Syst. Biol. 49(4):808–817, 2000

SINE Evolution, Missing Data, and the Origin of Whales

ANDREW M. SHEDLOCK,[1,2] MICHEL C. MILINKOVITCH,[3] AND NORIHIRO OKADA[1,4]

[1]*Tokyo Institute of Technology, Molecular Evolution Laboratory, Graduate School of Bioscience and Biotechnology, Yokohama, Japan*

[2]*The Institute of Statistical Mathematics, Molecular Systematics Group, Tokyo, Japan*

[3]*Free University of Brussels, cp300, Unit of Evolutionary Genetics, Institute of Molecular Biology and Medicine, 6041 Gosselies, Belgium*

The dramatic transformations necessary to evolve from a four-legged terrestrial mammal into a fully aquatic modern whale have generated both widespread fascination and sharp disagreement among biologists as to how such extreme specializations occurred over some 50 million years of vertebrate evolution. Evidence from DNA and protein sequences (Graur and Higgins, 1994; Gatesy, 1998; Milinkovitch et al., 1998) and, most recently, from short interspersed nuclear elements (SINEs; Milinkovitch and Thewissen, 1997; Shimamura et al., 1997, 1999; Nikaido et al., 1999) disrupts the traditional morphology-based view of a monophyletic Artiodactyla (even-toed ungulates) and suggests instead that cetaceans (whales, dolphins, and porpoises) are actually derived members nested deeply within a single "cetartiodactyl" clade. These results make a hippopotamus more closely related to a whale than to a camel or pig, and prompt serious reconsideration of interpretations of fossil data suggesting that mesonychians (an extinct group of hooved terrestrial mammals) gave rise to the extinct cetacean suborder Archaeoceti (Thewissen et al., 1998; Thewissen and Madar, 1999).

Central to the controversy is the scrutiny of molecular character sets for sources of error (Luckett and Hong, 1998; Heyning, 1999a, 1999b; O'Leary, 1999) such as homoplasy attributable to backmutation, inadequate taxon sampling, and inappropriate outgroup rooting. Recent reviews from morphologists on the subject of whale origins warn us that "we should be cautious about wholeheartedly embracing such provocative [molecular] hypotheses of relationships" (Heyning, 1999a). Carefully reviewing molecular data and analyses is generally a healthy practice, and statistical analysis of some DNA sequences, such as those of mitochondrial genes like cytochrome *b*, may indeed be suspect relative to that of the "hard" anatomical synapomorphies such as teeth, postcranial bones, and the astragalus (double-trochleated condition in the skeleton of the hindfoot) that have been used to support artiodactyl monophyly for more than a century (Owen, 1848; Cope, 1888; Schaeffer, 1947; Rose, 1996). However, none of these criticisms levied against DNA sequence studies apply to SINE insertion analysis.

In their recent assessment of morphological and molecular evidence for inferring cetacean ancestry, Luckett and Hong (1998) badly misconstrue how SINEs evolve, misinterpret how they are used for phylogeny inference, and incorrectly dismiss evidence from SINEs that conclusively demonstrate paraphyly of the Artiodactyla. We believe this is a serious problem that is counterpoductive to advancing an accurate understanding of cetacean evolution and unfortunately promotes a negative view regarding the value of molecular data that goes well beyond the current debate about whale origins.

WHAT ARE SINES AND HOW DO THEY EVOLVE?

Clearly, properly understanding the nature of SINE retroposons and how they evolve is fundamental to their intelligent application as phylogenetic tools. Several comprehensive reviews of retroposon evolution, their experimental characterization,

[4]Address correspondence to this author at: Tokyo Institute of Technology, Graduate School of Bioscience and Biotechnology, 4259 Nagatsuta-cho, Midori-ku, Yokohama 226-8501, Japan. E-mail: nokada@bio.titech.ac.jp

and intrinsic value as molecular systematic markers are available (Weiner et al., 1986; Deininger and Batzer, 1993; Schmid, 1996; Shedlock and Okada, 2000). Nevertheless, it is instructive to review here selected aspects of SINE molecular biology and evolution that are especially relevant to their use for phylogeny inference.

Roughly half of the higher eukaryotic genome is composed of a variety of repetitive sequences with no obvious function (Smit and Riggs, 1995). Of these, SINEs are among the most numerous (Kazazian and Moran, 1998) and form a class of dispersed, repetitive, mobile molecules that typically range between 75 and 500 bases in size and may be present at well $>10^4$ total copies in eukaryotes, making the hundreds of thousands of SINE movements per organism a powerful potential agent of genome evolution (Brosius, 1991; Schmid, 1996). Dispersed, mobile repeat sequences can amplify and move from a parent locus to a target locus and are fundamentally different from tandemly repeated sequences, such as microsatellites, which may arise by gene duplications (Singer and Berg, 1991). Unlike DNA-mediated transposons, or so-called "jumping genes," the retroposons, including SINEs, require an RNA intermediate for their indirect movement about the genome, and hence "retroposition" themselves by way of the reverse flow of genetic information from RNA back into chromosomal DNA (Rogers, 1983; Weiner et al., 1986).

Retroposons are further subdivided into the viral and nonviral superfamilies based on common structure and the important feature of whether they encode for reverse transcriptase (RTase), an essential enzyme for self-amplification. SINEs are nonviral elements typically derived from cellular tRNA that do not encode for RTase and thus are dependent on other viral superfamily retroelements, namely, long interspersed elements (LINEs), for their amplification and relocation about the genome (Okada et al., 1997). Briefly, SINEs may acquire recognition sites for RTase from specific partner LINEs, which allows them to amplify and reintegrate back into the genome at nicked sites by a mechanism known as target-primed reverse transcription, whereby RTase primes reverse transcription of the cDNA copy directly onto the target locus (Luan et al., 1993). Although hotspots of insertion have been observed in exceptional cases (H. A. Wichman, unpublished data), and human Alu repeats may preferentially integrate into locally AT-rich regions in the R-bands of chromatin (Korenberg and Rykowski, 1988; Matera et al., 1990), SINEs are generally found dispersed throughout the entire genome. Furthermore, no process that specifically removes SINEs is evident from extensive comparative study, including the close scrutiny for their strict retention at specific loci in distantly related taxa (Koop et al., 1986; Deininger and Batzer, 1993; Shedlock and Okada, 2000). A schematic of a proposed model for SINE amplification is shown in Figure 1 to illustrate the irreversible nature of SINE movement between parent and target loci.

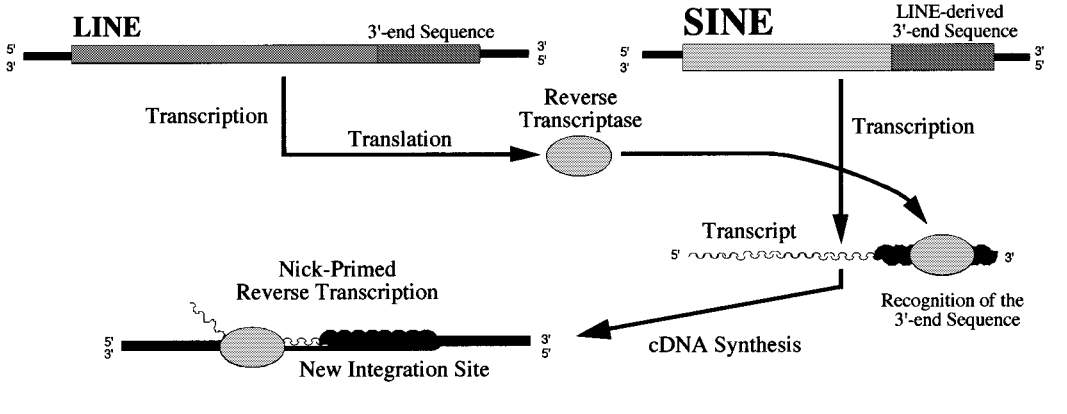

FIGURE 1. General model for irreversible SINE amplification by retrotransposition. The 3′-end sequences of SINEs are derived from specific partner LINEs, which can encode for reverse transcriptase (RTase). RTase recognizes the shared 3′ tail on the SINE transcript, allowing cDNA synthesis and insertion at a new locus of the host genome by way of target-primed reverse transcription.

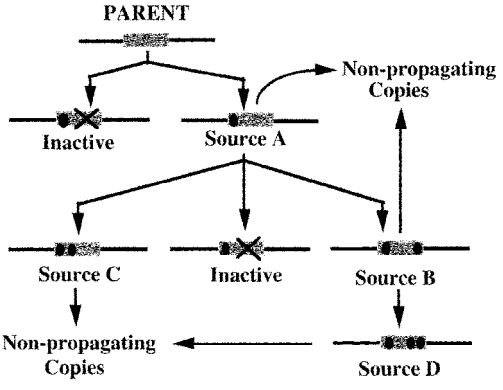

FIGURE 2. Multiple source gene model for SINE evolution. Some offspring of a parent SINE become inactive from accumulated mutation, and others can propagate in the same capacity of their parents, hence serving as multiple sources (A, B, C, D). In this model, SINEs evolve as parasites in the host genome with the amplification rate determined by a balance between deactivation of loci and overall production of new copies.

These aspects of SINE retroposition make them especially stable in the host genome relative to more well-known DNA-mediated transposons, such as P and mariner (Kidwell, 1993; Clark and Kidwell, 1997; Hartl et al., 1997; Robertson, 1997). SINEs use a replicative mechanism of integration, and empirical evidence suggests that they typically evolve by a multiple-source gene model that allows active copies to be maintained in a lineage over evolutionary time (Schmid and Maraia, 1992). In this model, illustrated by the schematic in Figure 2, SINE offspring copies can potentially propagate with the same capacity of parent loci and can thus serve as "multiple sources" for subsequent SINE amplification. The chromosomal environment at the site of insertion can influence the rate of amplification of each element, making each SINE unique in terms of its capability for amplification. Hence, the amplification rate will increase or decrease over time, depending on whether or not accumulated mutations deactivate loci faster or slower than the overall production of new copies.

The long-term stability of retrotransposons over hundreds of millions of years and their clearly vertical pattern of inheritance has been demonstrated nicely for LINEs (Malik et al., 1999). Little empirical evidence for horizontal transfer of SINEs has been observed, despite characterization of these molecules in a wide variety of taxa (see Hamada et al., 1997; Shedlock and Okada, 2000, for examples). The overall evidence suggests that copies of SINEs typically insert once into the genome and then simply stay put, only to decay over millions of years by random mutation. Repetitive elements clearly seem to have played a major role in the emergence and architecture of eukaryotic genomes (Charlesworth et al., 1994). From a macroevolutionary perspective, perhaps retroposon-like elements were crucial to the origin of DNA from the "RNA world," such that the behavior of SINEs observed today simply represents a continuation of that ancient genome-building process (Brosius, 1991; Jurka, 1998), much older than more modern and commonly appreciated DNA-mediated events.

Characterization of all SINE families to date clearly indicate that they arise in and are restricted to particular phylogenetic groups of organisms. This pattern provides the essential foundation for their application to systematic biology: *In the absence of incomplete lineage sorting, copies of the same fixed SINE at a given locus in different host taxa define a clade, and the absence of SINE insertion at that exact locus is an ancestral condition that defines an outgroup*. There is no compelling empirical or theoretical reason to expect SINEs to commonly insert at the exact same locus or be transferred or removed from the genome in any way that is ambiguous by standard methods of experimental detection. We emphasize that absence of SINE detection is not to be confused with detection of SINE absence, and aspects of this important methodological issue are discussed below. Obviously, as more SINE data accumulate for various taxonomic groups, estimating the level at which these basic assumptions are met should become possible. Certainly, based on all available evidence, SINE markers should continue to present levels of homoplasy resulting from reversals, convergence, and parallelism that are far less than those typically observed in DNA sequence data and many morphological comparisons.

SINE LIFESPANS

The active lifespan of SINEs is another important aspect of their evolution that is directly relevant to their use as systematic tools. The sequence comparison and

secondary structure of most SINEs indicate that they are typically derived from cellular tRNA and historically have probably arisen through recombination with transcriptionally active parent LINEs, which have been shown to persist stably in the eukaryotic genome well > one billion years (Terai et al., 1998; Malik et al., 1999; Ogiwara et al., 1999). Under the multiple-source gene model, SINEs born into the genome may experience variable rates of amplification and be eventually rendered dead or inert if they find themselves in unfavorable chromosomal environments for amplification or if they accumulate enough deleterious mutations. Furthermore, the parasitism of SINEs on specific partner LINEs to acquire retropositional activity subjects them to eventual certain death if parent LINEs in the same host become inactive (Okada et al., 1997).

The taxonomic scope of a given SINE family may be broad or restricted as can be the extent of apparent divergence among its members. SINE elements themselves are generally considered to be neutrally evolving parasites on the host genome (Weiner et al., 1986; Okada, 1991). Thus, if the average sequence divergence among members of a family is relatively small, then presumably they amplified in the recent past and the family is relatively young; conversely, large divergences among family members suggest that they are much older elements and may have died in the host genome in the distant past. It is therefore possible to have two different detectable SINE families where one is active and the other is dead. SINE families specific to particular taxonomic groups can be organized into subfamilies based on sequence comparison. The same principles described above for SINE families also apply at the subfamily. Because it is the historical pattern of proliferation of SINEs that allows them to resolve topologies of common ancestry among host taxa, the life span of a given SINE subfamily, not just its absolute age from birth to the present, defines the historical timeframe within which a given member of that subfamily is likely to prove phylogenetically informative. In other words, living SINEs may be able to resolve very recent lineage splits, whereas dead SINEs potentially can resolve only divergence events that occurred during their historical period of active proliferation.

Fixation, Incomplete Lineage Sorting, and Sampling

It is important to distinguish between the amplification of SINE copies in the germ line of an individual organism and the fixation or loss of particular loci in a species by way of genetic drift over successive generations of individuals in a population. If a SINE is very young, it may not be fixed in all individuals of a given species. Any locus that exhibits a polymorphic pattern of insertion in multiple individuals of the same species can be considered unfixed, and its status as a synapomorphy remains unclear. Thus, unfixed SINEs are clearly inappropriate for diagnosing common ancestry among species. Confirming fixation of a given SINE locus places an emphasis on sampling numerous individuals within a species. In cases where the species being diagnosed for common ancestry have diverged in the distant past (e.g., millions of years ago), there is little concern that all loci being examined will not be completely fixed. However, very young species demand more careful attention to the need for sampling multiple population samples to be able to confirm fixation with good certainty. In such cases, searching for insertions of multiple independent loci that diagnose the same clade is particularly valuable for establishing a final hypothesis of relationships. Also, the time to fixation under neutral conditions (Kimura, 1983; Nei, 1987) is typically far shorter than the length of time between lineage splits. Hence, it is reasonable to presume that the great majority of SINEs studied to date have been fixed before successive speciation events. Nevertheless, if the time to fixation transcends species boundaries formed in rapid succession, such as can occur during explosive radiations, inconsistent patterns of SINE insertions may be observed because of ancestral polymorphism (Nei, 1987; Takahata, 1989; Wu, 1991).

Figure 3 shows a diagram of the potential situation leading to inconsistent SINE insertion results. In such cases, SINEs will not reflect the true organismal phylogeny and are subject to the same limitations that have been described for other commonly used molecular markers such as mitochondrial DNA (mtDNA) and major histocompatibility loci (Figueroa et al., 1988; Avise, 1994). In this sense, incomplete lineage sorting of SINEs may lead to homoplasy, as broadly defined

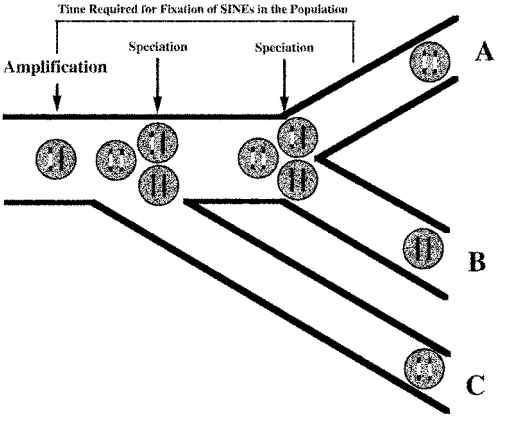

FIGURE 3. Schematic depicting potential for incomplete lineage sorting of SINEs. Diploid individuals (circles) in a population are unfixed for SINE insertions at a given locus (white tags). If speciation occurs rapidly after initial amplification (arrows) and before fixation (bracket), ancestrally polymorphic patterns of SINE insertions may be evident for taxa A, B, and C.

by the empirical presence of character conflict in a data matrix, even in the absence of ambiguity due to convergence and reversal. For some closely related species and for clades that are defined by a rapid succession of speciation events, ancestral polymorphism may be an important concern, and the assumption that SINE insertion at a particular locus defines a clade must be considered carefully with respect to evidence from multiple independent loci.

Recently, Hillis (1999) and Miyamoto (1999), although both supportive of the SINE method, highlighted the importance of lineage-sorting effects on the application of SINE insertion analysis. Although SINE data are clearly not exempt from this problem, the suggestion that they are particularly sensitive to it is not justified. The confounding effects of lineage sorting as encountered with DNA sequence results cannot be directly weighed against SINE insertion analysis, because each SINE locus considered represents the equivalent of an entirely separate independent gene sequence study (e.g., data from the entire linked mtDNA genome may be confounded by ancestral polymorphism and is thus equivalent to a single SINE insertion event in this respect). Furthermore, with DNA sequences, one can only speculate as to whether ancestral polymorphism is in fact responsible for ambiguous results, whereas SINE data can clearly identify and diagnose the problem. Because SINE polymorphisms are themselves valuable indicators of alternative pathways of alleles that are identical by descent, they have been employed to examine the scope and historical profile of this important evolutionary phenomenon with precision unavailable using other standard markers (Hamada et al., 1998; Takahashi, Terai, Nishida, and Okada, in prep.). Evidence to date from the majority of loci analyzed across a wide taxonomic spectrum does not suggest that character incongruence related to differential lineage sorting is a common, prohibitive problem for SINE insertion analysis (Shedlock and Okada, 2000). However, as new data emerge for groups that show patterns of inconsistency (e.g., at a single locus in felid carnivores, J. P. Slattery, pers. comm.), the examination of additional loci and careful evaluation of such sources of homoplasy within the context of how SINEs evolve will be essential for properly interpreting inconsistent SINE results and for maintaining a clear perspective on the proven value of these genetic markers (Shedlock and Okada, in prep.).

With respect to sampling genetic loci, Miyamoto (1999) suggested that SINEs derived from the same historical amplification event could be considered part of a single character complex and thus are not truly independent in nature. He thus advocated using only elements from different SINE subfamilies to construct cladograms. As outlined below, SINE families and subfamilies are established on the basis of sequence similarity. Although the taxonomic distribution of SINEs is uneven by the nature of their historical amplification profile, this fact, along with family classification, is irrelevant to the independent nature of SINE retroposition at sites throughout the eukaryotic genome and its indication of common ancestry. Each inserted SINE at a particular locus evolves independently, regardless of its amplification origin, and is essentially not different from an irreversible point mutation that will eventually become fixed or lost by random genetic drift over time.

EXPERIMENTAL DIAGNOSIS OF SINE INSERTIONS

Detailed protocols for isolating novel SINE loci and experimentally characterizing them for phylogeny inference has been published in numerous studies reviewed elsewhere (Shedlock and Okada, 2000). Briefly, new SINEs are isolated by use of in vitro

transcription of total genomic DNA (Endoh and Okada, 1986), which provides a basis for subsequent hybridization probing for phylogenetically informative loci in genomic libraries of selected taxa that are pertinent to the systematic question of interest. Once novel loci are isolated from a genomic library, they can be subcloned, sequenced, characterized into subfamilies on the basis of patterns of similarity, and used for designing locus-specific polymerase chain reaction (PCR) primers necessary for SINE insertion analysis.

The primary data for SINE insertion analysis are obtained by three basic experimental steps: (1) amplification of specific SINE loci by using PCR primers that anneal to the flanking sequences immediately adjacent to inserted SINE elements; (2) Southern hybridization of a blot of the PCR products by using a unit of the SINE element sequence as a probe; and (3) Southern hybridization of the same blot of products by using the SINE flanking sequence as a probe. Steps 2 and 3 confirm PCR target amplification fidelity in 1, and all PCR products in the analysis can easily be sequenced and examined for diagnostic features at the nucleotide level.

Phylogenetic Considerations and Published SINE Cetartiodactyl Data

The above experimental process draws attention to several practical issues for systematists. First of all, choosing an optimal species from which to create a genomic library for isolation of informative SINE loci is often based on results from other phylogenetic studies of morphology, allozymes, or DNA sequences. Hence, the design of SINE analyses clearly benefits from the careful consideration of other types of character data. Second, designing PCR primers in neutrally evolving SINE flanking sequences will be limited by the extensive divergence between taxa (~20–30% empirical sequence difference), which practically limits SINEs to systematic problems that do not involve ancient lineage splits dating back much more than 75–100 million years or so. Thus although they are ideal for addressing many species, genera, family, and interordinal questions, they are clearly inappropriate for resolving relationships among much older clades. This practical limit to acquiring primary data should not be confused with the phylogenetic value of retroposon insertions at specific loci that have been carefully characterized by the method. Recently, Luo (2000) confused this issue by suggesting SINE "estimates" may be inappropriate for inferring artiodactyl relationships involving divergences much older than 50 million years. As stated, Luo's misleading concern about the effects of mutational decay in SINE flanking sequences is logically inconsistent with the very existence of the published data in question (Shimamura et al., 1997; Nikaido et al., 1999). Lastly, the PCR-based assay will typically involve the absence of amplification at specific loci in specific taxa being examined, usually because of inefficiency of PCR priming at annealing sites or perhaps to something less common, such as gross deletion of an entire locus. In such cases, we can say nothing about the nature of SINE insertion at that locus in that particular taxon, and the lack of information can be considered a form of missing data.

We have found this latter issue regarding missing data to be a particularly common point of confusion regarding the proper interpretation of evidence for SINE insertion and its methodological integrity. Figure 4 diagrams the problem of confusing the absence of SINE evidence with the evidence of SINE absence, and how this can lead to drawing erroneous conclusions about SINE results and sampling design. Such confusion is a fundamental problem in the comments published by Luckett and Hong (1998) that are now being cited inappropriately (e.g., Heyning, 1999a) as a careful reconsideration of molecular evidence regarding artiodactyl paraphyly.

As reviewed above, it is well established that copies of the same SINE shared in two different taxa are derived from the same historical insertion event in the germ line of a common ancestor, and thus define a monophyletic group. Comparative studies suggest that the likelihood of SINEs being independently inserted at the same locus in different lineages (or precisely excised) is exceedingly small; furthermore, cases of horizontal transfer are also not problematic for SINE insertion analysis because diagnosis is focused on the presence or absence of a SINE at a particular locus. Hence, these characters can be expected to show exceptionally low levels of noise from reversals, parallelisms, and convergences. For instance, under the approximate assumption that a given SINE copy could be inserted at any locus in a

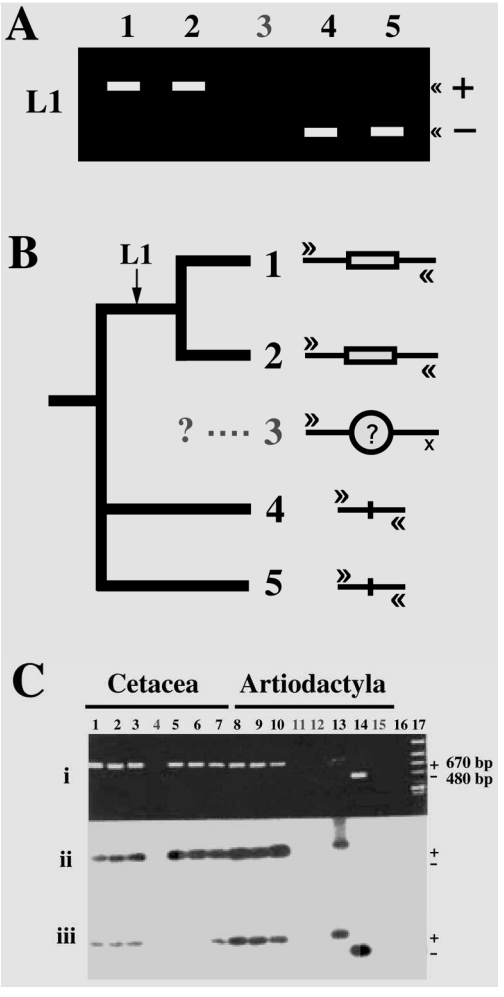

FIGURE 4. Interpreting SINE data: Absence of PCR amplification does not equal absence of SINE insertion. (A) Schematic of gel banding patterns for SINE PCR products for locus L1 and five taxa. Sizes of expected products for presence (+) and absence (−) of L1 insertion are indicated. Lane 3 depicts no PCR amplification. (B) Cladogram and historical point of initial L1 isertion diagnosed by banding patterns in (A). For each taxon, boxes indicate L1 elements inserted into the genome, confirmed experimentally by Southern hybridizations and sequencing of all + and − PCR products; arrows indicate PCR primer sites in flanking sequences. Taxa 1 and 2 form a clade diagnosed by L1 insertion (+ bands); 4 and 5 lack insertions (− bands) and are thus defined as outgroups. Nothing about L1 insertion can be concluded for taxon 3 (circle with "?") because of possible inadequate priming in flanks ("x"). Absence of PCR amplification indicates neither absence of SINE insertion nor a possible false signal in the SINE data set. (C) Published SINE evidence for artiodactyl paraphyly from Shimamura et al. (1997): (i) PCR products; (ii and iii) Southern hybridizations of gel blot from (i) by using the SINE element sequence as a probe (ii) and the locus-specific primer as a probe (iii); see original report for details of methods used and taxa and loci examined. For the single locus illustrated, four taxa show no amplification (gray lane numbers 4, 11, 12, 15). These patterns and others not shown are mistaken by Luckett and Hong (1998) to indicate absence of SINE insertion, leading to misunderstanding of retroposon evolution, SINE analysis, and its unambiguous indications of cetacean ancestry.

mammalian genome (~10^9 bp; Lewin, 1994), the possibility that the four independent SINE loci diagnosing the hippo-cetacean clade (Shimamura et al., 1997; Nikaido et al., 1999) would be inserted at the same site by chance would be on the order of 1 in 10^{36}. The probability of false-negative signals from specific deletions would be similarly remote and would be obvious anyway from the examination of multiple loci. Suggesting that low taxon sampling can produce spurious phylogenetic groupings (Luckett and Hong, 1998:165) is irrelevant for SINE data because taxon sampling has obviously no effect on phylogeny inference if the dataset is free from the effects of lineage sorting. Concerns that "inappropriate" outgroups can have adverse effects on character polarization and ingroup relationships is equally irrelevant to cladogram construction with SINEs, because the absence of SINE insertion is the ancestral condition and unambiguously identifies outgroups when there is no ancestral polymorphism evident. These points also underscore the dismissal of published SINE results of Shimamura et al. (1997) by O'Leary and Geisler (1999:461) that are highly pertinent to evaluating outgroup rooting problems and the systematic importance of long-extinct fossil taxa that diverged early from cetaceans.

STATISTICAL CONCERNS

The above issues are well illustrated by encoding artiodactyl SINE insertion data published by Shimamura et al. (1997) into a binary character matrix consisting of irreversible insertion (1), lack of insertion (0), and missing data where no PCR amplification occurred or was not assayed (?) and then running a maximum parsimony analysis with PAUP* (Swofford, 1996; Nikaido et al., 1999). Logically, with no character conflict in the dataset, unambiguous support for the same relationships among major "cetartiodactyl" groups is obtained by the parsimony search; the consistency index is 1.0, the retention index is 1.0, and the homoplasy index is 0.0. This fact underscores the erroneous interpretations by Luckett and Hong (1998),

who advocate the need for "hard" molecular synapomorphies that can be considered with the same criteria for homology as morphological characters (sensu Hennig, 1966) but then ironically dismiss SINEs, which provide a fine solution to this fundamental problem in modern systematic biology.

Clearly an increase in the amount of missing character information could eventually lead to loss of phylogenetic resolution—this logically holds true for any type of dataset. However, potential loss of resolution because of a lack of available SINE information should not be confused with the nature of how SINE characters evolve and can be rigorously diagnosed experimentally. In this respect, the missing data in the character matrix presented by Nikaido et al. (1999) require clarification: The high proportions of missing data (?) for the peccary and chevrotain are due to the lack of completed experimental assays versus failed PCR priming in locus-specific flanking regions, although this is not distinguished in the text (M. Nikaido and N. Okada, pers. comm.). Hence, this confusing oversight makes the promptly noted concerns of Hillis (1999:9980) and Miyamoto (1999:817), regarding loss of PCR priming efficiency with increasing divergence among artiodactyl host taxa, overstated. Both a high proportion of missing data and the presence of character conflict due to incomplete lineage sorting create situations that invite statistical evaluation of confidence in SINE-based hypotheses. However, the optimal method for exploring such cases is not obvious from either philosophical or numerical standpoints. In the absence of character conflict, the meaning of a bootstrap test for irreversible SINE insertions is questionable. In this respect, the comparison Hillis (1999:Fig. 1) made of published datasets is difficult to justify and should be interpreted cautiously because of the large error bias and the loss of information expected for fewer SINE data relative to those from the numerous nucleotide substitutions evaluated (Sanderson, 1995). The statistical treatment of inconsistent SINE data and the intelligent integration of high-confidence insertion results with other character types are important areas for new methods development that will be critical to the expanded use of SINE analysis in general.

In conclusion, confusion between the statistical nature of inferring phylogeny from DNA sequences and constructing cladograms based on SINE insertion analysis can lead to dangerous generalizations in the literature about the unreliability of molecular data, exemplified by mistakes published by Luckett and Hong (1998) and recent statements by Heyning (1999a, 1999b) and O'Leary and Geisler (1999). Reconstructing artiodactyl–cetacean relationships is a challenging evolutionary puzzle that for the first time in >150 years can be understood by using shared, derived molecular characters that are virtually free of noise from reversals, convergence, and rooting artifacts and thus allow for powerful evaluation of conflicting results from maximum likelihood or parsimony analyses of DNA sequences and cladistic studies of morphology. SINEs deserve careful attention from paleontologists, mammalogists, and molecular biologists alike as one of the most important references available for developing an accurate, multidisciplinary explanation for the fascinating origin of whales.

ACKNOWLEDGMENTS

We thank M. Hasegawa for encouragement, advice, and critical review of the manuscript. Two anonymous reviewers, and R. Olmstead, provided numerous suggestions that improved the final version of the paper. M. Nei and A. Rooney gave helpful input concerning matrix analysis. The Ministry of Education, Science, Sports, and Culture of Japan (Monbusho) and the Japan Society for the Promotion of Science provided generous fellowship support to A.M.S. M.C.M. is supported by grants from the National Fund for Scientific Research, Belgium; The Free University of Brussels; the Defay Fund; and the Van Buuren Fund. Monbusho supplies Grant-in-Aid for Specially Promoted Research to N.O.

REFERENCES

AVISE, J. C. 1994. Molecular markers, natural history and evolution. Chapman and Hall, New York.

BROSIUS, J. 1991. Retroposons—seeds of evolution. Science 251:753.

CHARLESWORTH, B., P. SNEIGOWSKI, AND W. STEPHAN. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371:215–220.

CLARK, J. B., AND M. KIDWELL. 1997. A phylogenetic perspective on P transposable element evolution in *Drosophila*. Proc. Natl. Acad. Sci. USA 94:11428–11433.

COPE, E. D. 1888. The Artiodactyla. Am. Nat. 22:1079–1095.

DEININGER, P. L., AND M. A. BATZER. 1993. Evolution of retroposons. Evol. Biol. 27:157–196.

ENDOH, H., AND N. OKADA. 1986. Total DNA transcription *in vitro*: a procedure to detect highly repetitive and transcribable sequences with tRNA-like structures. Proc. Natl. Acad. Sci. USA 83:251–255.

FIGUEROA, F., E. GUNTHER, AND J. KLEIN. 1988. MHC polymorphism pre-dating speciation. Nature 355:265–267.

GATESY, J. 1998. Molecular evidence for the phylogenetic affinities of Cetacea. Pages 63–111 *in* The emergence of whales, evolutionary patterns in the origin of Cetacea (J. G. M. Thewissen, ed.), Plenum, New York.

GRAUR, D., AND D. G. HIGGINS. 1994. Molecular evidence for the inclusion of cetaceans within the order Artiodactyla. Mol. Biol. Evol. 11:357–364.

HAMADA, M., Y. KIDO, M. HIMBERG, M. HASEGAWA, AND N. OKADA. 1997. Characterization of a newly isolated family of short interspersed repetitive elements (SINEs) in coregonid fish (whitefish) with sequences that are almost identical to those of the Sma I family of repeats: possible evidence for the horizontal transfer of SINEs. Genetics 146:369–380.

HAMADA, M., N. TAKASAKI, J. D. REIST, A. L. DECICCO, A. GOTO, AND N. OKADA. 1998. Detection of the ongoing sorting of ancestrally polymorphic SINEs toward eventual fixation or loss in populations of two species of charr during speciation. Genetics 150:301–311.

HARTL, D. L., A. R. LOHE, AND E. R. LOZOVSKYA. 1997. Modern thoughts on an ancyent marinere: function, evolution, regulation. Annu. Rev. Genet. 31:337–358.

HENNIG, W. 1966. Phylogenetic systematics. Univ. Illinois Press, Urbana-Champaign.

HEYNING, J. E. 1999a. Whale origins—conquering the seas. Science 283:943.

HEYNING, J. E. 1999b. Whale origins, response. Science 283:1641.

HILLIS, D. M. 1999. SINEs of the perfect character. Proc. Natl. Acad. Sci. USA 96:9979–9981.

JURKA, J. 1998. Repeats in genomic DNA: mining and meaning. Curr. Opin. Struct. Biol. 8:333–337.

KAZAZIAN, H. H. JR., AND J. V. MORAN. 1998. The impact of L1 retrotransposons on the human genome. Nat. Genet. 19:19–24.

KIDWELL, M. G. 1993. Lateral transfer in natural populations of eukaryotes. Annu. Rev. Genet. 27:235–256.

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge Univ. Press, Cambridge.

KOOP, B. F., M. M. MIYAMOTO, J. E. EMBURY, M. GOODMAN, J. CZELUSNIAK, AND J. L. SLIGHTCOM. 1986. Nucleotide sequence and evolution of the orangutan epsilon globin gene region and surrounding Alu repeats. J. Mol. Evol. 24:94–102.

KORENBERG, J. R., AND M. C. RYKOWSKI. 1988. Human genome organization: Alu, LINE and the molecular structure of metaphase chromosome bands. Cell 53:391–400.

LEWIN, B. 1994. Genes V. Oxford Univ. Press, New York.

LUAN, D. D., M. H. KORMAN, J. L. JAKUBCZAK, and T. H. EICKBUSH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-RTL retrotransposition. Cell 72:595–605.

LUCKETT, W. P., AND N. HONG. 1998. Phylogenetic relationships between the orders Artiodactyla and Cetacea: A combined assessment of morphological and molecular evidence. J. Mammal. Evol. 5:127–182.

LUO, Z. 2000. Evolution: in search of the whales' sisters. Nature 404:235–239.

MALIK, H. S., W. D. BURKE, AND T. H. EICKBUSH. 1999. The age and evolution of non-LTR retrotransposable elements. Mol. Biol. Evol. 16:793–805.

MARAIA, J. (ed.). 1995. The impact of short interspersed elements (SINEs) on the host genome. R. J. Landes, Austin, Texas.

MATERA, A. G., U. HELLMAN, AND C. W. SCHMID. 1990. A transpositionally and transcriptionally competent Alu subfamily. Mol. Cell. Biol. 10:5424–5432.

MILINKOVITCH, M. C., AND J. G. M. THEWISSEN. 1997. Even-toed fingerprints on whale ancestry. Nature 388:622–624.

MILINKOVITCH, M. C., M. BERUBE, AND P. POLSBØLL. 1998. Cetaceans are highly derived artiodactyls. Pages 113–131 *in* The emergence of whales, evolutionary patterns in the origin of Cetacea (J. G. M. Thewissen, ed.), Plenum, New York.

MIYAMOTO, M. M. 1999. Perfect SINEs of evolutionary history? Curr. Biol. 9:816–819.

NEI, M. 1987. Molecular evolutionary genetics. Columbia Univ. Press, New York.

NIKAIDO, M., A. P. ROONEY, AND N. OKADA. 1999. Phylogenetic relationships among cetartiodactyls based on evidence from insertions of SINEs and LINEs: hippopotamuses are the closest extant relatives of whales. Proc. Natl. Acad. Sci. USA 96:10261–10266.

OGIWARA, I., M. MIYA, K. OHSHIMA, AND N. OKADA. 1999. Retropositional parasitism of SINEs on LINEs: identification of SINEs and LINEs in elasmobranchs. Mol. Biol. Evol. 16:1238–1250.

OKADA, N. 1991. SINEs: short interspersed repeated elements of the eukaryotic genome. TREE 6:358–361.

OKADA, N., M. HAMADA, I. OGIWARA, AND K. OHSHIMA. 1997. SINEs and LINEs share common 3′ sequences: a review. Gene 205:229–243.

O'LEARY, M. A. 1999. Whale origins. Science 283:1641.

O'LEARY, M. A., AND J. H. GEISLER. 1999. The position of Cetacea within Mammalia: phylogenetic analysis of morphological data from extinct and extant taxa. Syst. Biol. 48:455–490.

OWEN, R. 1848. Descriptions of teeth and portions of jaws of two extinct anthracotheroid quadrupeds (*Hyopotamus vectianus and H. bovinus*) discovered by the Marchioness of Hastings in the Eocene deposits on the N.W. coast of the Isle of Wight: with an attempt to develope Cuvier's idea of the classification of pachyderms by the number of their toes. Q. J. Geol. Sci. 4:103–141.

ROBERTSON, H. M. 1997. Multiple Mariner transposons in flatworms and hydras are related to those of insects. J. Hered. 88:195–201.

ROGERS, J. 1983. Retroposons defined. Nature 301:460.

ROSE, K. D. 1996. On the origin of the order Artiodactyla. Proc. Natl. Acad. Sci. USA 93:1705–1709.

SANDERSON, M. J. 1995. Objections to bootstrapping phylogenies: a critique. Syst. Biol. 44:299–320.

SCHAEFFER, B. 1947. Notes on the origin and function of the artiodactyl tarsus. Am. Mus. Novit. 1356:1–24.

SCHMID, C. 1996. Alu: structure, origin, evolution, significance and function of one-tenth of human DNA. Prog. Nucl. Acid Res. Mol. Biol. 53:283–319.

SCHMID, C. W., AND R. MARAIA. 1992. Transcriptional regulation and transpositional selection of active SINE sequences. Curr. Opin. Genet. Devl. 2:874–882.

SHEDLOCK, A. M., AND N. OKADA. 2000. SINE insertions: powerful tools for molecular systematics. Bioessays 22:148–160.

SHIMAMURA, M., H. YASUE, K. OHSHIMA, H. ABE, H. KATO, T. KISHIRO, M. GOTO, I. MUNECHIKA, AND N. OKADA. 1997. Molecular evidence from retroposons that whales form a clade within even-toed ungulates. Nature 388:666–670.

SHIMAMURA, M., H. ABE, M. NIKAIDO, K. OHSHIMA, AND N. OKADA. 1999. Genealogy of families of SINEs in cetaceans and artiodactyls: the presence of a huge superfamily of $tRNA^{Glu}$-derived families of SINEs. Mol. Biol. Evol. 16:1046–1060.

SINGER, M., AND P. BERG. 1991. Genes and genomes. Univ. Science Books, Mill Valley.
SMIT, A. F. A., AND A. D. RIGGS. 1995. MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. Nucleic Acids Res. 23:98–102.
SWOFFORD, D. 1996. PAUP*: Phylogenetic analysis using parsimony (and other methods), version 4.0. Sinauer Associates, Sunderland, Massachusetts.
TAKAHATA, N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics 122:957–966.
TERAI, Y., K. TAKAHASHI, AND N. OKADA. 1998. SINE cousins: the 3′ end tails of the two oldest and distantly related families of SINEs are descended from the 3′ ends of LINEs with the same geneological origin. Mol. Biol. Evol. 15:1460–1471.
THEWISSEN, J. G. M., AND S. I. MADAR. 1999. Ankle morphology of the earliest cetaceans and its implications for the phylogenetic relations among ungulates. Syst. Biol. 48:21–30.
THEWISSEN, J. G. M., S. I. MADAR, AND S. T. HUSSAIN. 1998. Whale ankles and evolutionary relationships. Nature 395:452.
WEINER, A., P. L. DEININGER, AND A. EFSTRATIADIS. 1986. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu. Rev. Biochem. 55:631–661.
WU, C.-I. 1991. Inferences of species in relation to segregation of ancient polymorphism. Genetics 127:429–435.

Received 27 July 1999; accepted 4 December 1999
Associate Editor: R. Olmstead

Syst. Biol. 49(4):817–829, 2000

Parametric Phylogenetics?

MICHAEL J. SANDERSON[1] AND JUNHYONG KIM[2]

[1]*Section of Evolution and Ecology, University of California, Davis, California 95616, USA; E-mail: mjsanderson@ucdavis.edu*
[2]*Departments of Ecology and Evolutionary Biology; Molecular, Cellular, and Developmental Biology; and Statistics, Yale University, New Haven, Connecticut 06511, USA; E-mail: junhyong.kim@yale.edu*

The number of ways to infer phylogenies is large (Swofford et al., 1996) and getting larger (e.g., Larget and Simon, 1999). The proliferation of techniques poses a challenge to phylogeneticists concerned with justification of these techniques and their relative performance. Historically, debates over methodology have centered on philosophical issues (Farris, 1983; Felsenstein, 1983; Sober, 1988; Siddall and Kluge, 1997), many of which involve unprovable assertions about the proper way to perform inductive inference in science (which is better: Ockham's razor, or Fisher's "likelihood principle"? Sober, 1988; Edwards, 1992; Royall, 1997). In the last decade much of the debate has shifted over to performance evaluations based on computer simulations (reviewed in Hillis et al., 1994; Li, 1997) and studies of "known phylogenies" (Russo et al., 1996; Naylor and Brown, 1998; Leitner and Fitch, 1999). However, little consensus has emerged, except that a few methods that are not widely used anyway, such as UPGMA, perform poorly.

One way to classify these methods is by the degree to which they rely on a parametric model of the evolutionary process to estimate the tree. A parametric estimation method has an explicit deductive relationship to a family of sampling distributions characterized by one or more parameters. Parametric methods such as maximum likelihood (ML) have received much attention in phylogenetics recently because of extensions of models of molecular evolution in new directions through the addition of more parameters (e.g., Tillier and Collins, 1995; Goldman et al., 1998; Thorne et al., 1998; Huelsenbeck and Nielsen, 1999). Computational and algorithmic advances have contributed to this interest (Rogers and Swofford, 1998; Schadt et al., 1998), and the publication of PAUP* 4.0 (Swofford, 1999) presents systematists with an elegant, user-friendly platform to use parametric methods. This may be an appropriate time to take a critical look at likelihood-based methods in phylogenetic estimation as one end of an important methodological spectrum.

Much has been written about the relative strengths and weaknesses of model-based methods versus "model-free" methods such as parsimony. However, most of this has been concerned with philosophical rather than statistical arguments (e.g., Farris, 1983; Sober, 1988). Here we raise explicitly statistical

2001. Origin of whales from early artiodactyls: hands and feet of Eocene Protocetidae from Pakistan. Science 293:2239-2242. Goodman, M., J. Czelusniak, And J. E. Beeber.Â Palaeontologische Zeischrift 42: 83-104. Shedlock A. M., M.C. Milinkovitch, and N. Okada. 2000. SINE Evolution, Missing Data, and the Origin of Whales. Syst. Biol. Our data suggest that survey sequencing and genome mining are valuable tools to investigate SINE evolution among related lineages and can provide substantial information about the ability of SINEs to proliferate in diverse genomes. This method would also be a useful first step in determining which subfamilies would be the best to target when developing SINEs as markers for phylogenetic and population genetic analyses.Â Shedlock AM, Milinkovitch MC, Okada N. SINE evolution, missing data, and the origin of whales. Syst Biol. 2000;49(4):808â€“17.