Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 9979–9981, August 1999

Commentary

SINEs of the perfect character

David M. Hillis[†]

Section of Integrative Biology, School of Biological Sciences, University of Texas, Austin, TX 78712

In their quest to reconstruct the tree of life, evolutionary biologists are constantly looking for new sources of data. Morphological data have been applied to the phylogeny problem since the 1800s, and morphology continues to be the only source of available data on relationships for the vast majority of species on Earth, past and present. However, since the 1960s, molecular data have contributed an increasing fraction of the new information used to reconstruct phylogenetic history (1). Each new source of molecular information has provided a new perspective on evolutionary history, and each technique has a set of advantages and disadvantages. As with morphological approaches, most of these molecular techniques continue to be useful for specific kinds of problems. Nonetheless, almost every new molecular approach to phylogenetic inference has been ballyhooed as capable of "revolutionizing" the field. In truth, no one technique is a perfect solution for all phylogenetic problems, even though each provides us with a new perspective on evolution. Even the oft-perceived superiority of molecular data is largely a matter of timing. As Avise (2) has noted, imagine if we had studied the DNA sequences of organisms for decades without ever seeing a phenotype, and then someone suddenly discovered the morphologies and behaviors that DNA sequences specify! The sense that we were finally making real progress in our understanding of evolution would be at least as great as the sense of wonder and excitement that has accompanied advances in molecular biology.

Despite the usefulness of a varied approach to the study of phylogenetic history, the power of DNA sequencing, especially for "deep-time" problems, has led to the dominance of sequencing in phylogenetic studies in recent years. Thus, it comes as somewhat of a surprise to see a new kind of phylogenetic character presented as a better source of information about evolutionary history. In this issue of the *Proceedings*, Nikaido *et al.* (3) report that insertion events of SINEs and LINEs (short and long interspersed elements, respectively) provide a perfect record of the evolutionary history of the major lineages of artiodactyl mammals and confirm that hippopotamuses are the closest extant relatives of whales.

The relationship between whales and ungulates was first proposed in 1883 (4), but, until the last two decades, there has been little evidence to group whales with any particular group of ungulates. In the process of becoming highly adapted to an aquatic environment, whales have undergone enormous modification to many of the morphological structures that have been used to classify artiodactyls ("even-toed" ungulates, such as hippos, deer, cattle, pigs, and camels) and perissodactyls ("odd-toed" ungulates, such as horses, rhinoceroses, and tapirs). Even the common names of these two groups demonstrate the problem: the hind limbs of whales have been lost, and the fore limbs have been modified into flippers. Nonetheless, numerous morphological and molecular studies have shown that whales are related to, or actually imbedded within, the artiodactyls (5, 6). However, some of the similarities that exist between whales and various artiodactyls, such as several aquatic adaptations shared between whales and hippos [nursing of offspring underwater, communication by underwater vocalizations, lack of hair, and lack of sebaceous glands (7, 8)] have been dismissed, until recently, as convergence to a common environment. Convergence was assumed because, traditionally, hippos were thought to be more closely related to pigs and peccaries than to whales.

In 1985, Sarich (9) proposed that hippos are more closely related to whales than to other artiodactyls, based on a study of serum immunology. Over the past 15 years, this initially controversial hypothesis has been supported by sequence studies of 15 different DNA and protein data sets (5). The relationships of the major lineages of artiodactyls were once considered highly controversial (10, 11), but now artiodactyl (including whale) relationships are the best-resolved portion of the mammalian tree (Fig. 1). The support for the tree is so strong that many mammalogists now consider this a "virtually known" phylogeny (5, 12, 13).

Actually known phylogenies (i.e., groups in which the divergence of lineages has been directly observed by humans) provide a unique way of testing the methods and assumptions of phylogenetic analysis (14). However, the number of truly known phylogenies is rather small and taxonomically limited. Thus, systematists often turn to groups whose relationships are so well supported by multiple analyses that no reasonable person would question their resolution (15). It is in this context that the sequence-based tree shown in Fig. 1 provides a testing ground for the use of insertion events of SINEs and LINEs to infer phylogenetic history.

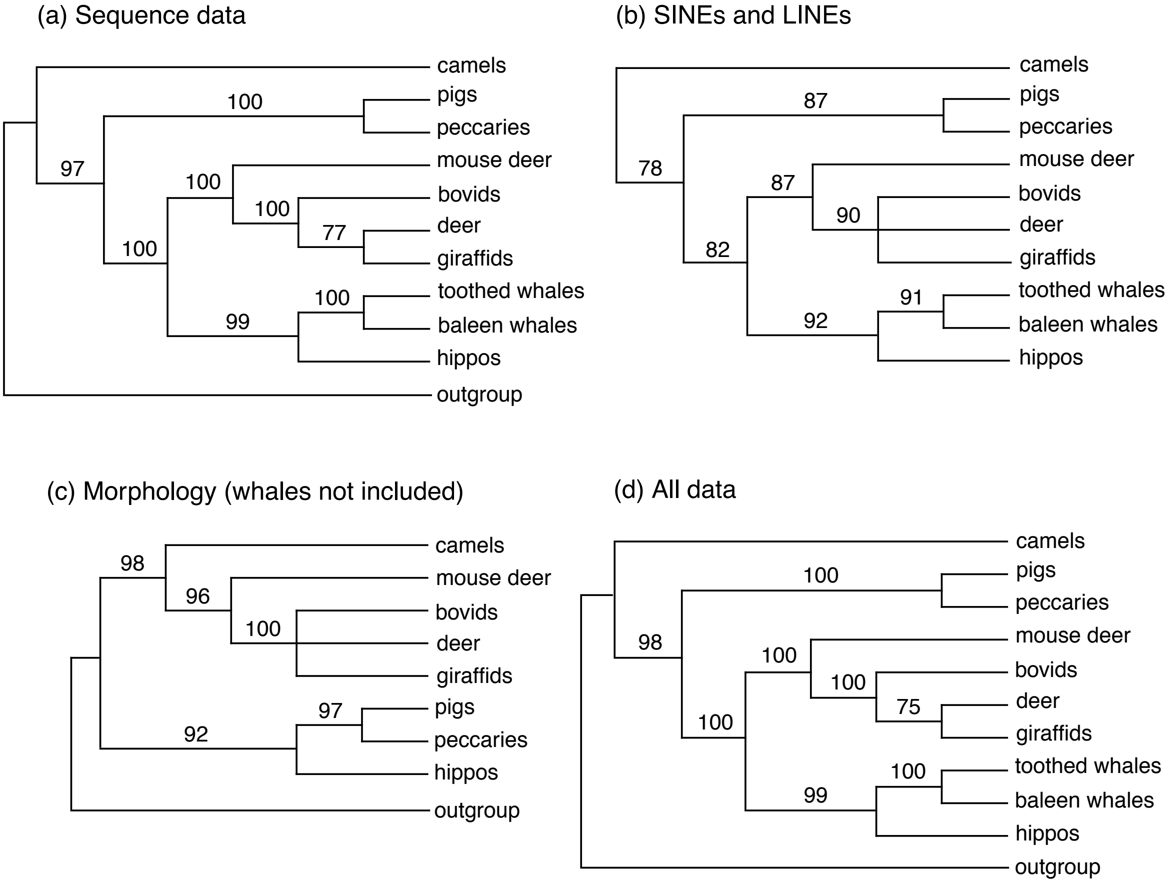
Norihiro Okada and his colleagues (3, 16–19) have argued that insertions of SINEs and LINEs are irreversible events that are unlikely to occur independently in multiple lineages at exactly the same chromosomal locations. Because they believe the probabilities of reversal and convergence to be very small, they state that ". . . the probability that homoplasy will obscure phylogenetic relationships [based on SINEs and LINEs] is, for all practical purposes, zero" (ref. 3, p. 10264). Thus, they argue, ". . . there is no need for statistical analysis and definitive conclusions [about phylogeny] can almost always be drawn" (ref. 19, p. 923). The perfect correspondence between the reported SINE/LINE insertions (3) and the virtually known phylogeny of artiodactyls (5) supports the suggested low level of homoplasy for the insertion events and their usefulness for phylogenetic inference.

The argument that SINEs are never lost once they are inserted has its caveats. For instance, Nikaido *et al.* (3) note that SINEs are often lost as part of larger deletions. However, in this case, they would not score the SINE as absent; instead, they would score the site as "missing information" because the entire locus is lost. If a locus is missing entirely, this is the appropriate way to handle the situation because the SINE may or may not have been there before the deletion. Since the SINEs are examined by amplifying a given locus with the use of conserved flanking sequences, it is also often true that data on a given locus cannot be obtained because of mutations in the flanking sequences. These problems highlight one of the big disadvantages of using SINEs and LINEs: the comparisons are restricted to relatively closely related species. As one attempts to discover SINE/LINE insertion events that mark old evolutionary events, more and more of the taxa need to

PNAS is available online at www.pnas.org.

The companion to this Commentary begins on page 10261.

[†]To whom reprint requests should be addressed. E-mail: hillis@phylo.zo.utexas.edu.



(a) Sequence data

(b) SINEs and LINEs

(c) Morphology (whales not included)

(d) All data

FIG. 1. Phylogenetic estimates of artiodactyl mammals. Numbers above the branches represent bootstrap percentages, which are used as a conservative measure of phylogenetic confidence (32, 33). All trees are based on equally weighted parsimony analyses and were conducted by using the phylogenetic analysis software PAUP* (34). (*a*) Analysis based on sequence data alone. The dataset for 15 genes was obtained from http://www.utexas.edu/ftp/depts/systbiol/48_1/vol48_1.html; the Whippo-1 dataset of Gatesy *et al.* (5) was modified by excluding the data for SINEs and morphological characters from the analysis. (*b*) Analysis based on SINE and LINE data from Nikaido *et al.* (3). (*c*) Analysis based on morphological characters (skeletal and dental characters) presented by Gentry and Hooker (11). (*d*) Combined analysis of sequence data, SINE and LINE data, and morphology (data from refs. 3 and 5, with redundant characters removed).

be scored as "missing data," eventually reaching the point where the analysis is completely uninformative. In the data matrix of Nikaido *et al.* (3), 21% of the elements are missing. For some loci, the loss of information on even one more species would yield the respective locus completely uninformative. The situation is worse for some species than others; the peccary, for instance, is missing data for 90% of the loci. This greatly reduces the confidence of the placement of such taxa based on SINE and LINE data alone.

Even if convergence and reversal are extremely rare for SINE/LINE insertion events, the characters are not immune from problems of lineage sorting of ancestral polymorphisms (20). Lineage sorting can introduce homoplasy when a polymorphism becomes fixed in some, but not all, of the descendants of a polymorphic ancestor. Although this is a potential problem for any kind of data, SINE/LINE analyses are particularly sensitive because of the very small number of independent loci that usually characterize any one clade. As the time between speciation events decreases toward zero, the probability that lineage sorting of a given ancestral polymorphism will provide misleading information about phylogenetic relationships approaches 0.5 (20). As the time between speciation events increases, the probability of incongruence between the data and the true phylogeny decreases. However, the rate of decrease also depends on the effective population size of the species and its generation time. Even if the speciation events are separated by N generations, where N is the effective population size, the probability of incongruence between a single locus and the true phylogeny may still be as high as 0.3 (20). For species with large population sizes and/or long generation times, the probability that a single locus may be misinformative because of lineage sorting is high even when branching events are fairly well separated in time. Thus, even in the absence of any homoplasy, it is highly inadvisable to dispense with statistical analysis altogether.

What of the claim that the SINE/LINE insertion events are perfect markers of evolution (i.e., they exhibit no homoplasy)? Similar claims have been made for other kinds of data in the past, and in every case examples have been found to refute the claim. For instance, DNA–DNA hybridization data were once purported to be immune from convergence (21), but many sources of convergence have been discovered for this technique (22). Structural rearrangements of genomes were thought to be such complex events that convergence was highly unlikely (23), but now several examples of convergence in genome rearrangements have been discovered (e.g., ref. 24). Even simple insertions and deletions within coding regions have been considered to be unlikely to be homoplastic (25), but numerous examples of convergence and parallelism of these events are now known (e.g., ref. 26). Although individual nucleotides and amino acids are widely acknowledged to exhibit homoplasy, some authors have suggested that widespread simultaneous convergence in many nucleotides is virtually impossible (27). Nonetheless, examples of such convergence have been demonstrated in experimental evolution studies (28). All of these sources of data remain useful and important for the inference of phylogeny. Therefore, the presence of homoplasy is not, in itself, terribly problematic, as long as appropriate statistical assessments are made of the inferences based on the data (29).

In the case of the SINE/LINE data presented by Nikaido *et al.* (3), no homoplasy is exhibited among 20 insertion events. Therefore, it is probably safe to assume that homoplasy of SINE/LINE insertions is uncommon, at least among some groups of mammals. Similar data have also been presented for salmonid fishes (18, 19). Although SINEs (or SINE-like elements) are known from several other diverse groups of organisms (17), their usefulness in phylogenetic analysis has not been widely explored, and it is too early to conclude that the studies will be easy to extend to other groups. Moreover, a recent study‡ showed that insertion events of SINEs are sometimes convergent in independent lineages.

Besides the low levels of observed homoplasy, the other principal advantage of SINE/LINE insertion analysis is related to the expected irreversibility of the insertion events. With most types of data, changes from one state to another are possible in either direction, often with equal probability (i.e., a C → G change might be just as likely as a G → C change within a gene sequence). This means that systematists usually must make reference to an outgroup (a group known to be outside the group currently under study) to root the inferred phylogeny. Outgroup rooting may produce artifacts, however, especially if the outgroup is relatively distantly related to the group under study. The correct unrooted tree of ingroup species may be inferred, but the long branch leading to the outgroup may then attach to another long branch within the ingroup [a problem known as "long-branch attraction" (30)]. This problem led, for example, to the incorrect rooting of the initial phylogenetic studies of whales based on mitochondrial DNA, which suggested that the toothed-whales were paraphyletic (31). In contrast, if SINE/LINE insertions are irreversible, then the characters are polarized, and the resulting trees can be rooted without reference to an outgroup. This holds great promise for rooting trees of closely related species that have no close outgroups or for which the close relatives are unavailable for analysis.

Studies of SINEs and LINEs clearly are an important new source of evolutionary information, and the study by Nikaido *et al.* study (3) should do much to stimulate additional development and investigation of the technique. SINE/LINE insertion studies have some advantages for phylogenetic analysis over other approaches, such as the apparent low levels of homoplasy and the expected irreversible nature of the characters. But will studies of SINEs and LINEs make widespread and significant inroads on the current dominance of DNA sequencing studies for phylogenetic inference? In the short term, the answer is no. First, the amount of time, money, and effort needed to collect data on relatively few characters will be prohibitive for most investigators. For most biologists, DNA sequencing will continue to be a far more efficient and cost-effective means of making robust inferences about phylogeny. For instance, as shown in Fig. 1, the support for the artiodactyl tree comes mostly from sequence data. SINEs and LINEs are congruent with the sequence tree but provide weaker support and only support the already well supported clades. Second, considerable background work is needed to transfer the technique to new groups of organisms, even if SINEs and LINEs are found to evolve by similar mechanisms and to be present in high enough numbers to be useful as evolutionary markers in these groups. Third, the use of SINE/LINE insertions is limited to relatively closely related species, and the problems with missing data become severe as investigations extend backward in phylogenetic time. In contrast, genes evolving at different rates can be used to reconstruct phylogenetic relationships from a few years ago to the earliest divergences of life. Fourth, given that SINE/LINE data are not immune from the problems of lineage sorting, homoplasy, or missing data, statistical analyses are still necessary to make robust phylogenetic inferences. The difficulty of collecting data from multiple insertions per clade will make the achievement of statistically robust conclusions relatively difficult, at least compared with DNA sequencing studies. Finally, DNA sequences provide other information of intrinsic interest to many biologists, including the possibility of inferring ancestral sequences, information on sequence evolution, and a means for estimating relative branch lengths and times of divergence. Thus, although SINE/LINE insertion studies provide a welcome, useful, and important new source of phylogenetic data, they are not magic bullets. They do show enough promise for evolutionary studies, however, that it would be highly worthwhile to study SINEs and LINEs more widely in other groups of organisms.

‡Cantrell, M. A. & Wichman, H. A., Society of Systematic Biologists 1999 Meeting, June 22–26, 1999, Madison, WI, p. 12.

1. Hillis, D. M., Moritz, C. & Mable, B. K., eds. (1996) *Molecular Systematics* (Sinauer, Sunderland, MA).
2. Avise, J. C. (1994) *Molecular Markers, Natural History and Evolution* (Chapman & Hall, New York).
3. Nikaido, M., Rooney, A. P. & Okada, N. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 10261–10266.
4. Flower, W. H. (1883) *Proc. Zool. Soc. London* **1883,** 466–513.
5. Gatesy, J., Milinkovitch, M., Waddell, V. & Stanhope, M. (1999) *Syst. Biol.* **48,** 6–20.
6. Thewissen, J. G. M. & Madar, S. I. (1999) *Syst. Biol.* **48,** 21–30.
7. Gatesy, J. (1997) *Mol. Biol. Evol.* **14,** 537–543.
8. Barklow, W. (1995) *Nat. Hist.* **104,** 54.
9. Sarich, V. (1985) in *Evolutionary Relationships Among Rodents: A Multidisciplinary Approach*, eds. Luckett, W. & Hartenberger, J. (Plenum, New York), pp. 423–452.
10. Simpson, G. G. (1945) *Bull. Am. Mus. Nat. Hist.* **85,** 1–350.
11. Gentry, A. W. & Hooker, J. J. (1988) in *The Phylogeny and Classification of the Tetrapods, Vol. 2, Mammals*, ed. Benton, M. J. (Clarendon, Oxford), pp. 235–272.
12. Waddell, P., Okada, N. & Hasegawa, M. (1999) *Syst. Biol.* **48,** 1–5.
13. Liu, F.-G. R. & Miyamoto, M. M. (1999) *Syst. Biol.* **48,** 54–64.
14. Hillis, D. M. (1995) *Syst. Biol.* **44,** 3–16.
15. Cummings, M. P., Otto, S. P. & Wakeley, J. (1995) *Mol. Biol. Evol.* **12,** 814–822.
16. Shimamura, M., Yasue, H., Ohshima, K., Abe, H., Kato, H., Kishiro, T., Goto, M., Munechika, I. & Okada, N. (1997) *Nature (London)* **388,** 666–670.
17. Okada, N. (1991) *Trends Ecol. Evol.* **6,** 358–361.
18. Murata, S., Takasaki, N., Saitoh, M. & Okada, N. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 6995–6999.
19. Murata, S., Takasaki, N., Saitoh, M., Tachida, H. & Okada, N. (1996) *Genetics* **142,** 915–926.
20. Tachida, H. & Iizuka, M. (1993) *Genetics* **133,** 1023–1030.
21. Sibley, C. G. & Ahlquist, J. E. (1987) in *Molecules and Morphology in Evolution: Conflict or Compromise?*, ed. Patterson, C. (Cambridge Univ. Press, Cambridge, U.K.), pp. 95–121.
22. Werman, S. D., Springer, M. S. & Britten, R. J. (1996) in *Molecular Systematics*, eds. Hillis, D. M., Moritz, C. & Mable, B. K. (Sinauer, Sunderland, MA), pp. 169–203.
23. Palmer, J. D., Jansen, R. K., Michaels, H. J., Chase, M. W. & Manhart, J. R. (1988) *Ann. Mo. Bot. Gard.* **75,** 1180–1206.
24. Mindell, D. P., Sorenson, M. D. & Dimcheff, D. E. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 10693–10697.
25. Rivera, M. C. & Lake, J. A. (1992) *Science* **257,** 74–76.
26. Cunningham, C. W., Jeng, K., Husti, J., Badgett, M., Molineux, I. J., Hillis, D. M. & Bull, J. J. (1997) *Mol. Biol. Evol.* **14,** 113–116.
27. Hedges, B. S. & Maxson, L. R. (1996) *Mol. Phylogenet. Evol.* **6,** 312–314.
28. Bull, J. J., Badgett, M. R., Wichman, H. A., Huelsenbeck, J. P., Hillis, D. M., Gulati, A., Ho, C. & Molineux, I. J. (1997) *Genetics* **147,** 1497–1507.
29. Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. (1996) in *Molecular Systematics*, eds. Hillis, D. M., Moritz, C. & Mable, B. K. (Sinauer, Sunderland, MA), pp. 407–514.
30. Hendy, M. D. & Penny, D. (1989) *Syst. Zool.* **38,** 297–309.
31. Messenger, S. L. & McGuire, J. A. (1998) *Syst. Biol.* **47,** 90–124.
32. Felsenstein, J. (1985) *Evolution (Lawrence, Kans.)* **39,** 783–791.
33. Hillis, D. M. & Bull, J. J. (1993) *Syst. Biol.* **42,** 182–192.
34. Swofford, D. L. (1998) *Phylogenetic Analysis Using Parsimony and Other Methods* (Sinauer, Sunderland, MA).