Supplemental Information

February 18, 2010

Note 1: Notation for the TEB method

Bayes' theorem (Eq. 1) states the posterior probability of a hypothesis H given data D equals the conditional probability of D given H, weighted by the prior probability of H, divided by the total probability of the data. (The total probability of the data can also be expressed as the prior-weighted conditional probability of the D summed over all hypotheses.)

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)} = \frac{P(D|H) \times P(H)}{\sum\limits_{H} P(D|H) \times P(H)} \tag{1}$$

The TEB method for ancestral state reconstruction, described in the section "Materials and Methods" of the main text, applies Bayes' theorem to calculate the posterior probability that some ancestral node contained state a at a sequence site of interest, given the data d at that site, an evolutionary model, and its parameters. The TEB posterior probability of ancestral state a is the weighted average of the posterior probability of a over all possible trees, with the posterior probability of a on each tree t being weighted by the empirical Bayes posterior probability of t. The empirical Bayes posterior probability P_{EB} of a tree assumes the maximum likelihood estimate of branch lengths and other model parameters $\hat{\theta}_t$ on each tree [Kolaczkowski and Thornton, 2008], [Kolaczkowski and Thornton, 2009].

$$P_{TEB}(a|d, m) = \sum_{t} P(a|d, t, m, \hat{\theta}_t) \times P_{EB}(t|d, m, \hat{\theta}_t) \tag{2}$$

In Eq. 2, the first factor inside the summation is the posterior probability of observing the ancestral state a, given the data at that site and topology t. This subexpression can be rewritten using Bayes' theorem:

$$P(a|d, t, m, \hat{\theta}_t) = \frac{P(d|a, t, m, \hat{\theta}_t)\pi_a}{\sum\limits_{a} P(d|a, t, m, \hat{\theta}_t)\pi_a} = \frac{P(d|a, t, m, \hat{\theta}_t)\pi_a}{P(d|t, m, \hat{\theta}_t)} \tag{3}$$

where π_a is the prior probability of hypothesis a, defined as the equilibrium state frequency of state a.

The second factor inside the summation of Eq. 2 is the empirical Bayes posterior probability of tree t, given the data at the site of interest. We use Bayes' theorem to rewrite this subexpression:

$$P_{EB}(t|d,m,\hat{\theta}_t) = \frac{P(d|t,m,\hat{\theta}_t)P(t)}{\sum\limits_t P(d|t,m,\hat{\theta}_t)P(t)} \tag{4}$$

We can use Equations 3 and 4 to expand Eq. 2 as follows:

$$P_{TEB}(a|d,m) = \sum_t \left[\frac{P(d|a,t,m,\hat{\theta}_t)\pi_a}{P(d|t,m,\hat{\theta}_t)} \times \frac{P(d|t,m,\hat{\theta}_t)P(t)}{\sum\limits_t P(d|t,m,\hat{\theta}_t)P(t)} \right] \tag{5}$$

The product of fractions in Eq. 5 can be simplified by canceling like factors in numerator and denominator:

$$P_{TEB}(a|d,m) = \sum_t \left[\frac{P(d|a,t,m,\hat{\theta}_t)\pi_a \times P(t)}{\sum\limits_t P(d|t,m,\hat{\theta}_t)P(t)} \right] \tag{6}$$

$$= \frac{\sum\limits_t P(d|a,t,m,\hat{\theta}_t)\pi_a P(t)}{\sum\limits_t P(d|t,m,\hat{\theta}_t)P(t)} \tag{7}$$

The denominator in Eq. 7 can be rewritten as as the total probability of the data given the maximum likelihood model parameters on each tree, summed over all possible ancestral states for a:

$$\sum_t P(d|t,m,\hat{\theta}_t)P(t) = \sum_t \sum_a P(d|a,t,m,\hat{\theta}_t)\pi_a P(t) \tag{8}$$

Substituting, we have

$$P_{TEB}(a|d,m) = \frac{\sum\limits_t P(d|a,t,m,\hat{\theta}_t)\pi_a P(t)}{\sum\limits_t \sum\limits_a P(d|a,t,m,\hat{\theta}_t)\pi_a P(t)} \tag{9}$$

Eq. 9 formulates the TEB method in form similar to that used by [Pagel et al., 2004], [Huelsenbeck and Bollback, 2001].

extant state pattern	**clade correct**	**mem. +**	**mem. -**	**all**
all: ML	**10.201**	4.823	3.344	**9.439**
all: EB	17.762	**2.191**	**2.532**	14.896
xxx: ML	**0.051**	0.002	0.001	**0.037**
xxx: EB	0.275	**0.001**	0.001	0.141
xxy: ML	**2.014**	0.081	0.081	**3.022**
xxy: EB	14.210	**0.013**	**0.024**	17.835
xyx: ML	1.809	**0.209**	0.464	**1.676**
xyx: EB	**1.565**	0.464	**0.274**	2.109
yxx: ML	2.645	**0.062**	0.214	**4.766**
yxx: EB	**2.630**	0.081	**0.205**	5.355
xyz: ML	**10.347**	4.475	**2.333**	**7.839**
xyz: EB	19.703	**1.597**	2.568	15.317

Table 1: χ^2 statistics for the ultrametric four-taxon simulation measures the fit between the function $f(x) = y$ and the points in published Figure 7 of the main text. The chi-square calculation is weighted because the bins (along the X axis) each contain different numbers of inferences; some bins contain more than 10,000 state predictions, while other bins contain less than 100 predictions. We calculated the weighted chi-square statistic as follows: $\chi^2 = \sum_{i=1}^{n} \frac{B_i(O_i - E_i)^2}{E_i}$, where n is the number of bins, B_i is the number of inferences within bin i, O_i is the observed proportion of correct inferences for bin i, and E_i is the expected proportion of correct inferences for bin i. Lower χ^2 scores correspond to more accurate posterior probability values. In this table, every ancestral state inferences from every replicate was sorted according to the same criteria in published Figure 4B of the main text. The top row expresses χ^2 values across all descendant state patterns. The right-most column express χ^2 values across all membership patterns.

References

[Huelsenbeck and Bollback, 2001] Huelsenbeck, J. P. and Bollback, J. P. (2001). Empirical and heirarchical bayesian estimation of ancestral states. *Systematic Biology*, 50(3):351–366.

[Kolaczkowski and Thornton, 2008] Kolaczkowski, B. and Thornton, J. W. (2008). A mixed branch length model of heterotachy improves phylogenetic accuracy. *Molecular Biology and Evolution*, 25(6):1054–1066.

[Kolaczkowski and Thornton, 2009] Kolaczkowski, B. and Thornton, J. W. (2009). Long-branch attraction bias and inconsistency in bayesian phylogenetics. *PLoS ONE*, 4(12):e7891.

extant state pattern	**all**	**clade correct**	**mem. +**	**mem. -**
all	1.0000	0.6770	0.1613	0.1617
xxx	0.3098	0.2280	0.0399	0.0421
xxy	0.1864	0.1330	0.0273	0.0261
xyx	0.1324	0.0848	0.0233	0.0243
yxx	0.1324	0.0849	0.0239	0.0237
xyz	0.2390	0.1464	0.0460	0.0466

Table 2: The proportion of sites from the ultrametric four-taxon simulations that exhibit particular state patterns and descendant membership patterns. The data is binned according to rows and columns as described in Figure 8C of the main text.

Node	χ^2	**PP(node)**
62: ML	2.032	0.65
62: EB	**1.699**	
63: ML	4.547	1.00
63: EB	**3.198**	
64: ML	2.820	1.00
64: EB	**2.690**	
82: ML	**3.967**	1.00
82: EB	6.525	
88: ML	0.201	0.94
88: EB	**0.082**	
89: ML	0.573	0.72
89: EB	**0.061**	
90: ML	0.024	0.99
90: EB	**0.023**	
94: ML	**0.428**	1.00
94: EB	1.310	
95: ML	**0.758**	0.75
95: EB	1.953	
104: ML	**0.874**	0.98
104: EB	1.044	
118: ML	**0.842**	1.00
118: EB	1.051	
All nodes: ML	**4.169**	
All nodes: EB	4.477	

Table 3: χ^2 statistics for the steroid-hormone simulation. χ^2 values were calculated as described for Table 1. Lower χ^2 scores correspond to more accurate ASR posterior probability values. The left-most column lists node numbers corresponding to phylogenetic labels in figure 2 of the main text. The right-most column lists the posterior probability (PP) of the corresponding node.

Node	χ^2	**PP(node)**
33: ML	**2.536**	0.61
33: EB	3.463	
36: ML	6.939	0.87
36: EB	**2.989**	
37: ML	**7.682**	1.00
37: EB	8.871	
38: ML	2.477	0.74
38: EB	**2.063**	
39: ML	**2.827**	0.42
39: EB	2.921	
40: ML	1.909	1.00
40: EB	**1.385**	
41: ML	**3.606**	1.00
41: EB	3.725	
42: ML	**3.989**	1.00
42: EB	4.281	
47: ML	**6.129**	0.83
47: EB	6.137	
48: ML	4.496	0.45
48: EB	**3.377**	
52: ML	**7.255**	1.00
52: EB	7.644	

Table 4: χ^2 statistics for the ADH simulation. χ^2 values were calculated as described for Table 1. Lower χ^2 scores correspond to more accurate ASR posterior probability values. The left-most column lists node numbers corresponding to phylogenetic labels in figure 3 of the main text. The right-most column lists the posterior probability (PP) of the corresponding node.

Node	χ^2	**PP(node)**
54: ML	**6.471**	0.62
54: EB	8.928	
55: ML	**3.976**	0.9
55: EB	5.163	
74: ML	**0.015**	0.98
74: EB	0.017	
75: ML	0.016	0.36
75: EB	0.016	
78: ML	0.062	0.99
78: EB	**0.047**	
79: ML	**0.051**	0.2
79: EB	0.057	
95: ML	**3.772**	–
95: EB	4.033	
All nodes: ML	**6.282**	
All nodes: EB	7.746	

Table 5: χ^2 statistics for the GFP simulation. χ^2 values were calculated as described for Table 1. Lower χ^2 scores correspond to more accurate ASR posterior probability values. The left-most column lists node numbers corresponding to phylogenetic labels in figure 4 of the main text. The right-most column lists the posterior probability (PP) of the corresponding node.

[Pagel et al., 2004] Pagel, M., Meade, A., and Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology*, 53(5):673–684.

Node	χ^2	**PP(node)**
54: ML	1.018	0.99
54: EB	1.131	
98: ML	0.016	1.0
98: EB	0.015	
55: ML	0.099	0.29
55: EB	0.103	
56: ML	0.077	0.02
56: EB	0.063	
92: ML	0.247	0.99
92: EB	0.247	
93: ML	0.275	0.99
93: EB	0.275	
91: ML	0.007	0.73
91: EB	0.007	
90: ML	0.260	1.0
90: EB	0.260	
57: ML	0.018	0.37
57: EB	0.018	
59: ML	0.187	0.45
59: EB	0.194	
76: ML	0.000	0.69
76: EB	0.000	
75: ML	0.000	1.0
75: EB	0.000	
69: ML	0.011	0.33
69: EB	0.011	
68: ML	0.039	–
68: EB	0.039	
65: ML	0.137	1.0
65: EB	0.137	
66: ML	0.513	0.23
66: EB	0.513	
64: ML	0.101	–
64: EB	0.098	
88: ML	5.265	1.0
88: EB	5.388	
89: ML	1.012	1.0
89: EB	1.012	
87: ML	3.771	0.69
87: EB	3.854	
86: ML	1.689	0.99
86: EB	1.657	
All nodes: ML	**4.448**	
All nodes: EB	4.465	

Table 6: χ^2 statistics for the EF-Tu simulation. χ^2 values were calculated as described for Table 1. Lower χ^2 scores correspond to more accurate ASR posterior probability values. The left-most column lists node numbers corresponding to phylogenetic labels in figure 5 of the main text. The right-most column lists the posterior probability (PP) of the corresponding node.

Simulation	n	j	df	T-value	P-value
ultrametric four-taxon	3200	2	6399	-0.06911	0.9449
non-ultrametric four-taxon	1000	2	1999	-0.07420	0.9409
ADH	100	12	1199	-0.07997	0.9363
steroid hormone receptor	100	12	1199	-0.08099	0.9355
EF-Tu	100	28	2799	-0.07161	0.9428
GFP	100	14	1399	-0.07540	0.9399

Table 7: T-values testing the hypothesis that the mean accuracy of the ML method is significantly different than the mean accuracy of the TEB method. The column titled n shows the number of replicates, j shows the number of ancestral nodes reconstructed on the given phylogeny, and *df* shows the degrees of freedom (calculated as $n \times j$). The column titled *T-value* shows the result of a paired two-sample T test. T-values were computed as the mean of paired differences among replicates, divided by the standard error of the mean of paired differences. The column titled *P-value* shows the statistical significance of the corresponding T value.