



Character-state space versus rate of evolution in phylogenetic inference

Mark P. Simmons,^{a,*} Aaron Reeves^a and Jerrold I. Davis^b

^a*Department of Biology, Colorado State University, Fort Collins, CO 80523, USA*

^b*L.H. Bailey Hortorium, Department of Plant Biology, Cornell University, Ithaca, NY 14853, USA*

Accepted 30 January 2004

Abstract

With only four alternative character states, parallelisms and reversals are expected to occur frequently when using nucleotide characters for phylogenetic inference. Greater available character-state space has been described as one of the advantages of third codon positions relative to first and second codon positions, as well as amino acids relative to nucleotides. We used simulations to quantify how character-state space and rate of evolution relate to one another, and how this relationship is affected by differences in: tree topology, branch lengths, rate heterogeneity among sites, probability of change among states, and frequency of character states. Specifically, we examined how inferred tree lengths, consistency and retention indices, and accuracy of phylogenetic inference are affected. Our results indicate that the relatively small increases in the character-state space evident in empirical data matrices can provide enormous benefits for the accuracy of phylogenetic inference. This advantage may become more pronounced with unequal probabilities of change among states. Although increased character-state space greatly improved the accuracy of topology inference, improvements in the estimation of the correct tree length were less apparent. Accuracy and inferred tree length improved most when character-state space increased initially; further increases provided more modest improvements.

© The Willi Hennig Society 2004.

Limited character-state space is one of the disadvantages of using nucleotide characters for phylogenetic inference (Lanyon, 1988; Mishler et al., 1988; Brooks and McLennan, 1994). With only four possible character states, parallelisms and reversals are expected to occur frequently. The potential character-state space can be particularly limited at nucleotide positions for which there are strong selective constraints to code for a particular amino acid, as well as in lineages with radically skewed base compositions (Meyer, 1994), such as the mitochondrial genomes of honeybees and ticks (Black and Roehrdanz, 1998). Because of their comparatively slowly evolving nature, first and second codon positions have often been favored over third codon positions for phylogenetic inference (e.g. Meyer and Wilson, 1990; Edwards et al., 1991). However, first and second codon

positions may retain less phylogenetic signal than third codon positions because of their selective constraints (Irwin et al., 1991). For example, Naylor et al. (1995) ascribed the limited potential character-state space of many mitochondrial second codon positions to their requirement of coding for hydrophobic residues—for which 24 of the 26 codons have a thymine or cytosine at the second codon position—in segments of genes encoding membrane-spanning amino acids. As Naylor et al. (1995, p. 565) stated:

If the “character state space” (the number of possible states a site can exhibit) of a slowly evolving site is highly constrained, it could retain less phylogenetic information than a more rapidly evolving site for which more character states are available, because the probability that multiple substitutions will result in chance matches (homoplasy) across taxa increases as character-state space becomes more tightly constrained.

Given sufficiently low constraints, even rapidly evolving characters may be useful for inferring ancient cladogenic events (Ortí and Meyer, 1996). Indeed,

*Corresponding author: Department of Biology, Colorado State University, Fort Collins, Colorado 80523, USA.
E-mail address: psimmons@lamar.colostate.edu.

several studies have documented a greater phylogenetic signal from third codon positions relative to their corresponding first and second codon positions—even for basal clades (e.g. Manhart, 1994; Lewis et al., 1997; Björklund, 1999; Källersjö et al., 1999; Wenzel and Siddall, 1999; Campbell et al., 2000; Sennblad and Bremer, 2000; Simmons et al., 2002).

The relative increase in character-state space is one of the three advantages (the others being less sensitive to changes in nucleotide composition, and avoiding the potential saturation of silent substitutions) of using amino acid characters (Simmons, in press), which have 20 potential alternative states, instead of nucleotide characters for phylogenetic inference (Albert et al., 1994; Frohlich and Parker, 2000). However, there are functional constraints on proteins that limit the available character-state space at any given amino acid position (Dayhoff et al., 1972; Miyamoto and Fitch, 1995; Naylor and Gerstein, 2000). For example, in their study of 35 empirical matrices, Simmons et al. (submitted) found that the percentage increase in character-state space for parsimony-informative amino acid characters relative to parsimony-informative nucleotide characters ranged from –5% to 160%, with an average of only 50.4%.

Limited character-state space has been described as restricting our ability to infer divergence among taxa (Chapman et al., 1979), a reason for greater homoplasy in transitions than transversions (Broughton et al., 2000), a basis for differential performance among genes (Davis et al., 1998; Pritchitko and Moore, 2000), and as a reason to explain our difficulty in reconstructing ancient phylogenetic relationships (Baldauf, 2003). With sufficiently large character-state space, parallelisms and reversals become rare, and parsimony will be statistically consistent (Steel and Penny, 2000).

We expect that the only way for faster evolving characters to have a greater phylogenetic signal than a greater number of slower evolving characters is if the faster evolving characters have a greater character-state space (assuming the same number of total substitutions for a greater number of slowly evolving characters). This expectation should hold given identical relative frequencies of character states, relative probabilities of change among states, and identical shifts (if any) in character-state frequencies among lineages. In this project, we addressed two questions. First, how do character-state space and rate of evolution affect the frequencies of convergences, parallelisms, and reversals—the bases for homoplasy and long-branch attraction (Felsenstein, 1978a)? For example, is twice the rate of evolution always offset by twice the character-state space with respect to the accuracy of phylogenetic inference? (Herein, accuracy is used to refer to the ability to reconstruct the model tree topology and/or the simulated amount of evolution.) Second, how is the relationship

between character-state space and rate of evolution affected by differences in: tree topology, branch lengths, rate heterogeneity among sites, probability of change among states, and frequency of character states? We addressed these two questions using simulations. These questions are relevant when considering the advantage of increased character-state space for third codon positions relative to first and second codon positions, as well as the increased state space for amino acids relative to nucleotides.

Materials and methods

For any given rate of evolution, given the same number of characters but different character-state space, the same number of substitutions occurs. However, our ability to infer those substitutions varies depending on the character-state space of those characters. We performed simulations in which we varied the number of potential character states while keeping the expected number of substitutions constant. The number of unobserved substitutions was determined by subtracting the most-parsimonious tree lengths from the expected number of substitutions that occurred when simulating the evolution of the characters for that matrix. The relative frequencies of convergences and reversals were estimated using the ensemble consistency (CI; Kluge and Farris, 1969) and retention indices (RI; Farris, 1989) for each matrix.

Matrices were simulated using the Evolver program within the PAML suite (Yang, 1997). The “MCaa.dat” parameter file, which allows up to 20 character states, was used with a proportional model of evolution (i.e., the probability of change from one character state to another is proportional to their frequencies). Character-state space was varied from two to 20 states, in increments of one for two through six, and increments of two from six through 20. One hundred replicate matrices were simulated for each set of model parameters. One-thousand characters were simulated for each matrix to reduce stochastic errors caused by use of fewer characters, and to emulate the number of characters that are generally available in single gene-tree analyses. The initial simulation was based on equal rates among sites, equal frequencies of character states, and equal probabilities of change among character states, following the Jukes and Cantor (1969) model.

Characters were simulated onto an unrooted, fully symmetrical tree with 32 terminals, in which all branches were of equal length. Thirty-two terminals were selected to represent a biologically interesting sample size (i.e., representative of many empirical studies), and to allow thousands of separate tree searches to be computationally tractable, given the exponential increase in numbers of possible trees with increasing numbers of terminals

Table 1
Simulations performed using different model parameters and character-state spaces

Simulation model			
Rates among sites	Frequencies of states	Prob. of change among states	State spaces examined
equal	equal	equal	all
gamma $\alpha = 5.0$	equal	equal	all
gamma $\alpha = 0.5$	equal	equal	all
equal	half + 25%, half - 25%	equal	2, 4, 6–20
equal	half + 50%, half - 50%	equal	2, 4, 6–20
equal	equal	$\frac{1}{2}$ rate between groups	4, 6–20
equal	equal	$\frac{1}{4}$ rate between groups	4, 6–20

(Felsenstein, 1978b). Characters also were simulated onto an unrooted, completely asymmetrical tree in which all branches were of equal length, as well as a rooted, completely asymmetrical tree in which the terminal branch lengths varied according to a molecular clock (all internal branches were of equal length, at one branch-segment long [see below]). The latter produced very long branches (i.e., branches consisting of many branch segments) leading to “basal” (i.e. early derived) terminals relative to those leading to the “distal” (i.e. recently derived) terminals. This discrepancy in branch lengths, coupled with the comparatively short internal branch lengths, made long-branch attraction likely to occur. Each of the symmetrical and asymmetrical trees had 61 total branches, of which 32 were terminal branches and 29 were internal branches.

The rate of evolution per branch segment was varied from 0.01 to 0.5, in increments of 0.05 (except for the first increment from 0.01 to 0.05), for a total of 11 rates. There were 61 branch segments (corresponding to the 61 total branches) for the symmetrical tree and the unrooted, asymmetrical tree with equal branch lengths. In contrast, there were 526 branch segments for the rooted, asymmetrical tree with unequal branch lengths according to a molecular clock. The expected number of changes per character ranged from 0.61 to 30.5 for the symmetrical tree and asymmetrical tree with equal branch lengths, and 5.26–263 for the asymmetrical tree with unequal branch lengths. These numbers of changes per character were selected so as to bracket the biologically interesting (i.e. reflective of variation in empirical data) number of character-state changes for a matrix of 32 terminals, and to produce matrices for which the correct topology would be difficult to reconstruct. At extremely low rates of evolution, in which parsimony-informative characters undergo a single character-state change across the entire tree, increases in character-state space are trivial because no convergences or reversals can occur.

In the context of differential character-state space, the complexity of the simulation model was increased to examine the effects of: (1) rate heterogeneity, (2) unequal frequencies of character states, and (3)

unequal probabilities of change among characters on the ability to accurately infer the amount of evolutionary change (i.e. tree lengths), homoplasy, and topological accuracy of inferred trees (Table 1). Rate heterogeneity, unequal frequencies of character states, and unequal probabilities of change among characters were examined independently of one another on the symmetrical tree with equal branch lengths. Gamma-distributed rates (Yang, 1993) were used to simulate rate heterogeneity. Moderate ($\alpha = 5$) and extreme ($\alpha = 0.5$, following Hillis’ 1998 and Yang’s 1998 simulations) rate heterogeneities were examined, with eight categories for the discrete gamma.

Unequal frequencies of character states were examined in which half of the character states were increased in frequency by 25% and the other half were decreased by 25%. More severe imbalances in character-state frequencies were then examined in which half of the character states were increased in frequency by 50%, and the other half were decreased by 50%. These simulations were only performed for even numbers of character states, from two to 20. Character-state frequencies were rounded to the nearest hundredth. Because a proportional model was used, character states were more likely to change to states represented in high frequency than to states represented in low frequency. These simulations are analogous to Felsenstein’s (1981) model in which unequal nucleotide frequencies are permitted, while only a single substitution rate is allowed.

Unequal probabilities of change among states were simulated while maintaining equal character-state frequencies, using the approach outlined by Yang et al. (1998). Character states for each simulation were separated into two groups of equal size (only even numbers of character states were examined, from four to 20). Character-state changes between groups were set at half the rate of changes within groups. To increase the discrepancy in rates, character-state changes between groups were separately set at one-quarter the rate of changes within groups. These simulations are analogous to Kimura’s (1980) two-parameter model in which transitions may occur at different rates than transversions.

Equally weighted parsimony tree searches were performed using PAUP* 4.0b10 (Swofford, 2002) with batch files running on PCs. Tree searches for each matrix entailed 100 tree-bisection-reconnection (TBR) searches, each with random taxon addition and a maximum of 1000 trees held. A maximum of 10 000 trees were potentially held for each matrix; trees found in the 100 separate TBR searches were not swapped to completion. PAUP* calculated a strict consensus tree (Schuh and Polhemus, 1980; Sokal and Rohlf, 1981), and output the most parsimonious tree length, CI (excluding parsimony-uninformative characters), and RI for each matrix.

Observed state space was determined for characters evolving at equal rates among sites, with equal frequencies of character states, and equal probabilities of change among character states on: (1) the unrooted, fully symmetrical tree, (2) the unrooted, completely asymmetrical tree in which all branches were of equal length, and (3) the rooted, completely asymmetrical tree in which the terminal branch lengths varied according to a molecular clock. The observed state space was determined using the formula: (number of parsimony-informative characters + minimum tree length)/number of parsimony-informative characters.

PEST v. 2.2 (Zujko-Miller and Miller, 2003) was used to determine the congruence and incongruence between the strict consensus tree and the reference tree (i.e., the tree topology on which the characters were simulated) for each matrix. The maximum congruence score was 29, for all clades correctly resolved. Likewise, the maximum incongruence score was -29, in which all clades from the reference tree were contradicted in the strict consensus tree. Following Simmons and Miya (in press), we measured the performance of phylogenetic inference for each matrix by subtracting the number of clades incorrectly resolved from the number of clades correctly resolved, resulting in a possible range of performance from +29 to -29. No penalty was assessed for unresolved clades in the strict consensus. Note that our use of PEST was not a particularly sensitive approach to measuring phylogenetic signal (i.e., character covariation; Archie, 1989; Faith and Cranston, 1991) in each matrix because the amount of branch support for each clade was not considered, but should be included (Källersjö et al., 1992). However, because 100 replicates were used for each set of model parameters, our approach was sensitive in the sense that weakly supported branches would be unlikely to be resolved in many replicates, whereas strongly supported branches would be consistently resolved in most or all of the 100 replicates. The average congruence, incongruence, tree length, CI, RI, and observed character-state space for each group of 100 replicates were determined using the program CONDENSE, which was written by Aaron Reeves

(available at: <http://www.biology.colostate.edu/Research/>).

Results and discussion

Supplemental data, including simulation parameter files, strict consensus trees, PAUP* log files, PEST outputs, and an Excel file of all of the data and figures are posted at: <http://www.biology.colostate.edu/Research/>. Results for characters with a state space from 2 to 20 with equal frequencies of character states, equal probability of change between character states, and without rate heterogeneity, simulated on a completely symmetrical tree topology with equal length branches, are shown in Fig. 1. Results for characters with a state space from 2 to 20, with equal frequencies of character states, equal probability of change between character states, and without rate heterogeneity, simulated on a completely asymmetrical tree topology with all branches of equal length are presented in Fig. 2. Results for characters with state space from 2 to 20, with equal frequencies of character states, equal probability of change between character states, and without rate heterogeneity, simulated on a completely asymmetrical tree topology with unequal length branches according to a molecular clock are presented in Fig. 3. The overall success of resolution (number of clades correctly resolved minus the number of clades incorrectly resolved) for the 29 clades relative to the rate of evolution, using characters simulated under different parameters, on a completely symmetrical tree topology, are shown in Fig. 4. Results for characters with a state space of four, simulated using different parameters on a completely symmetrical tree topology with equal length branches are presented in Fig. 5 (including results for the completely asymmetrical tree topology). Results for characters with state space of four, simulated using different parameters on a completely asymmetrical tree topology with unequal length branches according to a molecular clock are shown in Fig. 6. The results for characters with a state space of four were chosen as exemplars, because four states was the lowest number included in all sets of simulations, and was applicable to both nucleotide and amino acid characters. The observed state space for characters with a simulated state space from 2 to 20 with equal frequencies of character states, equal probability of change between character states, and without rate heterogeneity are presented in Fig. 7.

Caution is urged when directly extrapolating results from the simulations to empirical matrices because the simulations assume that character-state space remains constant in all lineages throughout time. This assumption is unrealistic as predicted by the covarion theory (Fitch and Markowitz, 1970), and because the

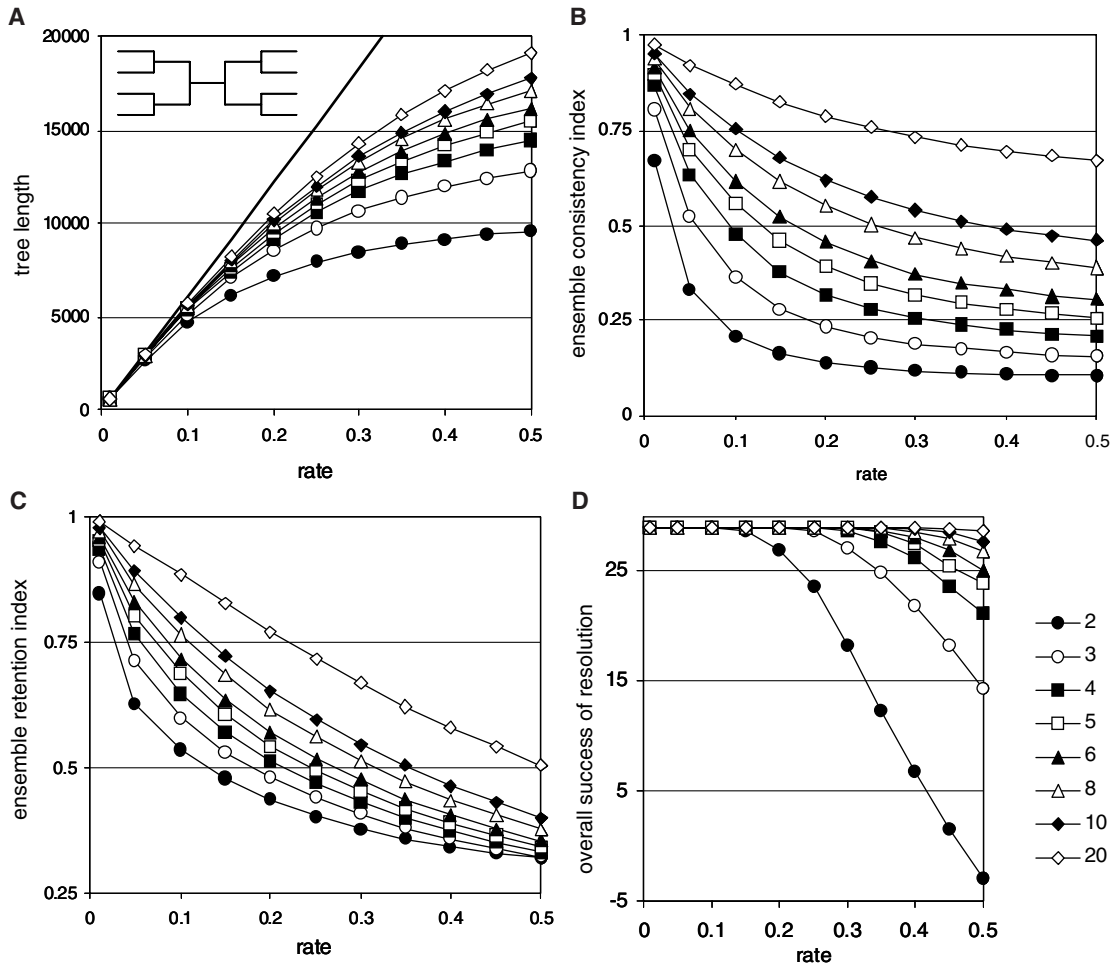


Fig. 1. Results for characters with state space from 2 to 20 with equal frequencies of character states, equal probability of change between character states, and without rate heterogeneity, simulated on a completely symmetrical tree topology with equal length branches. (A) Observed tree lengths plotted relative to the rate of evolution along each branch. The boldface line indicates the average simulated tree length. (B) CIs for parsimony-informative characters relative to the rate of evolution. (C) RIs relative to the rate of evolution. (D) Overall success of resolution (number of clades correctly resolved minus the number of clades incorrectly resolved) for the 29 clades relative to the rate of evolution.

degeneracy of first and third codon positions varies depending on which amino acid the codon specifies at any given time.

Relative advantages of increased character-state space

When the character-state space is low (i.e. two to three), relatively small increases can considerably improve the accuracy of phylogenetic inference. For example, at a rate of evolution of 0.5 per branch for a symmetrical tree with equal length branches, increasing the state space from two to three resulted in an increase from an overall success of resolution of -10% (-2.91 clades) to $+50\%$ (14.37 clades), and an increase from two to four resulted in an increase to $+73\%$ (21.13 clades; Fig. 1D). This impressive increase in accuracy arose despite more modest increases in estimations of

tree length. For example, at a rate of evolution of 0.5 per branch, increasing the state space from two to three resulted in 34% (3251.8) more steps inferred, and an increase from two to four resulted in 51% (4900.75) more steps inferred (Fig. 1A). Although similar results were obtained under a completely asymmetrical tree topology with all branches of equal length (Fig. 2A, D), our results for a completely asymmetrical tree topology with unequal branch lengths according to a molecular clock produced considerably less impressive increases in accuracy at higher rates of evolution, despite comparable increases in inferred tree length. For example, at a rate of 0.5 per branch segment, the overall success of resolution increased from -26.63 to -25.46 , accounting for an average increase in accuracy of just over a single clade (Fig. 3D). Likewise, although there were 43% (4227.16) more steps inferred when increasing the state

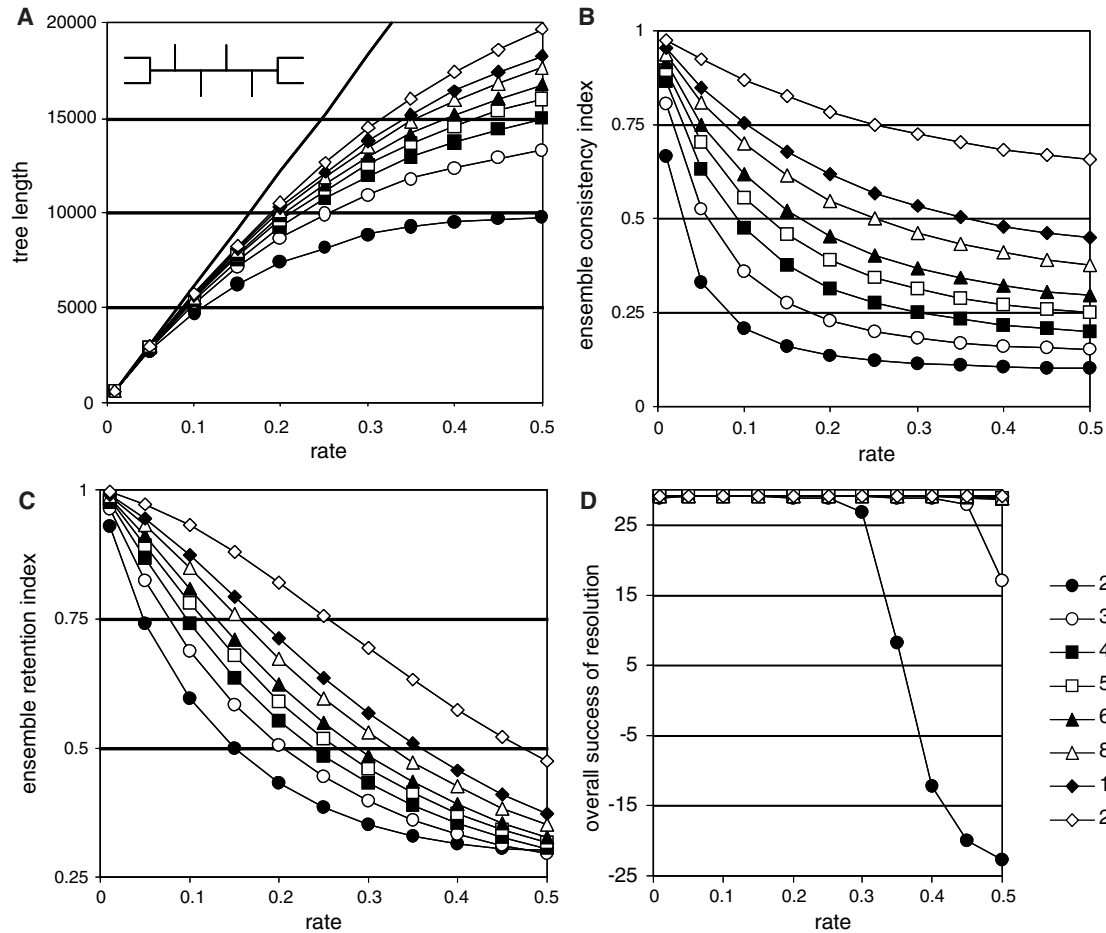


Fig. 2. Results for characters with state space from 2 to 20 with equal frequencies of character states, equal probability of change between character states, and without rate heterogeneity, simulated on a completely asymmetrical tree topology with all branches of equal length. (A) Observed tree lengths plotted relative to the rate of evolution along each branch. The boldface line indicates the average simulated tree length. (B) CIs for parsimony-informative characters relative to the rate of evolution. (C) RIs relative to the rate of evolution. (D) Overall success of resolution for the 29 clades relative to the rate of evolution.

space from two to three at a rate of 0.5 per branch segment, that increase accounted for less than 2% of the steps that actually occurred during the simulations (Fig. 3A). Further increases in character-state space provided more modest improvements in inferred tree lengths. For example, with four times the state space (from five to 20 states), only 34% (6035.93) more steps were inferred at a rate of 0.25 per branch segment on the asymmetrical tree with unequal branch lengths (Fig. 3A). Although dramatic increases in accuracy of resolution were observed at lower rates of evolution with increased state space (e.g. an increase in overall success of resolution from 2.47 to 16.13 at a rate of 0.05 per branch segment when increasing state space from 5 to 20), these increases declined dramatically at higher rates of evolution (e.g. an increase in overall success of resolution from -22.51 to -21.2 at a rate of 0.45 per branch segment when increasing state space from 5 to 20; Fig. 3D).

Given equal branch lengths, increases in state space were more beneficial for the accuracy of resolution at higher rates of evolution on asymmetrical trees than on symmetrical trees. For example, at a rate of 0.3 per branch, an increase of state space from two to four led to a 57% increase in overall success of resolution (to 98.8%) on the symmetrical tree, and an 8% increase in overall success (to 100%) on the asymmetrical tree. However, at a rate of 0.5 per branch, the same increase in state space lead to an increase from an overall success of resolution of -10% (-2.91 clades) to +73% (21.13 clades) on the symmetrical tree, and a still-more-dramatic increase from -78% (-22.69 clades) to +98% (28.55 clades) on the asymmetrical tree (Figs 1D and 2D). Based on inspection of the strict consensus tree files, the dramatic failure of accurate resolution at high rates of evolution on an asymmetrical tree with branches of equal length appears to be caused largely by long-branch attraction—both between closely related

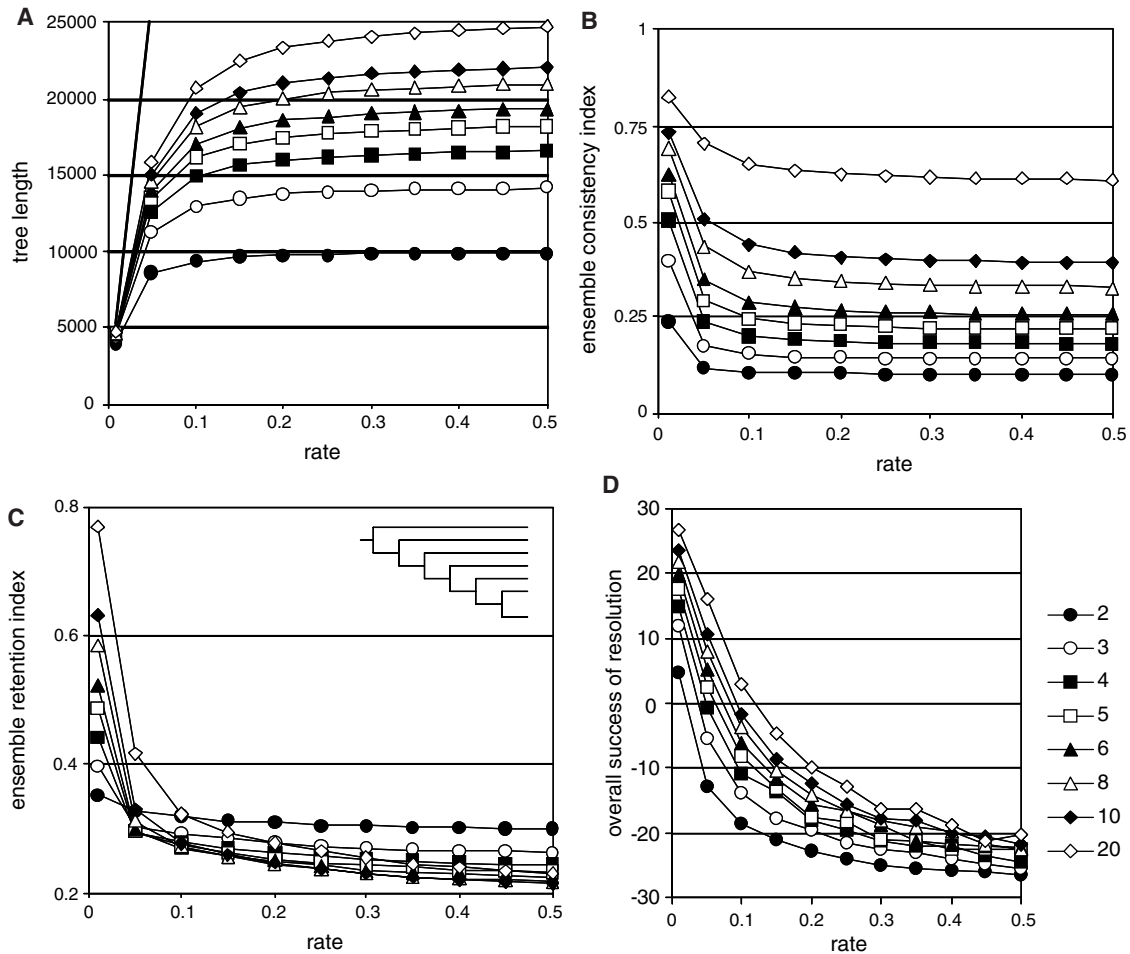


Fig. 3. Results for characters with state space from 2 to 20 with equal frequencies of character states, equal probability of change between character states, and without rate heterogeneity, simulated on a completely asymmetrical tree topology with unequal length branches according to a molecular clock (A–D). (A) Observed tree lengths plotted relative to the rate of evolution along each branch. The boldface line indicates the average simulated tree length. (B) CIs for parsimony-informative characters relative to the rate of evolution. (C) RIs relative to the rate of evolution. (D) Overall success of resolution for the 29 clades relative to the rate of evolution.

terminals as well as between distantly related terminals (results not shown; available as supplemental data).

Rate heterogeneity among characters and between states

As may be expected, increasing rate heterogeneity among characters led to progressively more severe underestimates of the number of substitutions that occurred (Figs 5A, 6A) because increasingly more changes are concentrated at particular characters, leading to multiple hits along individual branches (i.e. saturation). However, increased rate heterogeneity among characters actually can be beneficial for inferring the tree topology (Figs 4A,B, 5D, 6D), as predicted by Hillis (1987). That is, rapidly evolving characters may be useful for resolving recently diverged clades, whereas slowly evolving characters may be beneficial for resolving earlier diverging clades. This result holds for both

the symmetrical tree with equal length branches, as well as the asymmetrical tree with unequal length branches (albeit at higher rates of evolution only; Fig. 6D), and is therefore expected to be a general phenomenon. Despite this, there must be a threshold beyond which the rate heterogeneity becomes too extreme to be beneficial.

Progressively more extreme unequal probabilities of change among states actually can be advantageous for phylogenetic inference (Figs 4E,F, 5D). This may be explained by synapomorphies based on improbable changes being retained, while frequent changes simply cancel each other out as noise (Hillis, 1996; Wenzel and Siddall, 1999). This is the same basic phenomenon as seen for rate heterogeneity among characters, albeit with relatively infrequent changes among particular states of the characters, rather than different probabilities of change amongst the characters themselves. Note that unequal probabilities of change among states is not

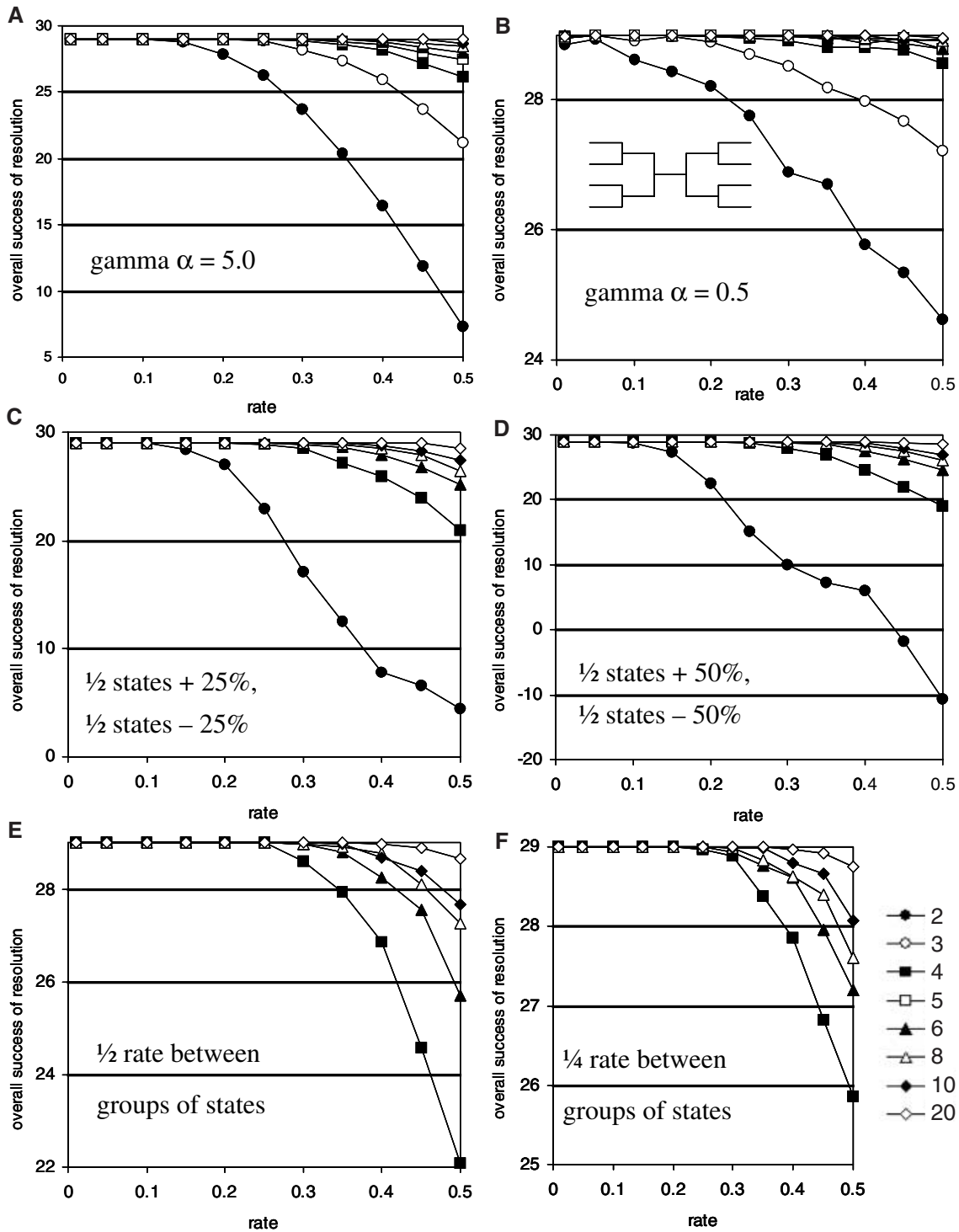


Fig. 4. Overall success of resolution for the 29 clades relative to the rate of evolution, using characters simulated under different parameters, on a completely symmetrical tree topology. (A) Rate variation among sites using a gamma distribution with $\alpha = 5.0$. (B) Rate variation among sites using a gamma distribution with $\alpha = 0.5$. (C) Unequal state frequencies with half of the character states increased in frequency by 25% and the other half were decreased by 25%. (D) Unequal state frequencies with half of the character states increased in frequency by 50% and the other half were decreased by 50%. (E) Unequal probabilities of change among states in which changes between groups were set at half the rate of changes within groups. (F) Unequal probabilities of change among states in which changes between groups were set at one-quarter the rate of changes within groups.

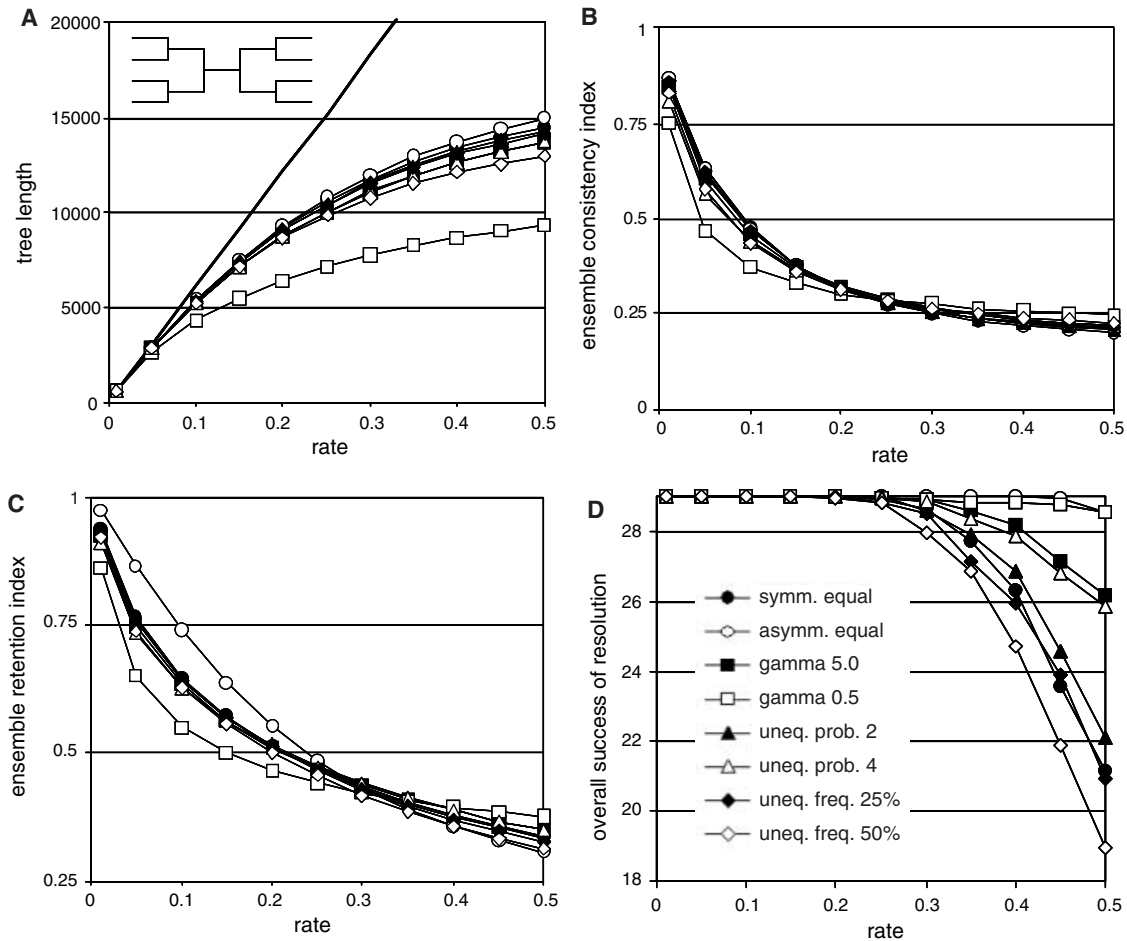


Fig. 5. Results for characters with state space of four simulated using different parameters on a completely symmetrical tree topology (with the exception of the open circles, for a completely asymmetrical tree topology) with equal length branches. (A) Observed tree lengths plotted relative to the rate of evolution along each branch. The boldface line indicates the average simulated tree length. (B) CIs for parsimony-informative characters relative to the rate of evolution. (C) RIs relative to the rate of evolution. (D) Overall success of resolution for the 29 clades relative to the rate of evolution.

universally advantageous for phylogenetic inference, as shown by the lower overall success of resolution at the slowest examined rate of evolution on the asymmetrical tree with unequal branch lengths (Fig. 6D; non-overlapping 95% confidence intervals at rate of 0.01 when comparing simulations in which the probability of change among all states was equal, relative to those in which character-state changes between groups were set at one-quarter the rate of changes within groups). As with rate heterogeneity among characters, there must be a point beyond which increased unequal probabilities of change among states is disadvantageous. Silent transitions generally occur more often than replacement transversions at third codon positions (e.g. Graur and Li, 2000). However, our results indicate that this factor does not necessarily counter the greater character-state space of third codon positions relative to first and second codon positions, at which the difference in

transition-transversion rate may be less extreme. Up to a point, this discrepancy may be advantageous for third codon positions.

Unequal frequencies of character states

In contrast to rate heterogeneity among characters and between character states, unequal frequencies of character states (with a proportional model of character-state change) appear to be generally disadvantageous for phylogenetic inference (Figs 4C,D, 5D, 6D). The only conditions explored here in which unequal state frequencies significantly outperformed equal state frequencies were at very high rates of evolution (0.45 and 0.5) for characters with a state space of two on the symmetrical tree with equal branch lengths, and then only when half of the character states were increased in frequency by 25% and the other half were decreased by

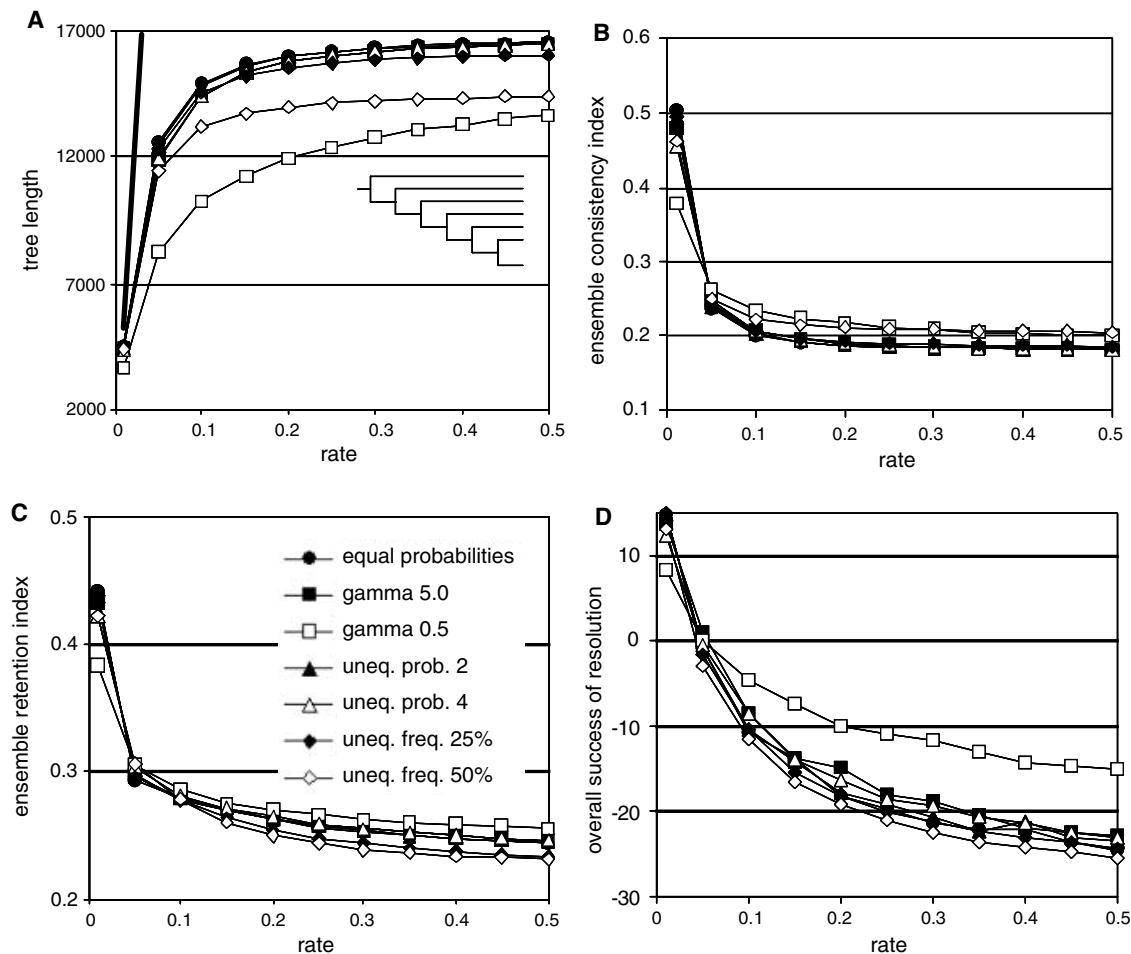


Fig. 6. Results for characters with state space of four simulated using different parameters on a completely asymmetrical tree topology with unequal length branches according to a molecular clock. (A) Observed tree lengths plotted relative to the rate of evolution along each branch. The boldface line indicates the average simulated tree length. (B) CIs for parsimony-informative characters relative to the rate of evolution. (C) RIs relative to the rate of evolution. (D) Overall success of resolution for the 29 clades relative to the rate of evolution.

25% (Fig. 4C; non-overlapping 95% confidence intervals at rates 0.45 and 0.5).

The simulations involving unequal frequencies of character states, with their proportional model of character-state change, share a common feature with the simulations of characters with unequal probabilities of change among states—differential substitution probabilities. The important differentiating factor between these simulations is that with unequal state frequencies, synapomorphies based on changes from the more frequently represented states to less frequently represented states are liable to be rapidly overwritten by reverse substitutions, thereby obscuring those synapomorphies. As such, unequal frequencies of character states are not expected to be a generally advantageous property of characters for phylogenetic inference—even when there is an equal frequency of change between rare and common states (Collins et al., 1994; Perna and Kocher, 1995; Eyre-Walker, 1998).

Shifts in nucleotide composition are generally most pronounced at third codon positions (e.g. Prager and Wilson, 1988; Hasegawa et al., 1993; Klenk and Zillig, 1994; Nishiyama and Kato, 1999; Wirth et al., 1999), at which most substitutions are silent (except in some cases of positive selection). So, in addition to convergent shifts in nucleotide composition leading to the same problem as long-branch attraction, wherein unrelated lineages that derive similar nucleotide compositions can be resolved as sister groups (Lockhart et al., 1992), these shifts in state frequencies can also decrease the advantage of increased character-state space at third codon positions.

Statistical consistency of parsimony

When the character-state space is sufficiently high, it is possible for the overall success of phylogenetic reconstruction to be $\geq 95\%$, even with extremely high rates of

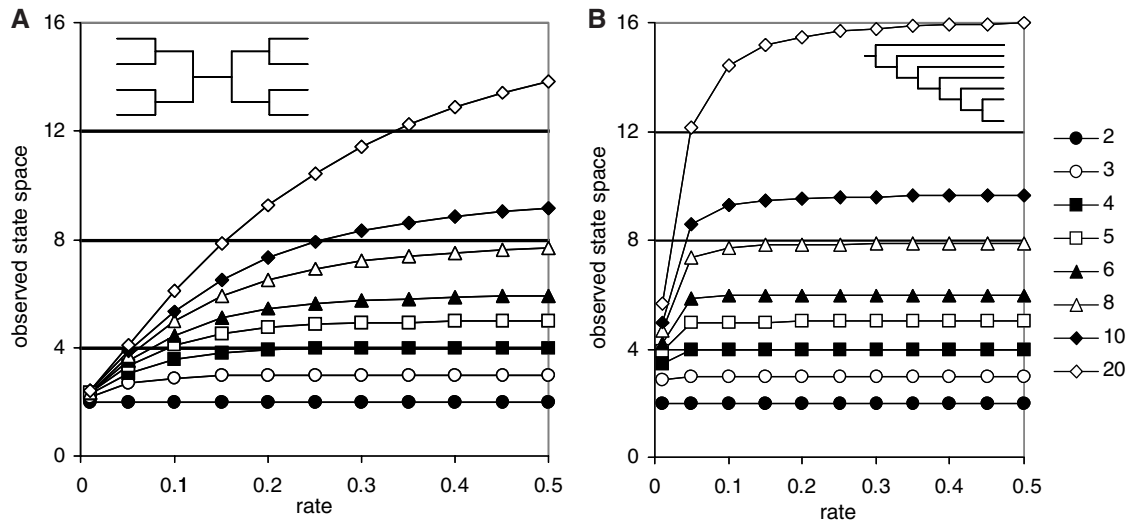


Fig. 7. Observed state space for characters with a simulated state space from 2 to 20 with equal frequencies of character states, equal probability of change between character states, and without rate heterogeneity, simulated on: (A) a completely symmetrical tree topology with equal length branches, and (B) a rooted, completely asymmetrical tree in which the terminal branch lengths varied according to a molecular clock.

evolution. For example, with a state space of 10, the overall success of resolution averaged 95% on the symmetrical tree with equal branch lengths, despite a rate of evolution of 0.5 along each branch (Fig. 1D). This result corresponds with Steel and Penny's (2000) proof that parsimony will be statistically consistent given a sufficiently large character-state space.

However, under certain topology and branch length combinations, an enormous state space may be required to achieve consistency. For example, despite 20 alternative character states, represented in equal frequency, an equal probability of change among states, and a relatively low rate of evolution (0.01 per branch segment), the overall success of resolution averaged 92%, rather than 100%, for the completely asymmetrical tree with unequal branch lengths according to a molecular clock (Fig. 3D). Although part of this (an average of 0.8 clades) could be attributed to a lack of resolution in the strict consensus trees, cases of long-branch attraction were evident (an average of 0.7 clades; results not shown; available as supplemental data).

Consistency and retention indices

Given the same character-state space and model of evolution, decreases in CI and RI were found to be predictive of decreased accuracy of tree length estimation (Figs 1A–C, 2A–C, 3A–C). This pattern also often applied to the accuracy of the overall success of phylogenetic resolution (Figs 1B–D, 2B–D, 3B–D), although relatively low CIs and RIs did not necessarily mean that the tree topologies would be inferred incorrectly. For example, although the CI averaged 0.46, and

the RI averaged 0.4, an average overall success of resolution was 95% with a state space of 10 simulated on the symmetrical tree with equal branch lengths at a rate of evolution of 0.5 (Fig. 1B–D). Likewise, although the CI decreased from 0.67 to 0.16, and the RI decreased from 0.85 to 0.48, when the rate of evolution was increased from 0.01 to 0.15 the average overall success of resolution was 99% with a state space of two simulated on the symmetrical tree with equal branch lengths (Fig. 1B–D). These results are consistent with the many empirical studies that have documented a strong phylogenetic signal from characters with relatively high levels of homoplasy (e.g. Manhart, 1994; Lewis et al., 1997; Björklund, 1999; Källersjö et al., 1999; Wenzel and Siddall, 1999; Campbell et al., 2000; Sennblad and Bremer, 2000; Simmons et al., 2002).

Given the same state space, overall rate of evolution, and underlying tree topology and branch lengths, similar CIs and RIs do not necessarily equate to equivalent performance for tree-topology estimation when groups of characters are evolving under different parameters. This was found to apply to differential state frequencies, probability of change among states, and rate heterogeneity among sites (Fig. 4A–F; CIs and RIs available as supplemental data). Consider the following examples for characters with a state space of two simulated on the symmetrical tree with equal branch lengths, and a rate of 0.5 along each branch. Under equal state frequencies, identical probabilities of change, and no rate heterogeneity, CI = 0.1, RI = 0.32, and the overall success of resolution was –2.91 clades (Fig. 1B–D). Yet, given moderate differential state frequencies and similar overall levels of homoplasy (CI = 0.11,

RI = 0.27), the overall success of resolution increased to +4.45 clades (Fig. 4C). Likewise, given a moderate rate heterogeneity among characters (gamma distribution with $\alpha = 5.0$; Fig. 4A) and similar overall levels of homoplasy (CI = 0.11, RI = 0.33), the overall success of resolution increased to +7.33 clades. These results confirm the importance of not comparing CIs and RIs between groups of characters (e.g. molecular versus morphological, rDNA versus protein-coding genes, nucleotides versus amino acids) to ascertain their relative phylogenetic signal, even when the same taxa are sampled and the same tree topology is examined.

Observed character-state space

Other than the trivial case in which the simulated state space was two, the average observed state space for the parsimony-informative characters was generally less than the simulated state space (Fig. 7). Regardless of the tree topology/relative branch lengths on which the characters were simulated, all characters showed an asymptotic increase in state space as the rate of evolution increased from 0.01 to 0.5 per branch segment. The greater the simulated state space, the higher the rate of evolution that was required before the asymptote leveled off. This indicates that, when counting the number of observed states in empirical data matrices, the state space is more likely to be underestimated for characters that are evolving slowly and for characters with a greater actual state space. Therefore, the actual state space for amino acid characters will generally be more severely underestimated than the actual state space for their corresponding nucleotide characters. The severity of the underestimate for first and second codon positions relative to third codon positions is less predictable, given that third codon positions generally evolve faster and have a greater state space. The severity of these underestimates may be decreased through increased taxon and/or paralog sampling.

Conclusions

Our results indicate that relatively small increases in character-state space, as evident in empirical data matrices when comparing first and second codon positions with third codon positions, and nucleotides with amino acids, can provide enormous benefits for the accuracy of phylogenetic inference. For example, if parsimony-informative second codon positions have a state space of two, while third codon positions have a state space of four, third codon positions could accommodate twice the rate of evolution while providing roughly equivalent phylogenetic signal (Figs 1D, 2D and 3D). This advantage may become more pronounced

with unequal probabilities of change among states, as may be expected to occur at twofold degenerate sites, as well as transitions relative to transversions at fourfold degenerate sites (e.g. Fitch, 1967; Collins and Jukes, 1994; Moriyama and Powell, 1997).

Likewise, although the observed increase in state space for amino acids relative to their underlying nucleotides in empirical data matrices (average of 50.4% for parsimony-informative characters reported by Simmons et al. in press) is considerably less than their potential fivefold increase, that increase can still be tremendously beneficial for the accuracy of phylogenetic inference. However, this benefit of using amino acid characters instead of nucleotide characters must be tempered by the loss of potential phylogenetic signal caused by discarding silent substitutions (Albert et al., 1994), convergence due to degeneracy of the genetic code (Simmons, 2000), and the use of composite characters that can create putative synapomorphies that are not present in any of the corresponding nucleotide characters (when considered individually; Simmons and Freudenstein, 2002). The phylogenetic signal from both nucleotide and amino acid characters for any given protein-coding gene may be incorporated into a simultaneous analysis by using the non-redundant-coding-of-dependent-characters method (Freudenstein et al., 2003).

Acknowledgments

We thank Arnold Kluge and two anonymous reviewers for their helpful suggestions; Jeremy Miller for detailed suggestions and help with PEST; Mike Antolin, John Freudenstein, Damon Little, Kevin Nixon, Chris Randle, Pat Reeves, the CSU Evolution Discussion Group, and the OSU Phylogenetics Discussion Group for helpful discussions; and Ziheng Yang for help implementing the simulations.

References

- Albert, V.A., Backlund, A., Bremer, K., Chase, M.W., Manhart, J.R., Mishler, B.D., Nixon, K.C., 1994. Functional constraints and *rbcL* evidence for land plant phylogeny. *Ann. Mo. Bot. Gard.* 81, 534–567.
- Archie, J.W., 1989. A randomization test for phylogenetic information in systematic data. *Syst. Zool.* 38, 219–252.
- Baldauf, S.L., 2003. The deep roots of eukaryotes. *Science* 300, 1703–1706.
- Björklund, M., 1999. Are third positions really that bad? A test using vertebrate cytochrome b. *Cladistics* 15, 191–197.
- Black, W.C., Roehrdanz, R.L., 1998. Mitochondrial gene order is not conserved in arthropods: prostriate and metastriate tick mitochondrial genomes. *Mol. Biol. Evol.* 15, 1772–1785.
- Brooks, D.R., McLennan, D.A., 1994. Historical ecology as a research programme: scope, limitations and the future. In: Eggleton, P.,

- Vane-Wright, R.I. (Eds.), *Phylogenetics and Ecology*. The Linnean Society of London, London, pp. 1–27.
- Broughton, R.E., Stanley, S.E., Durrett, R.T., 2000. Quantification of homoplasy for nucleotide transitions and transversions and a reexamination of assumptions in weighted phylogenetic analysis. *Syst. Biol.* 49, 617–627.
- Campbell, D.L., Brower, A.V.Z., Pierce, N.E., 2000. Molecular evolution of the *Wingless* gene and its implications for the phylogenetic placement of the butterfly family Riodinidae (Lepidoptera: Papilionoidea). *Mol. Biol. Evol.* 17, 684–696.
- Chapman, R.W., Avise, J.C., Asmussen, M.A., 1979. Character state restrictions and boundary conditions in the evolution of quantitative multistate characters. *J. Theor. Biol.* 80, 51–64.
- Collins, D.W., Jukes, T.H., 1994. Rates of transition and transversion in coding sequences since the human-rodent divergence. *Genomics* 20, 386–396.
- Collins, T.M., Wimberger, P.H., Naylor, G.J.P., 1994. Compositional bias, character-state bias, and character-state reconstruction using parsimony. *Syst. Biol.* 43, 482–496.
- Davis, J.I., Simmons, M.P., Stevenson, D.W., Wendel, J.F., 1998. Data decisiveness, data quality, and incongruence in phylogenetic analysis: an example from the monocotyledons using mitochondrial *atpA* sequences. *Syst. Biol.* 47, 282–310.
- Dayhoff, M.O., Eck, R.V., Park, C.M., 1972. A model of evolutionary change in proteins. In: Dayhoff, M.O. (Ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, pp. 89–99.
- Edwards, S.V., Arctander, P., Wilson, A.C., 1991. Mitochondrial resolution of a deep branch in the genealogical tree for perching birds. *Proc. Roy. Soc., London, Ser. B* 243, 99–107.
- Eyre-Walker, A., 1998. Problems with parsimony in sequences of biased base composition. *J. Mol. Evol.* 47, 686–690.
- Faith, D., Cranston, P., 1991. Could a cladogram this short have arisen by chance alone? *Cladistics* 7, 1–28.
- Farris, J.S., 1989. The retention index and the rescaled consistency index. *Cladistics* 5, 417–419.
- Felsenstein, J., 1978a. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27, 401–410.
- Felsenstein, J., 1978b. The number of evolutionary trees. *Syst. Zool.* 27, 27–33.
- Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376.
- Fitch, W.F., 1967. Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. *J. Mol. Biol.* 26, 499–507.
- Fitch, W.M., Markowitz, E., 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.* 4, 579–593.
- Freudenstein, J.V., Pickett, K.M., Simmons, M.P., Wenzel, J.W., 2003. From basepairs to birdsongs: phylogenetic data in the age of genomics. *Cladistics* 19, 333–347.
- Frohlich, M.W., Parker, D.S., 2000. The mostly male theory of flower evolutionary origins: from genes to fossils. *Syst. Bot.* 25, 155–170.
- Graur, D., Li, W.-H., 2000. *Fundamentals of Molecular Evolution*, 2nd edn. Sinauer Associates, Sunderland, MA.
- Hasegawa, M., Hashimoto, T., Adachi, J., Iwabe, N., Miyata, T., 1993. Early branchings in the evolution of eukaryotes: ancient divergence of *Entamoeba* that lacks mitochondria revealed by protein sequence data. *J. Mol. Evol.* 36, 380–388.
- Hillis, D.M., 1987. Molecular versus morphological approaches. *Ann. Rev. Ecol. Syst.* 18, 23–42.
- Hillis, D.M., 1996. Inferring complex phylogenies. *Nature* 383, 130–131.
- Hillis, D.M., 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Syst. Biol.* 47, 3–8.
- Irwin, D.M., Kocher, T.D., Wilson, A.C., 1991. Evolution of the cytochrome b gene of mammals. *J. Mol. Evol.* 32, 128–144.
- Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.), *Mammalian Protein Metabolism*. Academic Press, New York, pp. 21–132.
- Källersjö, M., Albert, V.A., Farris, J.S., 1999. Homoplasy increases phylogenetic structure. *Cladistics* 15, 91–93.
- Källersjö, M., Farris, J.S., Kluge, A.G., Bult, C., 1992. Skewness and permutation. *Cladistics* 8, 275–287.
- Kimura, M., 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120.
- Klenk, H.-P., Zillig, W., 1994. DNA-dependent RNA polymerase subunit B as a tool for phylogenetic reconstructions: branching topology of the archaeal domain. *J. Mol. Evol.* 38, 420–432.
- Kluge, A.G., Farris, J.S., 1969. Quantitative phyletics and the evolution of Anurans. *Syst. Zool.* 18, 1–32.
- Lanyon, S.M., 1988. The stochastic mode of molecular evolution: what consequences for systematic investigations? *Auk* 105, 563–573.
- Lewis, L.A., Mishler, B.D., Vilgalys, R., 1997. Phylogenetic relationships of the liverworts (Hepaticaceae), a basal embryophyte lineage, inferred from nucleotide sequence data of the chloroplast gene *rbcl*. *Mol. Phylogenet. Evol.* 7, 377–393.
- Lockhart, P.J., Howe, C.J., Bryant, D.A., Beanland, T.J., Larkum, A.W.D., 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. *J. Mol. Evol.* 34, 153–162.
- Manhart, J.R., 1994. Phylogenetic analysis of green plant *rbcl* sequences. *Mol. Phylogenet. Evol.* 3, 114–127.
- Meyer, A., 1994. Shortcomings of the cytochrome b gene as a molecular marker. *Trends Ecol. Evol.* 9, 278–280.
- Meyer, A., Wilson, A.C., 1990. Origin of tetrapods inferred from their mitochondrial DNA affiliation to lungfish. *J. Mol. Evol.* 31, 359–364.
- Mishler, B.D., Bremer, K., Humphries, C.J., Churchill, S.P., 1988. The use of nucleic acid sequence data in phylogenetic reconstruction. *Taxon* 37, 391–395.
- Miyamoto, M.M., Fitch, W.M., 1995. Testing the covarion hypothesis of molecular evolution. *Mol. Biol. Evol.* 12, 503–513.
- Moriyama, E.N., Powell, J.R., 1997. Synonymous substitution rates in *Drosophila*: mitochondrial versus nuclear genes. *J. Mol. Evol.* 45, 378–391.
- Naylor, G.J.P., Collins, T.M., Brown, W.M., 1995. Hydrophobicity and phylogeny. *Nature* 373, 565–566.
- Naylor, G.J.P., Gerstein, M., 2000. Measuring shifts in function and evolutionary opportunity using variability profiles: a case study of the globins. *J. Mol. Evol.* 51, 223–233.
- Nishiyama, T., Kato, M., 1999. Molecular phylogenetic analysis among bryophytes and tracheophytes based on combined data of plastid coded genes and the 18S rRNA gene. *Mol. Biol. Evol.* 16, 1027–1036.
- Ortí, G., Meyer, A., 1996. Molecular evidence of ependymin and the phylogenetic resolution of early divergences among euteleost fishes. *Mol. Biol. Evol.* 13, 556–573.
- Perna, N.T., Kocher, T.D., 1995. Unequal base frequencies and the estimation of substitution bias. *Mol. Biol. Evol.* 12, 359–361.
- Prager, E.M., Wilson, A.C., 1988. Ancient origin of lactalbumin from lysozyme: analysis of DNA and amino acid sequences. *J. Mol. Evol.* 27, 326–335.
- Prychitko, T.M., Moore, W.S., 2000. Comparative evolution of the mitochondrial cytochrome b gene and nuclear B-fibrinogen intron 7 in woodpeckers. *Mol. Biol. Evol.* 17, 1101–1111.
- Schuh, R.T., Polhemus, J.T., 1980. Analysis of taxonomic congruence among morphological, ecological, and biogeographic data sets for the Leptopodomorpha (Hemiptera). *Syst. Zool.* 29, 1–26.
- Sennblad, B., Bremer, B., 2000. Is there a justification for differential a priori weighting in coding sequences? A case study from *rbcl* and Apocynaceae s.l. *Syst. Biol.* 49, 101–113.
- Simmons, M.P., 2000. A fundamental problem with amino-acid-sequence characters for phylogenetic analyses. *Cladistics* 16, 274–282.

- Simmons, M.P., In press. Amino acid versus nucleotide characters for phylogenetic inference of the 'basal' angiosperms. In: Sharma, A.K., Sharma, A. (Eds.), *Plant Genome Biodiversity and Evolution: Vol. 1, Part B: Phanerogams*. Science Publishers, Enfield, NH.
- Simmons, M.P., Carr, T.G., O'Neill, K., In press. Relative character-state space, amount of potential phylogenetic information, and heterogeneity of nucleotide and amino acid characters. *Mol. Phylogenet. Evol.*
- Simmons, M.P., Freudenstein, J.V., 2002. Artifacts of coding amino acids and other composite characters for phylogenetic analysis. *Cladistics* 18, 354–365.
- Simmons, M.P., Miya, M., in press. Efficiently resolving the basal clades of a phylogenetic tree using Bayesian and parsimony approaches: a case study using mitogenomic data from 100 higher teleost fishes. *Mol. Phylogenet. Evol.* 00, 000–000.
- Simmons, M.P., Ochoterena, H., Freudenstein, J.V., 2002. Amino acid vs. nucleotide characters: challenging preconceived notions. *Mol. Phylogenet. Evol.* 24, 78–90.
- Sokal, R.R., Rohlf, F.J., 1981. Taxonomic congruence in the *Leptodomorpha* re-examined. *Syst. Zool.* 30, 309–325.
- Steel, M., Penny, D., 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol. Biol. Evol.* 17, 839–850.
- Swofford, D.L., 2002. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods) 4.0 beta 10. Sinauer, Sunderland, MA.
- Wenzel, J.W., Siddall, M.E., 1999. Noise. *Cladistics* 15, 51–64.
- Wirth, T., Le Guellec, R., Veuille, M., 1999. Directional substitution and evolution of nucleotide content in the *cytochrome oxidase II* gene in earwigs (dermapteran insects). *Mol. Biol. Evol.* 16, 1645–1653.
- Yang, Z., 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* 10, 1396–1401.
- Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* 13, 555–556.
- Yang, Z., 1998. On the best evolutionary rate for phylogenetic analysis. *Syst. Biol.* 47, 125–133.
- Yang, Z., Nielsen, R., Hasegawa, M., 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.* 15, 1600–1611.
- Zujko-Miller, C., Miller, J.A., 2003. PEST: Precision estimated by sampling traits. Program distributed by the authors, <<http://www.gwu.edu/~clade/spiders/pestDocs.htm>>.