Points of View

Syst. Biol. 54(3):493–500, 2005

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150590947339

Choosing the Best Genes for the Job: The Case for Stationary Genes in Genome-Scale Phylogenetics

TIMOTHY M. COLLINS,[1] OLIVIER FEDRIGO,[2] AND GAVIN J. P. NAYLOR[3]

[1]*Department of Biological Sciences, Florida International University, University Park, Miami, Florida 33199, USA; E-mail: collinst@fiu.edu*
[2]*Department of Biology, Duke University, Durham, North Carolina 27708, USA*
[3]*School of Computational Science and Information Technology 150-A Dirac Science Library, Florida State University Tallahassee, Florida 32306-4120, USA*

The advent of genomics has fueled optimism for improvement in the reliability and accuracy of phylogenetic trees. An implicit assumption is that there will be an inexorable improvement in phylogenetic accuracy as the number of genes used increases, and that this approach is necessary because there are no identifiable parameters that predict the phylogenetic performance of genes (Gee, 2003; Rokas et al. 2003). These issues were explored in the recent article by Rokas et al. who investigated the phylogenetic signal in a sample of 106 protein-encoding genes selected from the genomes of 8 species of yeast.

Rokas et al. (2003) analyzed these genes separately, and in combination, showing that individual genes sometimes support conflicting topologies. Although considerable character incongruence existed in the combined data set, simultaneous analysis of all genes resulted in one tree with 100% bootstrap proportions (BP) at all nodes. This "species tree" was taken to represent the true phylogeny (Fig. 1a topology). The authors then carried out a series of analyses with randomly concatenated data sets of varying size to determine the minimum amount of data required to establish confidence in the species tree at a given level of statistical significance. They concluded that a minimum of 20 randomly concatenated genes was required to infer relationships confidently and that "It is only through the analyses of larger amounts of sequence data that confidence in the proposed phylogenetic reconstruction can be obtained" and further "that analyses based on a single or a small number of genes provide insufficient evidence for establishing or refuting phylogenetic relationships." They also expressed the opinion that the result for these yeast species was likely to be typical for molecular phylogenetic studies: "... we believe that this group is a representative model for key issues that researchers in phylogenetics are confronting," with the clear implication that the majority of current molecular phylogenies must be considered unreliable.

Another important conclusion was that there are no predictors of phylogenetic performance of genes: "there were no identifiable parameters that could systematically account for or predict the performance of single genes." Similarly, Gee (2003), in discussing the Rokas et al. (2003) paper states, "there are no identifiable parameters that can predict the performance of genes in any systematic way." Finally, they noted that bootstrap values were lower and variance higher for contiguous gene sequences than for randomly sampled orthologous nucleotides and took this as evidence of the misleading signal in individual genes resulting from the nonindependence of nucleotides within genes.

These conclusions, if true, are sobering for those attempting to infer relationships using DNA sequences with limited time and budgets. Herein, we demonstrate that these conclusions require substantial revision. First we show that many genes in the yeast data set published by Rokas et al. (2003) have nucleotide frequencies that have shifted markedly among taxa at third positions of codons. These nucleotide sequences deviate significantly from the stationary condition (see also Phillips et al., 2004). Second, we illustrate through a series of analyses that the stationary gene partition is superior to the nonstationary partition, recovering the underlying phylogeny with many fewer genes. Finally, we show that the conclusion of Rokas et al. regarding the superiority of random sampling of orthologous nucleotides relative to contiguous sequences for phylogenetic analysis is largely an artifact of different bootstrap treatments for these two sampling schemes.

Rokas et al. (2003) used several criteria for sampling and retaining genes from seven species of *Saccharomyces* yeasts, and one outgroup species, *Candida albicans* (Fig. 1). Genes were spaced at approximately 40-kilobase intervals. Only protein-encoding genes with identifiable and generally alignable homologs in all eight

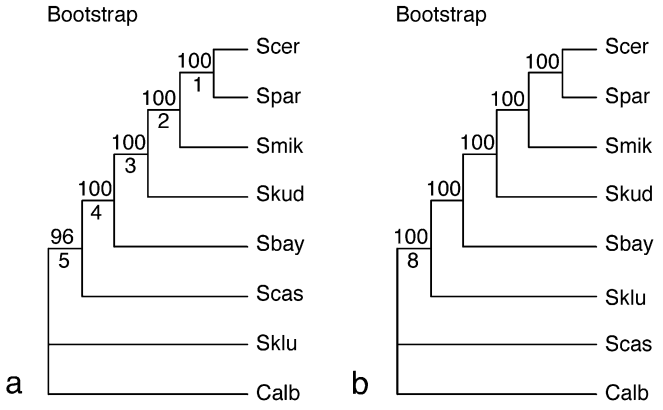

FIGURE 1. Results of branch-and-bound maximum parsimony bootstrap analysis (300 replicates) of stationary (a) and nonstationary (b) third-position partitions. The topology and branch numbers of the tree in (a) are the same as the species tree of Rokas et al. (2003). Bootstrap proportions are shown above the branch, branch numbers below. Calb = *Candida albicans*; Sklu = *Saccharomyces kluyveri*; Scas = *S. castellii*; Sbay = *S. bayanus*; Skud = *S. kudriavzevii*; Smik = *S. mikatae*; Spar = *S. paradoxus*; Scer = *S. cerevisiae*.

species were used. In addition, Rokas et al. restricted their selection to genes that had at least two homologous flanking syntenic genes. Regions of proteins that were considered difficult to align were discarded, with an average of 76% of each gene being retained. Trimmed gene size ranged from 390 to 2994 nucleotides, with an average of 1198 nucleotides. We carried out a series of further analyses with a copy of the data set kindly provided by the authors. Analyses of first and second codon positions using equally weighted parsimony yielded the species tree with 100% support at all nodes. Third positions, however, which represent the majority ($\approx$ 70%) of the parsimony-informative characters, yielded a conflicting tree, with 84% support for a basal position of *S. castellii* rather than *S. kluyveri* within the ingroup (branch 8). What is the source of this conflicting signal at third positions?

Most currently used phylogenetic methods assume that the frequencies of nucleotides do not change significantly among taxa. Deviations from this stationary condition at any codon position can result in systematic error (Saccone et al., 1989). We divided the data set into two halves: one half composed of genes that were not stationary at third positions in the ingroup at $P = 0.01$ (52 genes) and those that were stationary (54), based on the chi-square test in PAUP*. In subsequent analyses we refer to those genes that were not stationary at third positions in the ingroup at $P = 0.01$ as the nonstationary partition, and the remaining genes as the stationary partition. We performed parsimony bootstrap analysis of individual genes across all positions to compare the phylogenetic performance of these partitions. We considered a gene's performance as good if it yielded a tree identical to, or fully compatible with, the species tree. By fully compatible, we mean that a semistrict or combinable component consensus of the gene tree and the species tree is identical to the species tree. Note that this differs from the method employed by Rokas et al. (2003). Their approach to measuring topological incongruence among single-gene topologies involved trimming branches from 50% bootstrap trees until the topologies were identical (e.g., Fig. 3 of Rokas et al., 2003). Their approach implicitly treats all polytomies as hard, conflating conflict and lack of resolution (see also Taylor and Piel, 2004). Given that the fully resolved species tree is taken to be the true topology, and that one third of the genes analyzed are less than 900 nucleotides in length, we argue that polytomies should be treated as soft. Therefore, we measured only actual rather than potential topological conflict. When analyzed this way, 61% of the stationary genes were found to yield trees that were identical to or fully compatible with the species tree. By contrast, only 38% of the nonstationary genes yielded identical or fully compatible topologies. When third positions alone were analyzed, the stationary half yielded the species tree with 96% to 100% support at all nodes (Fig. 1a), whereas the nonstationary half of the data set yielded the incorrect *S. castellii*–basal tree with 100% support at all nodes (Fig. 1b). The incongruence between stationary and nonstationary third-position partitions was significant ($P = 0.002$, ILD test in PAUP*, 500 replicates).

We investigated incongruence between stationary and nonstationary partitions further by examining partitioned Bremer support. Partitioned Bremer support quantifies the partition-specific support for a given branch by counting the difference between the length of a data partition on the most parsimonious tree for the complete data set, and the length on the shortest tree that does not contain that branch (Baker and DeSalle, 1997). In this case, strongly positive values indicate strong support for the branch by a partition on the species tree. Partitioned Bremer support for the *S. kluyveri*–basal branch 5 in the species tree is positive for first positions (+418), second positions (+452), and stationary third positions (+82), but negative for all third positions (−72), and especially nonstationary third positions (−154) (Table 1). Concatenated gene analyses demonstrate that the nonstationary third-position partition is inconsistent under parsimony; as genes are added, the incorrect *S. castellii*–basal branch 8 is recovered with confidence increasing to 100% (Fig. 2a). A cluster analysis performed with SYSTAT indicates the source of the misleading signal for this branch: *C. albicans* and *S. castellii* have convergent extreme A+T-rich base compositions at nonstationary

TABLE 1. Partitioned Bremer support by codon position

Partition/ codon position	Length partition on species tree	Partitioned Bremer		
		Branch 2	Branch 3	Branch 5
Stationary				
1	8584	+86	+73	+192
2	3581	+55	+32	+262
3	33552	+220	+162	+82
Nonstationary				
1	8181	+27	+10	+226
2	3196	+25	+22	+189
3	38909	+96	+49	**−154**

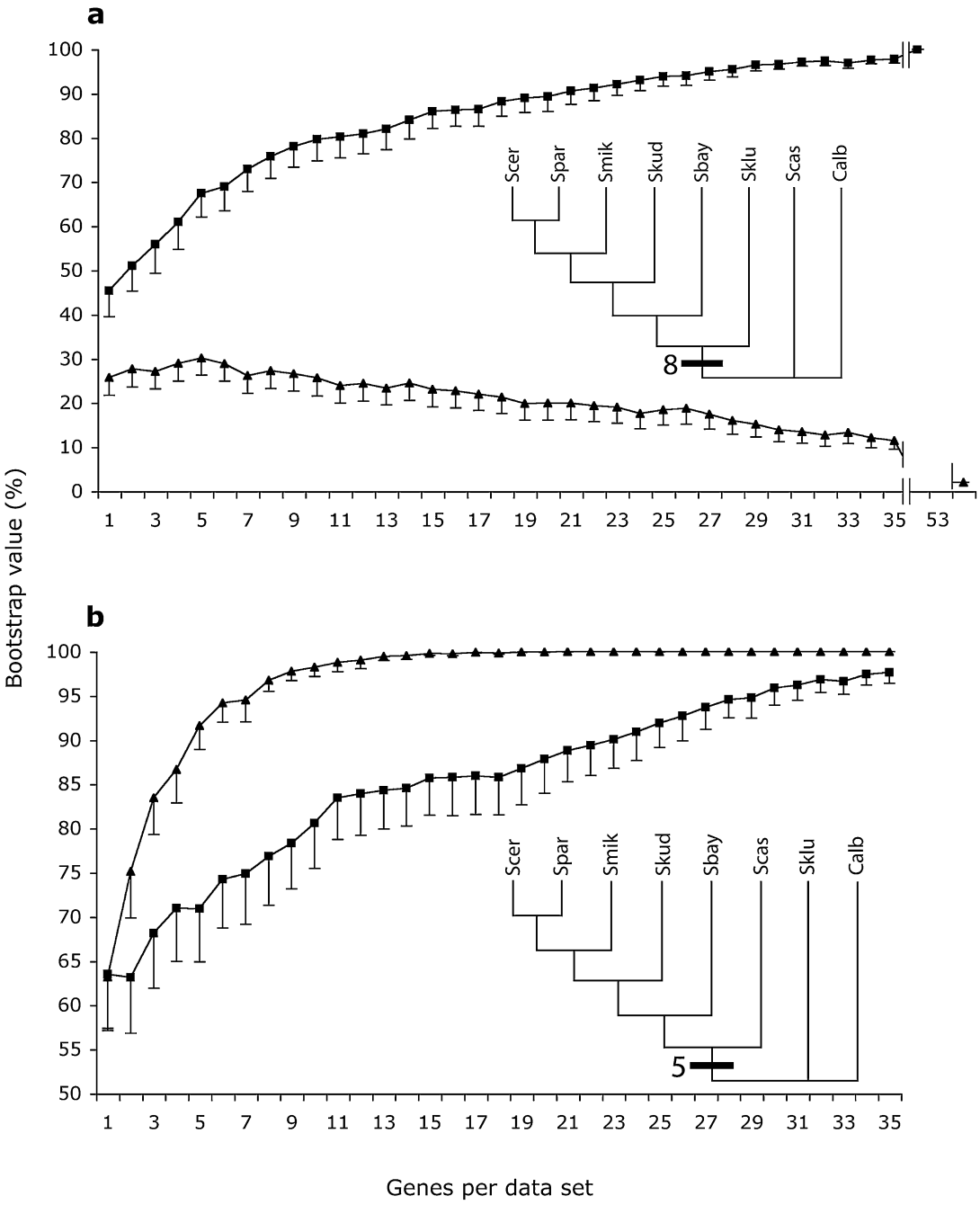

FIGURE 2. Stationary partition genes (triangles) are less prone to misleading phylogenetic signal than nonstationary partition genes (squares) using parsimony. Bootstrap proportions for the branches uniting the outgroup *C. albicans* to either *S. castellii* or *S. kluyveri* were determined by the analysis of 100 randomly concatenated subsets of genes for each data set size. These branches are equivalent to branches 8 and 5 of Rokas et al. Each data point represents the mean and minus one 95% confidence interval of 100 branch and bound parsimony bootstrap analyses for that data set size. (a) The third-codon position nonstationary data are inconsistent. As more genes are concatenated, the incorrect *S. castellii*–basal branch 8 is recovered with increasing statistical support. (b) Stationary partition genes using all codon positions recover the correct *S. kluyveri*–basal branch 5 with fewer genes for a given level of support. Compare to figure 5b of Rokas et al. (2003).

third positions (Fig 3a). *C. albicans* and *S. castellii* average 70.4% A+T at nonstationary third positions, while the remaining taxa average 57.4% A+T. This base compositional bias is more extreme than at all codon positions (Fig. 3b, note change in scale). *C. albicans* and *S. castellii* do not form a cluster when considering base composition at second positions (Fig. 3c), or first positions (data not shown).

A concatenated gene analysis across all codon positions demonstrates that the stationary partition recovers the *S. kluyveri*–basal branch 5 with significantly fewer sampled genes (Fig. 2b). Three genes from the stationary

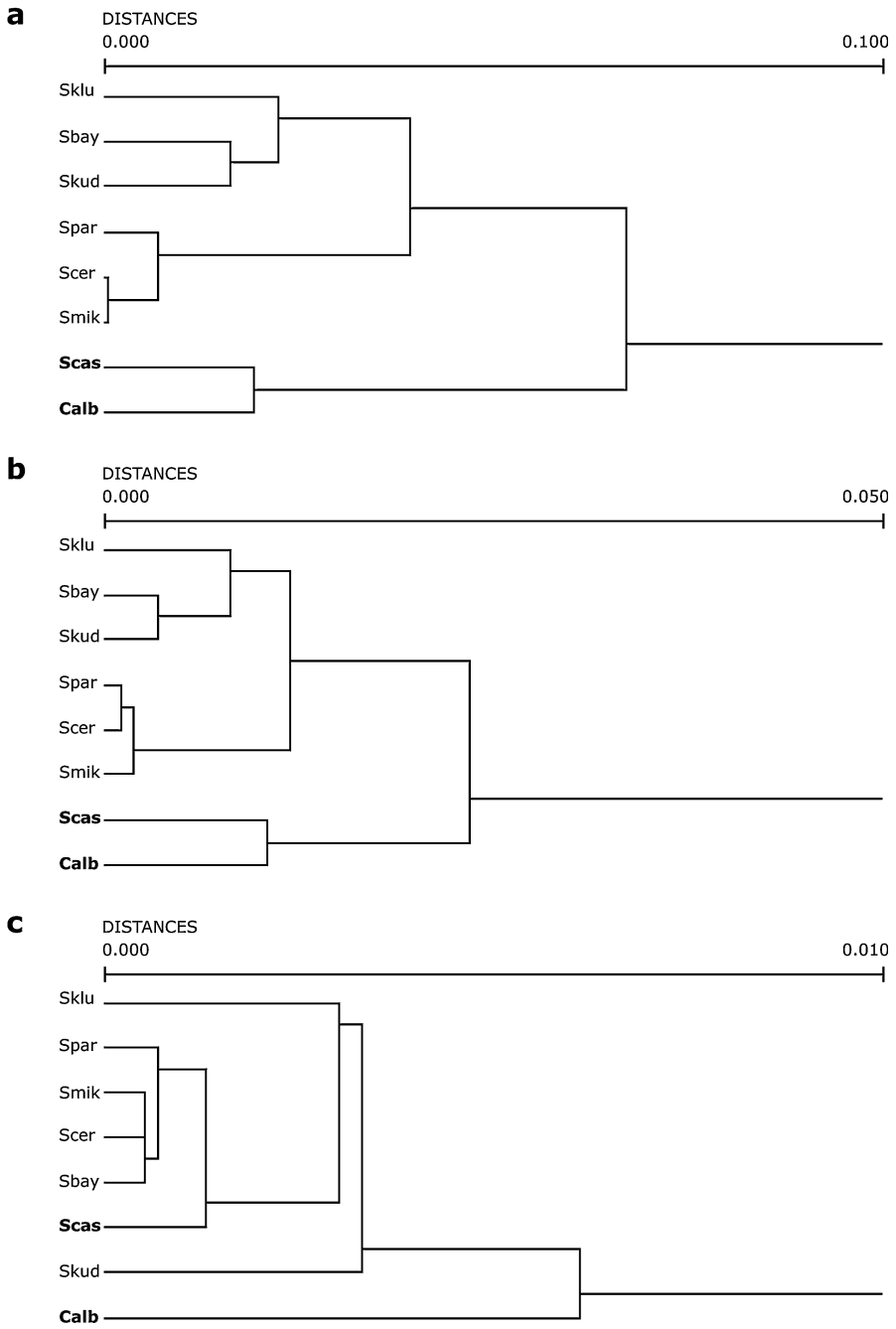

FIGURE 3. Euclidean distance average-linkage cluster analysis by nucleotide frequencies demonstrates the compositional attraction of *C. albicans* and *S. castellii* for select data partitions. (a) The third-position nonstationary partition. (b) All codon positions. (c) Second codon positions. Note differing scales.

partition were required to recover the *S. kluyveri* basal branch 5 with >70% BP and a 95% confidence interval; in contrast 8 genes from the nonstationary partition were required to meet that minimum. Similarly, 8 genes from the stationary partition versus 32 genes from the nonstationary partition were necessary for >95% BP and a 95% confidence interval. A similar pattern was found with maximum likelihood. The all-codon position stationary partition performed significantly better, requiring 8 stationary genes, but 21 nonstationary partition genes for >95% BP and a 95% confidence interval (Fig. 4b). The third-position nonstationary partition is no longer inconsistent under maximum likelihood using the HKY model, although it does not perform as well as stationary third positions (Fig. 4a).

Two other branches in the species tree, branches 2 and 3, had significant character conflict (Rokas et al., 2003). We compared the performance of the stationary and

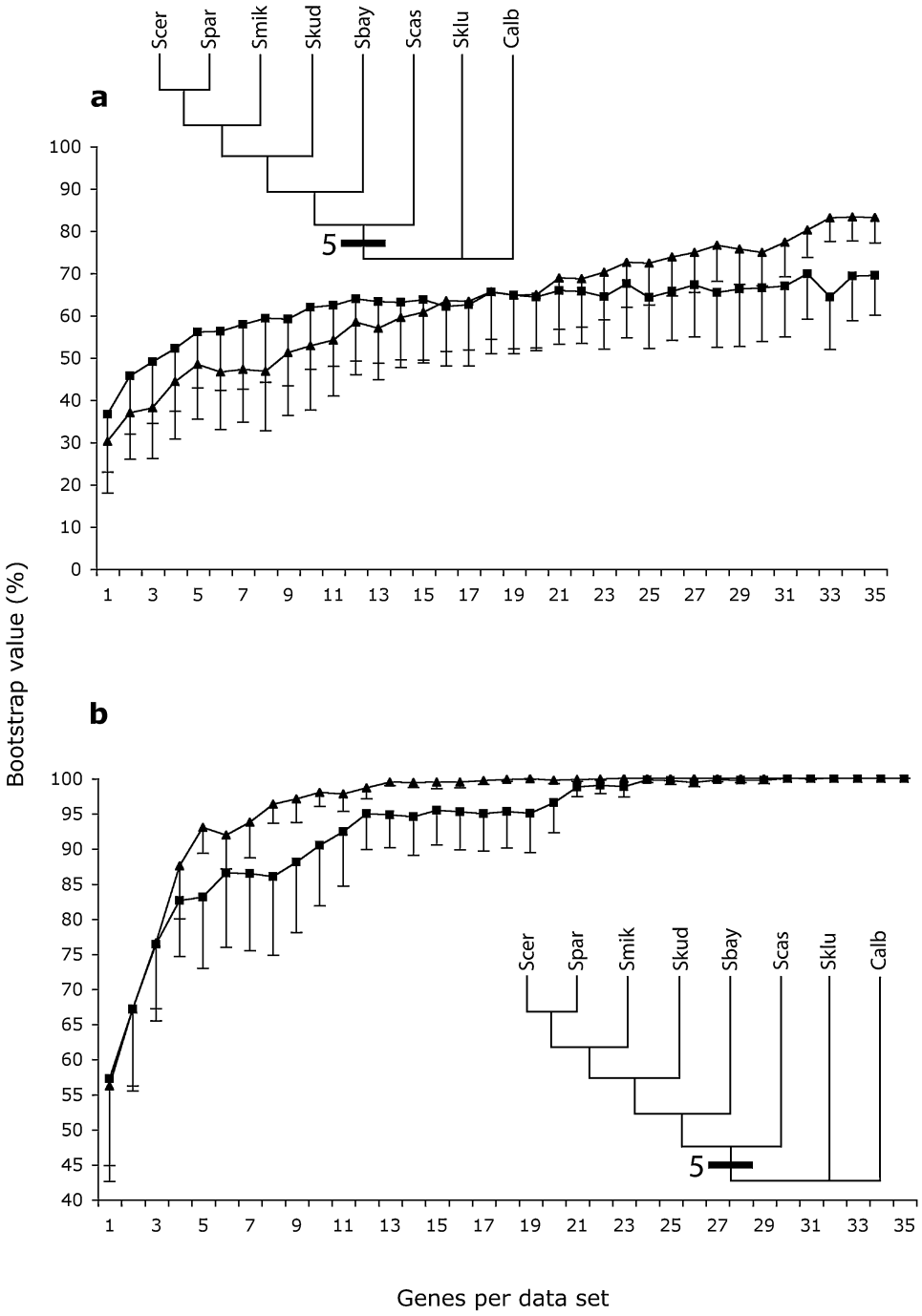

FIGURE 4. Stationary genes (triangles) are less prone to misleading phylogenetic signal than nonstationary genes (squares) using maximum likelihood. Bootstrap values for the branches uniting the outgroup *C. albicans* to *S. kluyveri* (branch 5) were determined by the analysis of 20 randomly concatenated subsets of genes for each data set size. Each data point represents the mean and minus one 95% confidence interval of 20 heuristic maximum likelihood bootstrap analyses using HKY as the model of sequence evolution. (a) The third-codon position stationary partition outperforms the nonstationary partition. (b) Stationary genes using all codon positions recover the correct *S. kluyveri*–basal branch 5 with fewer genes for a given level of support. Compare to Figure 5b of Rokas et al. (2003).

nonstationary partitions for these two branches, finding once again that the stationary partition performed significantly better for both branches (Fig. 5). Seven genes from the stationary partition were required to recover branch 2 with >95% BP and a 95% confidence interval; in contrast, 18 genes from the nonstationary partition were required to meet that minimum. For branch 3, 10, and 23 stationary and nonstationary partition genes were required for >95% BP and a 95% confidence interval. The partitioned Bremer support was lower for the nonstationary partition at all three codon positions for these branches (Table 1), consistent with the concatenated gene analysis.

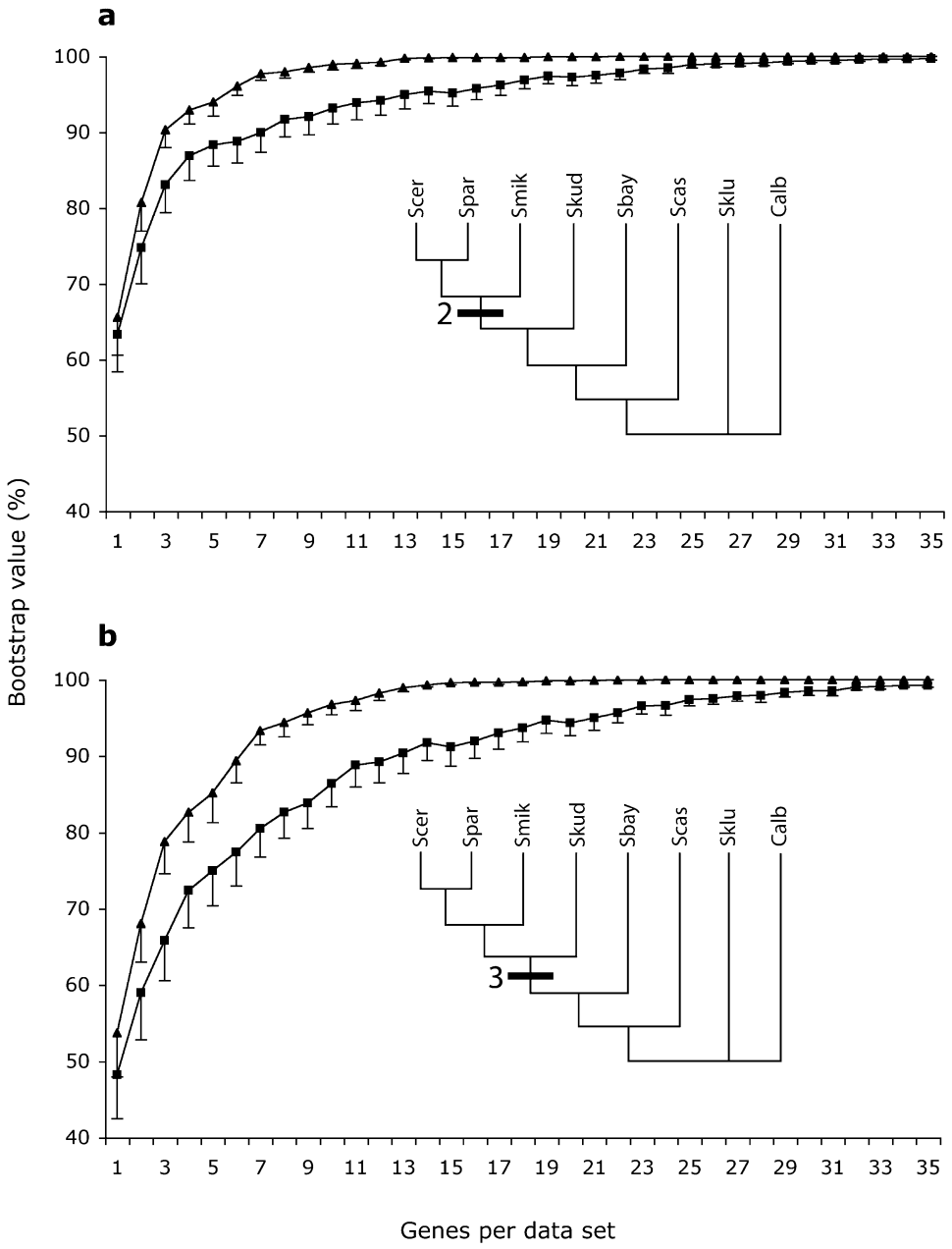

FIGURE 5. Stationary genes (triangles) are less prone to misleading phylogenetic signal than nonstationary genes (squares) using parsimony. The bootstrap values for branches 2 and 3 of the species tree by partition. Each data point represents the mean and minus one 95% confidence interval of 100 branch-and-bound parsimony bootstrap analyses. (a) Branch 2. (b) Branch 3.

A final question concerns methods of sampling nucleotide sequences for phylogenetic analysis. Rokas et al. (2003) suggested that randomly sampled orthologous nucleotides were superior to contiguous gene sequences. This superiority of randomly sampled nucleotides was thought to be due to avoidance of within-gene nonindependence of nucleotides linked in a functional gene. An examination of figure 5 from Rokas et al. suggests that randomly sampled nucleotides had considerably higher bootstrap values for similar levels of sampling, and remarkably low variance in bootstrap values. In fact, the confidence intervals are not visible in this figure, overlapping almost completely with the plotted average data points. This result is, however, largely an artifact of differing bootstrap resampling techniques applied to the randomly sampled and contiguous gene sequences. A typical nonparametric bootstrap was applied to the contiguous gene sequences: a sample (an individual gene) was taken and pseudosamples of the same size were generated from this sample by sampling with replacement. This gives the variance about the estimate of the phylogeny for that sample. Randomly sampled orthologous

nucleotides were sampled with a different strategy, using the variable-length-bootstrap option in PAUP*. In this case, a sample of a given size was taken from the complete data set and the phylogeny was estimated. Then a new sample of a given size was taken, and the phylogeny estimated. So, for even the smallest sample size of 1000 nucleotides, 1000 replicates would have sampled the vast majority of nucleotides from the complete data set. This sampling scheme did not, therefore, measure the variance on the estimate of the phylogeny from a particular random sample, as in the contiguous gene sequences, but is instead akin to the variance on the phylogeny for repeated sampling of a given size from the complete data set. This difference in treatment explains the extremely small confidence intervals for the randomly sampled nucleotides in the Rokas et al. analysis. When the randomly sampled orthologous nucleotides are bootstrapped in the same manner as the contiguous gene sequences, much greater variances are apparent (Fig. 6, open circles), although these variances are not directly comparable to the contiguous gene bootstrap variances because they do not include a variance component related to variation in gene size. We also find that orthologous nucleotides randomly sampled from the stationary partition are superior to those sampled from the nonstationary partition (Fig. 6, triangles and squares). It is certainly a reasonable expectation that nucleotides sampled from within a gene would have greater nonindependence than randomly sampled nucleotides, but the phylogenetic performance difference, if present, appears to be much more subtle than would be inferred from examination of figure 5 of Rokas et al.

Our reanalysis of Rokas et al.'s data indicates that their estimate of the number of genes required to infer a phylogeny confidently was inflated by signal heterogeneity caused by their inclusion of nonstationary genes. In addition, the conclusion that there are no useful predictors of phylogenetic performance does not hold. We have shown that the addition of a single criterion, base compositional stationarity of individual genes at third codon positions, to the selection criteria of Rokas et al. significantly improved performance with this data set. Genes from the stationary partition were superior, recovering the underlying phylogeny with a third of the number of genes from the nonstationary partition for a given level of significance. Twenty genes were not required for this data set, and, if one analyzed genes from the stationary partition, the improvement beyond about eight genes was very modest. It is important to note that we are not suggesting that genes come in two discrete classes, stationary and nonstationary, that can be unequivocally distinguished. The chi-square test in PAUP* is rough and approximate; it does not take phylogeny or the proportion of invariant sites into account (Foster, 2004). In addition, larger, more variable genes are more likely to fail the chi-square test of stationarity. The size effect would be a bias against our hypothesis, because one would assume that, all other things being equal, big genes would perform better than small genes, and the bias would be for including stationary genes in the nonstationary partition because of size. We should not overemphasize this effect. For example, all 42,342 second positions derived from the concatenated data set of 106 genes for the ingroup taxa are stationary ($P = 0.989$). Similarly, no individual genes

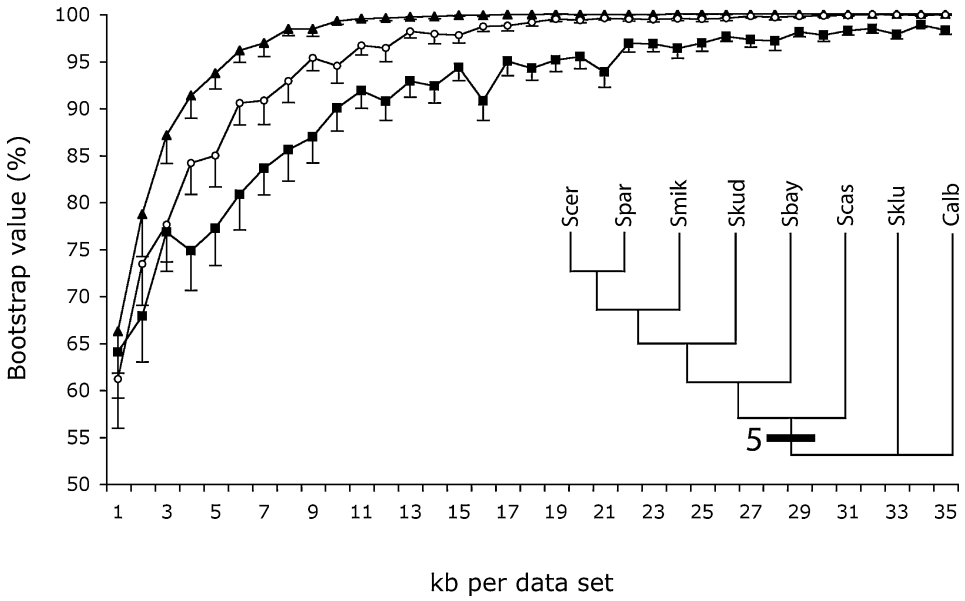

FIGURE 6. Randomly sampled orthologous nucleotides have significant variance in bootstrap values using a standard nonparametric bootstrap. Each data point represents the mean and minus one 95% confidence interval of 100 branch-and-bound parsimony bootstrap replicates. Randomly sampled orthologous nucleotides from the complete data set (circles) have significant variance. Randomly sampled nucleotides from the stationary partition (triangles) recover the correct *S. kluyveri*–basal branch 5 with fewer nucleotides for a given level of support than randomly sampled nucleotides from the nonstationary partition (squares).

were nonstationary at first positions for the ingroup at $P = 0.01$ using the chi-square test in PAUP*. The chi-square test in PAUP* appears to do a good job of identifying the worst offenders and although crude, is effective for a large data set when analyzing individual genes. We certainly anticipate that as more refined tests of stationarity are implemented (e.g., Foster, 2004), discrimination of genes that deviate from the stationary condition will improve. We note finally that although deviations from stationarity were detected at third positions, we are not suggesting that all misleading base compositional signal is at third positions. Since these nucleotides are linked as codons, we might expect that strong deviations at third positions would influence first and second positions of codons.

Broadly speaking, accurate phylogenetic trees can be recovered from correctly aligned sequences when the inference model is consistent with the process that gave rise to the data. When processes are stationary over lineages and time, relatively straightforward models can be designed to yield accurate inferences, even from short sequences (Steel and Penny, 2000). When processes differ across or within lineages, models must explicitly accommodate the nonstationarity involved. This is generally not straightforward, and even if it could be done, would require many more parameters and associated error terms (but see Foster, 2004). As such, at a given data set size, stationary sequences will prove to be more effective for recovering phylogeny. Stationary sequences will be less prone to the grouping of taxa with convergent base compositions. Of course, when taxa share an atypical base composition in a gene sequence because of shared history, nonstationary sequences may outperform stationary sequences in recovering that branch when using models that assume stationarity. Such instances are cases of obtaining the right answer for the wrong reason (e.g. Swofford et al. 2001) and are a poor argument for use. The criterion of stationarity should prove useful in selecting genes for phylogenetic analysis from completely sequenced genomes, and to the extent that genes that tend to remain stationary can be identified, will be useful for de novo sequencing studies. In general, avoidance of genes with strong deviations from base compositional equilibrium should prove to be a useful strategy for efficient recovery of accurate phylogenetic estimates with markedly fewer genes.

ACKNOWLEDGEMENTS

We thank Thomas Buckley, David Kizirian, Matt Osentoski, Rod Page, Matt Phillips, Tim Rawlings, Barry Williams, and an anonymous reviewer for helpful comments on the manuscript, and Dave Swofford and Mark Holder for discussions about the chi-square test in PAUP*. GN acknowledges the support of NSF grant DEB-0415486.

REFERENCES

Baker, R. H., and R. DeSalle. 1997. Multiple sources of character information and the phylogeny of Hawaiian drosophilids. Syst. Biol. 46:654–673.

Foster, P. G. 2004. Modeling compositional heterogeneity. Syst. Biol. 53: 485–495.

Gee, H. 2003. Ending incongruence. Nature 425:782.

Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21:1455–1458.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Saccone, C., G. Pesole, and G. Preparata. 1989. DNA microenvironments and the molecular clock. J. Mol. Evol. 29:407–411.

Steel, M., and D. Penny. 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), version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.

Swofford, D. L., P. J. Waddell, J. P. Huelsenbeck, P. G. Foster, P. O. Lewis, and J. S. Rogers. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50:525–539.

Taylor, D. J., and Piel, W. H. 2004. An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. Mol. Biol. Evol. 21:1534–1537.

First submitted 10 June 2004; reviews returned 31 August 2004; final acceptance 8 December 2004

Associate Editor: Thomas Buckley

Syst. Biol. 54(3):500–507, 2005

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150590947311

Estimating Divergence Times in Phylogenetic Trees Without a Molecular Clock

TOM BRITTON

Department of Mathematics, Stockholm University, SE-106 91 Stockholm, Sweden; E-mail: tom.britton@math.su.se

Rates of evolution often tend to vary between lineages in a phylogenetic tree, implying that the molecular clock assumption is not valid. In this article, we are therefore concerned with estimation of divergence times without assuming a constant molecular clock, where inference is based on DNA (or amino acid or protein) sequences from the species of interest. "Time" could here either be relative time, i.e., all divergence times are relative to the unknown age of the root of the tree, or absolute time if some fossil dating(s) relating the relative times to absolute time are available. Here we focus on relative times, but in either case such a tree is ultrametric and will be denoted the time-tree.