LETTERS
Genome-Scale Phylogeny and the Detection of Systematic Biases

Matthew J. Phillips,[1] *Frédéric Delsuc,*[2] *and David Penny*

The Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand

Phylogenetic inference from sequences can be misled by both sampling (stochastic) error and systematic error (nonhistorical signals where reality differs from our simplified models). A recent study of eight yeast species using 106 concatenated genes from complete genomes showed that even small internal edges of a tree received 100% bootstrap support. This effective negation of stochastic error from large data sets is important, but longer sequences exacerbate the potential for biases (systematic error) to be positively misleading. Indeed, when we analyzed the same data set using minimum evolution optimality criteria, an alternative tree received 100% bootstrap support. We identified a compositional bias as responsible for this inconsistency and showed that it is reduced effectively by coding the nucleotides as purines and pyrimidines (RY-coding), reinforcing the original tree. Thus, a comprehensive exploration of potential systematic biases is still required, even though genome-scale data sets greatly reduce sampling error.

Rokas et al. (2003) use eight yeast genomes to derive a data set of 106 nuclear genes (127,026 nucleotides). This data set gave a phylogeny of seven *Saccharomyces* species, rooted by *Candida albicans*, where all internal branches receive 100% bootstrap support under both maximum parsimony (MP) and maximum likelihood (ML). It is expected theoretically and empirically (from simulations) that with very long sequences sampling error should vanish, with bootstrap values going to 100%. However, the presence of other (nonhistorical) signals in the data that are not indicative of ancestry has long been recognized (Penny, Hendy, and Steel 1992; Hillis 1995; Lopez, Casane, and Philippe 2002). Bootstrap support values (Felsenstein 1985), convergence tests (Penny and Hendy 1986), and the Templeton (1983) and Shimodaira and Hasegawa (1999) tests assess sampling effects only but cannot indicate whether the trees are actually correct.

In the present example, the Rokas et al. (2003) tree appears correct, but it is an excellent data set for detecting whether there are nonhistorical signals (systematic biases) in the data and, if so, their potential influence. We used PAUP* version 4.0b8 (Swofford 2002) for minimum evolution (ME) general time-reversible (GTR [Yang 1994]) and LogDet (Lockhart et al. 1994) distances from the data, coded both as standard nucleotides (NT-coding) and as purines (A&G→R) and pyrimidines (C&T→Y). The latter regime (RY-coding) eliminates transition biases, a goal that may be traced back to methods such as transversion parsimony (e.g., Woese et al. [1991]). Strengths of conflicting signals were inferred using SpectroNet version 1.2 (Huber et al. 2002). The results in figure 1 show that although ML and MP support the Rokas tree with 100% bootstrap support, ME on both GTR and LogDet distances retaining all constant sites gives a different tree, also with full 100% bootstrap support.

Figures 1*A* and *B* cannot both be correct. With stochastic effects eliminated, if the strongest signal is not the historical signal, then the tree that emerges will be incorrect because the method is inconsistent under those conditions. Under the nomenclature of Rokas et al. (2003), the trees differ by selecting branch 3 (fig. 1*A*) or branch 6 (fig. 1*B*). These are two of the three local rearrangements at the node circled in figure 1. The remaining possibility (which we call X) is for *S. bayanus* and *S. kudriavzevii* to swap positions in figure 1*A*. We tested for potential effects from model misspecification by conducting ML searches using PAUP* with TBR branch swapping on Neighbor-Joining starting trees. The figure 1*A* tree was found for the 56 models tested in ModelTest version 3.06 (Posada and Crandall 1998), using nucleotide data and the parameters estimated by the program. Thus, the tree appeared relatively robust to model assumptions, but this does not account for all the signals in the data.

To help estimate the signals, we used the distance Hadamard transform (Hendy and Charleston 1993) in SpectroNet for comparing the three local rearrangements at the node circled in figure 1. If the model fits the data accurately, then we expect the ideal situation with one positive (phylogenetic) signal and the other two (nonhistorical) to be zero. Even if other processes such as positive selection had occurred frequently and randomly among lineages, then we expect the effects to be evenly distributed among the two nonhistorical signals. Unfortunately we do not find the ideal situation where the model fits the data accurately, but the latter (semi-ideal) situation occurs for the RY-coded data (fig. 2*B* [see below]). In contrast, two strong competing signals, and a third lesser one, are indicated for the NT-coded data (fig. 2*A*). Thus, there are three major signals in the data; they cannot all be phylogenetic (historical). Indeed, in figure 2*A* the sum of the two smaller signals exceeds the largest.

Our experience with mitochondrial genomes (Phillips and Penny 2003; Delsuc, Phillips, and Penny 2003) is that RY-coding increases historical signal relative to compositional bias. In the present case, it gives both a 75% reduction in relative composition variability (RCV [Phillips and Penny 2003]) and a marked increase in the signal on internal branches, as measured by the treeness statistic

[1] Present address: Ancient Biomolecules Centre, Zoology Department, University of Oxford, UK.

[2] Present address: Institut des Sciences de l'Evolution de Montpellier, Université Montpellier II, France.

Key words: genome-scale phylogeny, systematic error, *Saccharomyces*, RY-coding.

E-mail: matthew.phillips@zoo.ox.ac.uk.

Mol. Biol. Evol. 21(7):1455–1458. 2004
doi:10.1093/molbev/msh137
Advance Access publication April 14, 2004

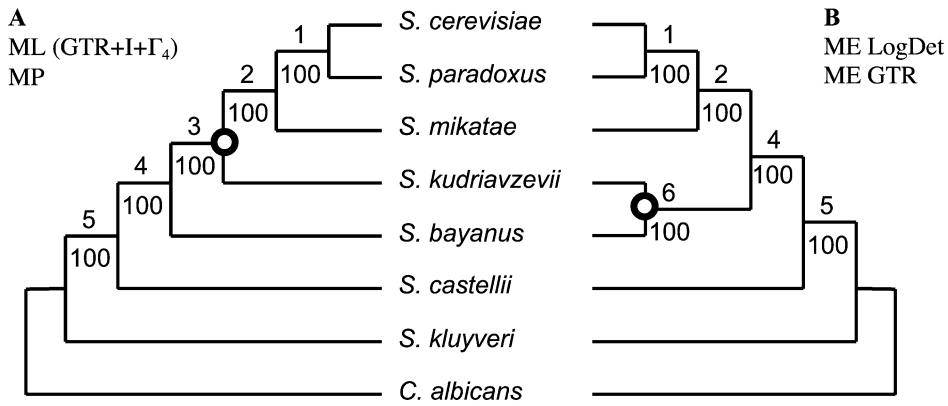

FIG. 1.—Contradictory trees, each with 100% bootstrap support. (*A*) MP and ML (GTR+I+Γ_4) and (*B*) ME with GTR and LogDet distances.

of Phillips et al. (2001) (see Supplementary Information online). In addition, under RY-coding, all four methods give the tree in figure 1*A*. Clearly, this agreement that emerges from the RY-coding is a desirable property but does not identify the competing signals in the data. So far, we can just hope the largest signal is phylogenetic. In the following section, we explore the hypothesis that an AT-GC bias is responsible for the apparent excess of signal for branch 6 with NT-coding (fig. 2*A*).

Initially, all constant sites were included for the ME analyses (fig. 1*B*). Theory predicts (Steel, Huson, and Lockhart 2000) that if there is a compositional bias and with a reduced effect from invariable sites, then the LogDet model is more likely to flip over to supporting branch 3 than is the symmetric base-frequency GTR model. Indeed, as the proportion of invariant sites deleted approaches 75%, support for branch 6 flips to support for branch 3 with LogDet, whereas branch 6 is retained under the GTR model. These results illustrate the importance of examining for nonhistorical signals.

To focus on the potential for composition bias to mislead phylogeny reconstruction, ME trees were constructed based solely on base frequencies. These new "base-frequency" distances were calculated as follows, for example the pairwise GC frequency distance between *S. mikatae* and *S. castellii* is the absolute value $|GC_{S.\ mikatea} - GC_{S.\ castellii}|$, where GC is the number of guanines plus cytosines (G+C) in the sequences. Three GC frequency distance trees with the alternative branches 3, 6, and X were constructed.

The optimal ME tree from these GC frequency distances is the same as the tree in figure 1*B*, which favors branch 6 (see Supplementary Material online). Thus, the base composition signal by itself is sufficient to give this tree. The alternatives with branches 3 and X both require an additional 688 GC/AT changes (table 1). This difference (688) based on GC bias is four and a half times greater than the 152 changes based on standard distances for the NT-coded data (see table 1). Thus, the GC bias is a candidate for explaining much of the nonhistorical signal (systematic bias) for branch 6 under NT-coding (shown in figure 2*A*).

Similarly, purine frequency distances also support branch 6 over branches 3 and X, although only marginally (by 10.9 and 13.5 changes respectively [table 1]). However, these are smaller than the differences between the three hypotheses for the standard ME trees from the RY-coded data. This suggests that resolving for branch 3 on the RY-coded data was influenced little by composition bias. Indeed, for a frequency bias to result in an incorrect phylogeny, the difference between frequency difference trees would have to be considerably greater than between those trees for standard distances. In such cases, many of the frequency differences would be unique or would be splits that do not favor one tree over another.

Stochastic tests, such as the composition homogeneity test, will almost always give highly significant results with large data sets (with constant sites excluded) and provide no indication of the reduced susceptibility to compositional bias that RY-coding appears to confer. In

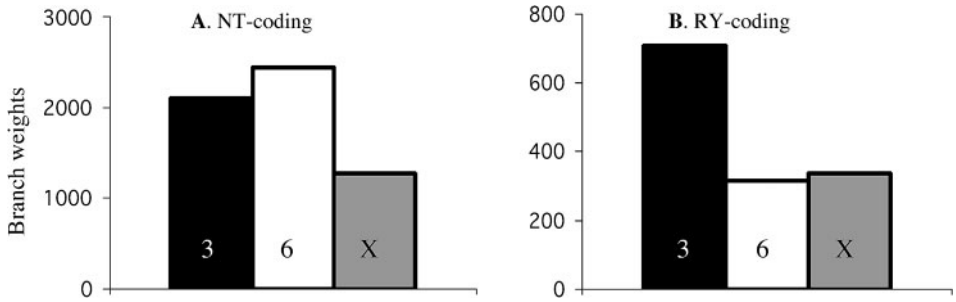

FIG. 2.—Signal estimates from Kimura-corrected weights for the three competing branch hypotheses (3, 6, X). The distance Hadamard transform and Lento plots were done in SpectroNet with the data as standard nucleotides (*A*) and coded as purines R and pyrimidines Y (*B*).

Table 1
Composition Effects, Base-Frequency Scores, and Absolute Distance ME Scores for the Three Alternatives

	Distances from Base Frequencies		Standard Absolute Distances	
	GC Frequency[a]	Purine Frequency[b]	NT-coding[c]	RY-coding[d]
Branch 3	+688.4	+10.9	+152.2	(51,959.0)
Branch 6	(12,530.4)	(2,113.4)	(116,764.4)	+60.0
Branch X	+688.4	+13.5	+289.1	+77.9

NOTE.—The remainder of the tree was constrained as in figure 1 for GC (i.e., G+C) and purine (i.e., A+G) base-frequency distances. In each column, the optimal value is in parentheses.

[a] GC scores may be compared with the NT standard (absolute distance).

[b] Purine ME scores may be compared with the NT standard (absolute distance).

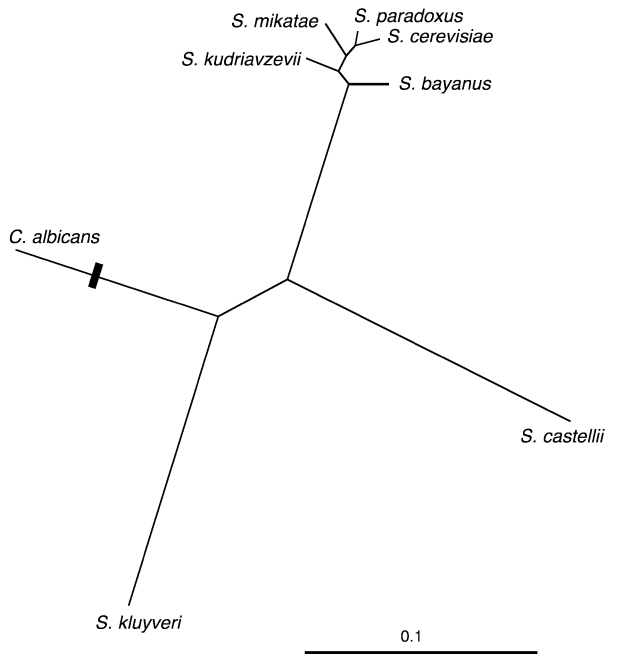

FIG. 3.—Maximum likelihood tree for the RY-coded 106-gene data set. The model used is that of Cavender and Felsenstein (1987) for two-state characters, with ML estimates for $I + \Gamma_4$. Note that the *Candida albicans* branch has been reduced by a factor of 10.

contrast, magnitude tests, such as RCV and treeness (see table S1 in Supplementary Material online), which focus on the size of biases, should be further developed for exploring the relationship between phylogenetic signal and nonhistorical biases.

Conclusions

A swing upon changing optimality criteria or model assumptions from 100% bootstrap support for one branch to 100% for a conflicting branch shows the need for models to fully account for the data. Bootstrap support of 100% is not enough; the tree must also be correct. If there are systematic biases, even phylogenetic analysis of complete genomes can be misled by inconsistency. In the present case, there are strong signals in the data additional to the historical one (fig. 2) that mislead ME, although they are insufficient to mislead ML and MP.

In the past, to reduce sampling error, emphasis has been placed on retrieving the maximum information from sequences. An advantage of genome-scale data sets is that sampling error is reduced so that more conservative approaches can be used to retrieve phylogenetic signal. A classic example with concatenated genes is reducing conflicting signal by excluding third codon positions (e.g., Delsuc et al. [2002]) and/or data partitions that fail tests for compositional heterogeneity (e.g., Springer et al. [1999]). As well as the present usage of RY-coding, focusing on the slowest evolving sites has also been effective (Brinkmann and Philippe 1999). The relative increase in historical versus nonhistorical signal is essential. The benefit of using the most conservative transformations and/or sites is twofold; both the loss of historical signal and build up of systematic biases are slower.

Composition variability and treeness are useful indicators, respectively, of the strength and potential effect of additional biases and support RY-coding as more reliable than standard NT-coding for the present data set. As such the ML tree in figure 3 is our best estimate of the tree and branch lengths for the yeast phylogeny. None of the internodes are especially short relative to the adjacent external branches compared with many deep-level phylogenetic problems such as land plants (Pryer et al. 2001), placental mammals (Amrine-Madsen et al. 2003), and birds (Harrison et al. 2004). This warns of the potential for nonhistorical signals to bias phylogeny among these groups when in the future they too are "fully resolved" with genome-scale data sets.

Traditionally, the tree with the highest likelihood is considered the best estimate, irrespective of any systematic biases. In the present case, models with conservative RY-coding and models resulting in the most even signal distribution for next-best trees all favor branch 3 over branch 6, in agreement with the MP and ML analyzes. Genome-scale data sets provide unprecedented potential for detecting and correcting for nonhistorical signals in real data. Simulation is not relevant here; it is real data that counts. The focus must now be on detecting any systematic biases in the data.

Acknowledgments

Antonis Rokas kindly sent us the aligned yeast data set. Three anonymous referees provided constructive comments. F.D. acknowledges the support of a Lavoisier Postdoctoral Grant from the French Ministry of Foreign Affairs, and MJP and DP acknowledge the support of the New Zealand Marsden Fund.

Literature Cited

Amrine-Madsen, H., K. P. Koepfli, R. K. Wayne, and M. S. Springer. 2003. A new phylogenetic marker, apolipoprotein B, provides compelling evidence for eutherian relationships. Mol. Phylogenet. Evol. **28**:225–240.

Brinkmann, H., and H. Philippe. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. **16**:817–825.

Cavender, J. A., and J. Felsenstein. 1987. Invariants of phylogenies in a simple case with discrete states. J. Classif. **4**: 57–71.

Delsuc, F., M. J. Phillips, and D. Penny. 2003. Comment on "Hexapod origins: monophyletic or paraphyletic?" Science **301**:1482d.

Delsuc, F., M. Scally, O. Madsen, M. J. Stanhope, W. W. de Jong, F. M. Catzeflis, M. S. Springer, and E. J. P. Douzery. 2002. Molecular phylogeny of living xenarthrans and the impact of character and taxon sampling on the placental tree rooting. Mol. Biol. Evol. **19**:1656–1671.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

Harrison, G. L., P. A. McLenachan, M. J. Phillips, K. E. Slack, A. Cooper, and D. Penny. 2004. Four new avian mitochondrial genomes help get to basic evolutionary questions in the late Cretaceous. Mol. Biol. Evol. **21**: (in press).

Hendy, M. D., and M. A. Charleston 1993. Hadamard conjugation—a versatile tool for modelling nucleotide-sequence evolution. N. Z. J. Bot. **31**:231–237.

Hillis, D. M. 1995. Approaches for assessing phylogenetic accuracy. Syst. Biol. **44**:3–16.

Huber, K. T., M. Langton, D. Penny, V. Moulton, and M. D. Hendy. 2002. Spectronet: a package for computing spectra and median networks. Appl. Bioinf. **1**:159–161.

Lockhart, P. J., M. Steel, M. D. Hendy, and D. Penny. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. **11**:605–612.

Lopez, P., D. Casane, and H. Philippe. 2002. Heterotachy, an important process in protein evolution. Mol. Biol. Evol. **19**: 1–7.

Penny, D., and M. D. Hendy. 1986. Estimating the reliability of evolutionary trees. Mol. Biol. Evol. **3**:403–417.

Penny, D., M. D. Hendy, and M. Steel. 1992. Progress with methods for constructing evolutionary trees. Trends Ecol. Evol. **7**:73–79.

Phillips, M. J., Y.-H. Lin, G. L. Harrison, and D. Penny. 2001. Mitochondrial genomes of a bandicoot and a brushtail possum confirm the monophyly of australidelphian marsupials. Proc. R. Soc. Lond. B Biol. Sci. **268**:1533–1538.

Phillips, M. J., and D. Penny. 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. Mol. Phylogenet. Evol. **28**:171–185.

Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14**:817–818.

Pryer, K. M., H. Schneider, A. R. Smith, R. Cranfill, P. G. Wolf, J. S. Hunt, and S. D. Sipes. 2001. Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants. Nature **409**:618–622.

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.

Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. **16**:1114–1116.

Springer, M. S., H. M. Amrine, A. Burk, and M. J. Stanhope. 1999. Additional support for Afrotheria and Paenungulata, the performance of mitochondrial versus nuclear genes, and the impact of data partitions with heterogeneous base composition. Syst. Biol. **48**:65–75.

Steel, M., D. Huson, and P. J. Lockhart. 2000. Invariable sites models and their use in phylogeny reconstruction. Syst. Biol. **49**:225–232.

Swofford, D. L. 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4.0b10. Sinauer Associates. Sunderland, Mass.

Templeton, A. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution **37**:221–244.

Woese, C. R., L. Achenbach, P. Rouviere, and L. Mandelco. 1991. Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts. Syst. Appl. Microbiol. **14**:364–371.

Yang, Z. 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. **39**:105–111.

William Martin, Associate Editor

Accepted March 16, 2004