMCs were allowed to run for 97,000 generations after 3,000 generations of burnin to achieve an effective sample size around 1,000. We integrated over phylogenetic uncertainty by sampling substitution data from PLEX every 100 generations, on average.

\[
P_{\text{converge}}(w) = \frac{R_w}{R_w + 1} \quad \text{(S.6)}
\]

\[
P(C, D|w) = \left( \frac{C + D}{C} \right) P_{\text{converge}}(w)^C \left(1 - P_{\text{converge}}(w)\right)^D \quad \text{(S.7)}
\]

**Linear Fit Common Ancestor C/D ratio**

We also wanted to quantify our confidence in the downward slope of the data in Figure 2A. We estimated the slope \((m)\) and intercept \((b)\) of the C/D ratios \((R)\) in Figure 2A using an MCMC determine the credible intervals (see Equation S.8). Again we modeled each data point as the result of C+D trials of a Bernoulli process and C events, however the probability used in the likelihood calculation was derived from the linear model of the C/D ratio. The probability of a data point with C
convergences, D divergences, and at a distance of t resulting from a model with slope m and intercept b is shown in Equation S.10. The MCMC ran for 14,000 generations after 1,000 generations of burnin to achieve an effective sample size of 173 for the slope and 144 for the intercept.

\[
R(t|m,b) = m \cdot t + b
\]

(S.8)

\[
P_{\text{converge}}(t|m,b) = \frac{R(t|m,b)}{R(t|m,b) + 1} = \frac{m \cdot t + b}{m \cdot t + b + 1}
\]

(S.9)

\[
P(C,D,t|m,b) = \left(\frac{C+D}{C}\right)P_{\text{converge}}(t)^C(1 - P_{\text{converge}}(t))^D
\]

(S.10)

**Table S1**

| Window | Distance Range | C/D ratio 99% Credible Interval | Effective Sample Size |
|--------|----------------|---------------------------------|-----------------------|
| 1      | 0 - 0.105      | 0.47 - 0.49                     | 1463                  |
| 2      | 0.105 - 0.21   | 0.32 - 0.34                     | 966                   |
| 3      | 0.21 - 0.315   | 0.26 - 0.27                     | 865                   |
| 4      | 0.315 - 0.42   | 0.23 - 0.23                     | 1020                  |
| 5      | 0.42 - 0.525   | 0.21 - 0.22                     | 1073                  |
| 6      | 0.525 - 0.63   | 0.20 - 0.21                     | 1025                  |
| 7      | 0.63 - 0.735   | 0.19 - 0.20                     | 839                   |
| 8      | 0.735 - 0.84   | 0.18 - 0.19                     | 936                   |
| 9      | 0.84 - 0.945   | 0.18 - 0.19                     | 965                   |
| 10     | 0.945 - 1.05   | 0.18 - 0.19                     | 979                   |
| 11     | 1.05 - 1.155   | 0.18 - 0.18                     | 937                   |
| 12     | 1.155 - 1.26   | 0.17 - 0.18                     | 939                   |
| 13     | 1.26 - 1.365   | 0.17 - 0.18                     | 914                   |
| 14     | 1.365 - 1.47   | 0.17 - 0.18                     | 1070                  |
| 15     | 1.47 - 1.575   | 0.17 - 0.18                     | 1050                  |
| 16     | 1.575 - 1.68   | 0.16 - 0.17                     | 1142                  |
| 17     | 1.68 - 1.785   | 0.16 - 0.17                     | 1132                  |
| 18     | 1.785 - 1.89   | 0.16 - 0.18                     | 1200                  |
| 19     | 1.89 - 1.995   | 0.15 - 0.18                     | 1517                  |
| 20     | 1.995 - 2.1    | 0.11 - 0.19                     | 885                   |
SUPPLEMENTARY FIGURE LEGENDS

Figure S1: The mitochondrial COI phylogenetic tree. The tetrapod mitochondrial tree shown was derived from COI sequences as described in the methods. This is the tree that was used in all analyses. Due to the large number of taxa, each individual taxon is not labeled, but clades that are labeled are squamates (blue), birds (green), crocodiles (cyan), turtles (yellow), mammals (orange), and amphibians (purple).

Figure S2: The 99% credible region of convergence ratio in mitochondrial proteins. The convergence ratios were estimated for the data in Figure 1, divided into 20 windows. The 99% credible regions (red) are extremely tight showing that we can estimate the C/D ratios for each window very well. The mean C/D is shown in black. The C/D for first few windows are clearly decreasing, while the rest of the windows slowly decrease.

Figure S3: The 99% credible linear fits of the common ancestor convergence ratio. The slope of a linear model fit to the mitochondrial data in Figure 2A was estimated. The slope never approaches zero and is always negative with a p-value of less than 0.01. The mean is shown in black and the 99% credible region is shown in red.

Figure S4: Randomized convergence ratios. The data and visualization are the same as in Figure 1, except that convergence events were randomized across sites.

Figure S5: Convergence ratios at conserved sites for common and different amino acids. The data and visualization are the same as in Figure 3A, except that events were separated into two categories depending on whether the ancestral amino acid at a site was the same (A) or different (B).

Figure S6: Convergence ratios at variable sites for common and different amino acids. The data and visualization are the same as in Figure 3B, except that events were separated into two categories depending on whether the ancestral amino acid at a site was the same (A) or different (B).

Figure S7: Comparison of alignment methods. The main analyses were performed using the program ClustalX, but we wanted to determine whether there were major differences using a different alignment method. We therefore ran an analysis of the COI gene with an alignment using PRANK for comparison. The downward sloping curve is also seen with the PRANK results on COI, although there are differences between the two curves early on. The 99% credible region is shown for the C/D ratios using ClustalX on the COI dataset (blue) and PRANK on COI (red). The C/D curves for ClustalX and PRANK are in dark blue and dark red, respectively.

Figure S8 Analysis of inference errors due to the phylogenetic tree. Here, we show the differences between the actual C/D ratios for data simulated on the mtDNA tree compared to inferred C/D ratios with ancestral inferences obtained using mtMAM, in the same fashion as the natural mtDNA data. The inferred data is the same as shown in Figure 4 and Figure 6 of the main manuscript. The data for the WAG, CAT and SF models are shown in A), B), and C), respectively. The actual C/D ratios are shown in dark blue (99% credible regions in blue), while the inferred C/D
ratios are shown in dark red (99% credible regions in red). We note that in many parts of the curves, especially near the beginning, the main trend line is obscured because the 99% credible region is not very wide.

**REFERENCES**

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725-736.

Jukes T, Cantor C. 1969. "Evolution of protein molecules.". In: Munro H, editor. Mammalian Protein Metabolism, vol. III. New York: Academic Press. p. 21-132.

Quang IS, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24: 2317-2323. doi: 10.1093/bioinformatics/btn445

Yang Z. 2014. Molecular Evolution: A Statistical Approach: Oxford University Press.
Distance Between Branches

Convergence / Divergence

Distance Between Branches

Convergence / Divergence
A

Convergence / Divergence

Distance Between Branches

B

Convergence / Divergence

Distance Between Branches

Density
