Points of View

Syst. Biol. 50(3):438–444, 2001

Popper and Likelihood Versus "Popper*"

JAMES S. FARRIS,[1] ARNOLD G. KLUGE,[2] AND JAMES M. CARPENTER[3]

[1]*Molekylärsytematiska laboratoriet, Naturhistoriska riksmuseet, Box 50007 SE-104 05 Stockholm, Sweden*
[2]*Division of Reptiles and Amphibians, Museum of Zoology, The University of Michigan, Ann Arbor, Michigan 48109, USA; E-mail: akluge@umich.edu (author for correspondence)*
[3]*Department of Invertebrates, American Museum of Natural History, New York, New York 10024, USA*

Faith (1999) regarded Kluge's (1997) and Carpenter et al.'s (1998) discussions of corroboration as examples of "sloganized Popper." That was as distinguished from "Popper*," which included (Faith, 1999:676; as throughout, italics are as in the original):

> ...the explicit Popperian severity/corroboration framework underlying the permutation tail probability test (PTP) and related tests. The stated degree of corroboration/severity for a PTP test of a phylogenetic hypothesis is equated with the degree of improbability, equal to the *P* value in evaluating a null model based on random character covariation...

PTP is just Faith's name for the permutation test that Archie introduced in 1985 (see Legendre, 1986:137; cf. Archie, 1989). Now seldom used, the test does not even provide a reliable indication of phylogenetic structure in data (Källersjö et al., 1992; Farris et al., 1994; Carpenter et al., 1998), and interpreting PTP as corroboration leads to multiple contradictions (Farris, 1995). Faith did not discuss any of those problems, however. Instead, he created a new one (Faith, 1999:678):

> Whereas a sloganized Popper has provided an exclusive philosophy, twisting and turning to uniquely justify cladistic parsimony, Popper* is relevant to phylogenetic analyses using parsimony and other methods.

That is not correct, for what Faith presented as a criticism of "sloganized Popper" is actually an objection to likelihood. The premises of Faith's "Popper*" are inconsistent with a likelihood approach, as we will show here by analyzing the connection between likelihood and corroboration. This cannot be blamed on Popper (1968, 1972, 1992), however, because each point of conflict between "Popper*" and likelihood corresponds to a difference between "Popper*" and Popper.

LIKELIHOOD

We begin by examining the relationship between likelihood and Popper's formulae for severity (strength of evidence) and corroboration. Popper (1972:391; as throughout, italics are as in the original) defined:

> ...the severity of the test e interpreted as *supporting evidence* of the theory h, given the background knowledge b [as]:

$$S(e,h,b) = ((p(e,hb) - p(e,b))/(p(e,hb) + p(e,b))$$

Popper (1968:400f) also called this quantity explanatory power E. Here $p(e,b)$ denotes the probability of e given b, while $p(e, hb)$ is the probability of e given both h and b. For a phylogenetic likelihood method, h would be a postulated phylogenetic tree, while e would be a matrix of character (sequence) data. Background knowledge b includes accepted (well-corroborated) theories that help guide the interpretation of e as evidence on h. Kluge (1997), for example, included descent with modification in background knowledge. In likelihood methods the stochastic evolutionary model is included in b (see Kluge, 1997).

Popper's (1992:240; cf. Popper, 1972:288, 1968:400f) formula for corroboration is almost the same, differing only in having an

additional term $p(eh,b)$ in the denominator:

$$C(h,e,b) = \frac{p(e,hb) - p(e,b)}{p(e,hb) - p(eh,b) + p(e,b)}$$

The difference can be ignored for present purposes, because $p(eh, b)$ is close to 0 for hypotheses with high content (Popper, 1968:401), a category that would surely include phylogenetic hypotheses. This is just as well, since Faith has never mentioned $p(eh,b)$.

The connection between C (or S, or E) and likelihood becomes clear on considering which tree h would be best corroborated on the basis of currently available information. For given e and b, $p(e,b)$ would have some set value, so that corroboration $C(h,e,b)$ for different trees h would vary just according to the other term $p(e,hb)$, and as is easily seen from the formula, the trees with greatest $p(e,hb)$ would have strongest corroboration C. But those same trees would also be the maximum likelihood trees, for as Popper (1968:410) pointed out, $p(e,hb)$ is the likelihood of h given evidence e and background knowledge b.

Likelihood does not always correspond so directly to corroboration. As Popper (1968) emphasized, any evidence e with high $p(e,hb)$ automatically makes likelihood high, whereas corroboration is strong only for critical evidence (severe tests), that is, when $p(e,hb)$ is much larger than $p(e,b)$. But this consideration arises only when assessments based on different evidence or backgrounds are compared. For given evidence and background, the hypotheses h with the greatest likelihood $p(e,hb)$ also have the strongest corroboration $C(h,e,b)$.

$p(e,hb)$ is thus an important part of Popper's corroboration, but one would not know that from Faith's discussion. Faith mentioned the likelihood $p(e,hb)$ only in passing, and then only to suggest that it can be effectively ignored, because it is merely set to 1 (Faith, 1999:678; notice that Faith always dropped the italics from Popper's p):

> That [idea of Mayo's (1996)] can be expressed as a high $p(e,hb)$ [*sic*]—the probability of e given h and b—a likelihood term also used in Popper's equations for severity corroboration (but usually set equal to 1).

That would be disastrous for maximum likelihood. Likelihood can be used to select among trees h only if the likelihood $p(e,hb)$ varies among trees, yet Faith treated $p(e,hb)$ as if it were fixed. Faith's "Popper*" would prevent application of the likelihood principle.

But that difficulty comes just from "Popper*," not Popper. Popper had no intention of keeping $p(e,hb)$ fixed at 1, but meant $p(e,hb)$ to vary. To be sure, $p(e,hb)$ may be high for evidence e that strongly favors hypothesis h (Popper, 1992:238):

> This leads us at once to realize that the smaller $p(e,b)$, the stronger will be the support which e renders to h—provided our first demand is satisfied, that is, provided e follows from h and b, or from h in the presence of b.

Under that provision, $p(e,hb)$ would be 1, but $p(e,hb)$ is instead low for evidence that undercuts the hypothesis, as Popper (1992:242) went on to emphasize:

> [But] what about an empirical evidence e which falsifies h in the presence of b? Such an e will make $p(e,hb)$ equal to zero.

"Popper*" seems to be based on simply ignoring the latter possibility, for Faith (1999:677) quoted the first of those passages himself, yet never mentioned the second, although it is part of the same discussion.

Point Probabilities

Rather than being based on Popper's discussion, treating the likelihood $p(e,hb)$ as if it were fixed was required by another part of Faith's own position. Faith identified PTP—the significance level or tail probability from the PTP test—with Popper's $p(e,b)$, so that he had to treat $p(e,hb)$ as fixed to conclude that PTP would determine corroboration.

Of course that provides no legitimate grounds for fixing $p(e,hb)$, but further, Faith's identification of PTP with $p(e,b)$ leads in itself to another incompatibility between "Popper*" and likelihood methods. Because PTP is a cumulative probability, equating PTP with $p(e,b)$ would mean that $p(e,b)$ must also be a cumulative probability; if so, then $p(e,hb)$ would have to be a cumulative probability as well, because $p(e,hb)$ differs from $p(e,b)$ only in the added condition h. Yet $p(e,hb)$ cannot be a cumulative probability, for it is a likelihood. Maximum likelihood estimation procedures always maximize point probabilities or densities (see Lindgren, 1962), not cumulative probabilities.

Again, the conflict with likelihood comes from 'Popper*,'' not Popper, for Popper used point probabilities. This is seen, for example, in his discussion of statistical hypotheses. Unfortunately, Popper did not use quite the same notation here as in the formulae seen earlier. Probability is written P, and b indicates a population, not background knowledge, the latter not being explicitly denoted. Thus $P(a,b)$ means the probability of property a in population b, while $P(e)$ corresponds to the $p(e,b)$ of the formulae above, and $P(e,h)$ corresponds to (e,hb). With all this in mind, the expression $P(e,h) - P(e)$, connected with C and E near the end of the passage, is readily recognized as the numerator of C and $S(= E)$ in the formulae first quoted. Popper (1968:410f) commented:

> Now let h be the statement $P(a,b) = r$ and let e be the statement 'In a sample which has size n and which satisfies the condition b (or which is taken at random from the population b), a is satisfied in $n(r \pm \delta)$ of the instances'. Then we may put, especially for small values of δ, $P(e) \cong 2\delta$. We may even put $P(e) = 2\delta$; for this would mean that we assign equal probabilities—and therefore, the probabilities $1/(n+1)$—to each of the $n+1$ proportions, $0/n,\ 1/n, \ldots n/n$, with which a property a may occur in a sample of size n ... (The equidistribution here described is ... adequate for assessing the absolute probability, $P(e)$, if e is a statistical report about a sample. But ... for assessing relative probability $P(e,h)$ of the same report ... in this case, it is adequate to assume a combinatoric, i.e., a Bernoullian rather than a Laplacean distribution.) ... We therefore find that $P(e,h) - P(e)$, and thus our functions E and C, can only be large if δ is small and n large; or in other words if e is *a statistical report asserting a good fit in a large sample*.

$P(e,h)$ is thus the Bernoullian (binomial) probability of obtaining so many a's with sample size n and parametric frequency r. That is a point, not a cumulative, probability and the same is obviously true of the discrete uniform distribution $P(e) = 2\delta = 1/(n+1)$. Faith's use of cumulative probabilities is not based on Popper's ideas, but only on Faith's own.

FIT

A further conflict between "Popper*" and likelihood arises from the seemingly innocent fact that the PTP test uses the length of the most-parsimonious tree or trees for the data as a test statistic. The cumulative probability PTP, which Faith identified with Popper's $p(e,b)$, is obtained from a null distribution of such lengths. Faith regarded that length as a measure of the fit of the data to the tree, and accordingly he maintained that e should refer to fit, not data (Faith, 1999:676):

> In the introductory quote from Mayo [(1996)] above, note that e refers to the data, whereas Popper uses e for the evidence, corresponding to what Mayo refers to as "fit" of data to hypothesis.

On Faith's view, then, $p(e,b)$ and $p(e,hb)$ would be not just cumulative probabilities but cumulative probabilities from distributions on degrees of fit. But in that case $p(e,hb)$ would not be suitable for a likelihood method, because estimating maximum likelihood works by choosing the hypothesis to maximize the probability of the *data* given the hypothesis.

Once more the difficulty comes from "Popper*," not Popper, although this time it is not hard to see how a superficial reading could have led Faith to his position. In the discussion of property a quoted above, for example, Popper (1968:411) noted that strong corroboration can be achieved only "if e is a *statistical report asserting a good fit in a large sample*." That might seem to suggest that $P(e)$ and $P(e,h)$ are distributions on some measure of fit, but this possible confusion disappears when the example is considered further. Popper's $P(e)$ and $P(e,h)$ are simply distributions on the number of a's in a sample of n independent observations, not on any variable that could be regarded as a measure of fit. Identifying evidence with fit, in fact, directly violates a rule that Popper (1972:288) discussed while emphasizing the importance of avoiding ad hoc hypotheses:

> My [formula for corroboration] does not automatically exclude ad hoc hypotheses, but it can be shown to give most reasonable results if combined with a rule excluding ad hoc hypotheses ... [This rule] may take the following form: the hypothesis must not repeat ... the evidence or any conjunctive component of it.

If the "evidence" were fit, and so calculated from the hypothesis as well as the data, then the hypothesis would ipso facto repeat a conjunctive component of the "evidence," namely, itself. That is just the situation that must be avoided if "most reasonable results" are to be obtained.

How unreasonable results could become if evidence were identified with fit can best be appreciated from some examples. For brevity, use x to denote the number of a's observed in a random sample of n independent observations. Let the hypothesis in

question be that $r = 1/2$, and suppose that $n = 1{,}000$. The observed count of a's that would most strongly favor the hypothesis is $x = 500$, for which the binomial probability $P(e,h)$ is ~0.025225. For $n = 1{,}000$, Popper's $P(e) = 1/(n+1)$ is 1/1,001. In present notation, Popper's (1968:400f) S (there called E) is just

$$S(e,h) = (P(e,h) - P(e))/(P(e,h) + P(e))$$

which in this case is

$$(0.025225 - 1/1{,}001)/(0.025225 + 1/1{,}001) = 0.9238.$$

Positive S indicates that the evidence favors the hypothesis (Popper, 1968:400f; cf. Popper, 1992:241), quite strongly in this example, given that the upper bound of S is +1. S should approach its upper bound when the evidence is ideally favorable and extensive, and in this case keeping $x = n/2$ as n is increased causes S to approach +1 in the limit, just as one would like.

Suppose, on the other hand, that $x = n = 1$. Now $P(e) = 1/(n+1) = 1/2$, while binomial $P(e,h)$ is also $1/2$, so that $S = (1/2 - 1/2)/(1/2 + 1/2) = 0$. $S = 0$ indicates irrelevance of the evidence to the hypothesis (Popper, 1968:400f), and this is obviously correct. $x = n = 1$ is not enough data to provide grounds for evaluating the hypothesis that $r = 1/2$.

Finally, return to $n = 1{,}000$ and suppose that $x = 1{,}000$, a count most unfavorable to the hypothesis that $r = 1/2$. $P(e)$, which depends only on n, is 1/1,001, but the binomial probability $P(e,h)$ is now $2^{-1,000}$ ($\sim 9.33 \times 10^{-302}$), so that S is

$$(2^{-1,000} - 1/1{,}001)/(2^{-1,000} + 1/1{,}001) \cong -1.$$

Negative S indicates that the evidence refutes the hypothesis, emphatically in this case, because −1 is the lower bound of S (Popper, 1968:400f).

That is also just as one would like, because with those data a two-tailed exact binomial test—which is a likelihood ratio test—would reject the hypothesis that $r = 1/2$ at a significance level of $\sim 1.9 \times 10^{-301}$. This satisfactory behavior of S, however, depends on using the probabilities that Popper intended. According to Faith, one should instead use cumulative probabilities of fit, and this leads to quite different results.

Write f for a measure of fit of the count x to the hypothesis h, and $K(f)$ and $K(f,h)$ for cumulative probability distributions on f. With K used in place of P, "severity" would be

$$S_K(f,h) = (K(f,h) - K(f))/(K(f,h) + K(f)).$$

Note that what Faith called "fit" is actually the opposite: Small values mean that the data conform to the hypothesis. That being understood, for the hypothesis that $r = 1/2$ and sample size $n = 1{,}000$, f should reach its maximum (worst) possible value f^* when $x = 1{,}000$. Because f^* is the maximum of f, the cumulative probabilities $K(f^*)$ and $K(f^*, h)$ must be unity, so that

$$S_K(f^*, h) = (1 - 1)/(1 + 1) = 0.$$

$S = 0$, as seen earlier, indicates that the evidence is irrelevant to the hypothesis, so that misinterpreting S_K as Popper's S makes the observation $x = n = 1{,}000$ seem irrelevant to evaluating the hypothesis that $r = 1/2$. Using cumulative probabilities of fit instead of point probabilities of observations in "Popper's" formulae has the thoroughly unreasonable result of making it impossible to identify even very strong evidence as unfavorable to a hypothesis.

Dependence

Faith's objection to "sloganized Popper" was based on his idea that "evidence" meant fit. Because in that case e would depend on h, he supposed $p(e,b)$ would also depend on h. Accordingly, he criticized other authors for taking "only properties of the data" as evidence and for failing to treat $p(e,b)$ as if it depended on h (Faith, 1999:678):

> ...recastings of Popperian corroboration, while explicitly considering the terms, $p(e,hb)$ [*sic*] and $p(e,b)$ [*sic*], actually assign $p(e,b)$ [*sic*] no role. For example,
>
> > the best supported hypotheses are those that assign highest probability to the evidence. Only $p(e,hb)$ [*sic*] can perform this role; the other term $p(e,b)$ [*sic*] does not involve h. (Carpenter et al., 1998:107)

> Similarly... [in Kluge (1997)] "evidence," e reflects only the properties of the data itself, implying that all tree hypotheses have the same value for $p(e,b)$ [*sic*]. So this term again plays no role in determining relative corroboration/severity for different tree hypotheses.

The role of $p(e,b)$, as Kluge (1997) stressed, and as was pointed out earlier, is to distinguish critical from non-critical evidence. But it is the likelihood $p(e,hb)$—not $p(e,b)$—that determines the relative corroboration of different trees h for the *same* evidence and background. Faith's complaint is actually an objection to likelihood, as becomes apparent when we quote Carpenter et al.'s (1998:107; underlining added) comments a little more fully than Faith did:

> [Faith] missed the point of Popper's formula. The function of the term $p(e,hb)$ is to relate the hypothesis h to the evidence e. As with the likelihood principle, the best-supported hypotheses are those that assign highest probability to the evidence. Only $p(e,hb)$ can perform this role; the other term $p(e,b)$ does not involve h. By pretending that $p(e,hb)$ could be frozen at unity "so that the first term can be ignored," Faith (1992:266) arrived at a formulation that would absurdly make the "corroboration of h" independent of h.

If, as Faith maintained, $p(e,b)$—not $p(e,hb)$—varied among trees h, maximum likelihood estimation would be exactly the wrong approach. Maximum likelihood uses only the likelihood $p(e,hb)$, ignoring any information on h contained in $p(e,b)$. But in fact there is no such information, because Popper's $p(e,b)$ does not depend on h. This is plain from Popper's property a example, discussed earlier. There, Faith's claim would mean that $P(e)$ would depend on h, whereas in fact Popper's $P(e) = 1/(n+1)$ depends only on the sample size n, not on the hypothesis, which is the value of r. This supposed fault of "sloganized Popper" is just another case in which conflict between likelihood and "Popper*" reflects a difference between "Popper*" and Popper.

RELEVANCE

An even worse difficulty stems from the permutation (randomization) null model Y used in the PTP test. Under that model, characters are distributed randomly among taxa and are completely independent of the phylogeny. To identify PTP with Popper's $p(e,b)$, Faith had to identify the background knowledge b with Y, and this choice of background is not beneficial to likelihood methods.

To see why that is so, consider the likelihood $p(e,hY)$ for data e and some given tree h with Y used in place of the background b. According to Y, e is independent of h. In that case, making the probability conditional on h would not change the distribution of e, so that $p(e,h\ Y) = p(e,Y)$. But the same would be true for any tree, so that all trees would have the same likelihood, and the data would never provide grounds for choosing one tree over another. Similarly,

$$\begin{aligned} S(e,h,Y) &= (p(e,hY) - p(e,Y))/(p(e,hY) \\ &\quad + p(e,Y)) = (p(e,Y) - p(e,Y))/2p(e,Y) \\ &= 0 \end{aligned}$$

would hold for any tree. Faith's choice of background would make the data—*any* data—irrelevant to inferring phylogeny.

To be sure, Faith meant to use fit in place of e, but that would do no good. If f is a measure of fit of data e to a given tree hypothesis h, f is a set function of the data, so that $p(f,h\ Y)$ and $p(f,Y)$ can naturally be computed from the corresponding distributions $p(e,h\ Y)$ and $p(e,Y)$ on the data. Then, because $p(e,h\ Y)$ is the same distribution as $p(e,Y)$, $p(f,h\ Y) = p(f,Y)$, in which case the corresponding cumulative distributions are obviously the same as well: $K(f,hY) = K(f,Y)$. Consequently, using cumulative probabilities of fit, as Faith would like, the "corroboration" of h would be

$$\begin{aligned} S_K(f,h,Y) &= (K(f,hY) - K(f,Y))/(K(f,h\ Y) \\ &\quad + K(f,Y)) = (K(f,Y) - K(f,Y))/ \\ &\quad 2K(f,Y) = 0 \end{aligned}$$

for *any* tree h and any data. With Y used in place of b, there would never be grounds for preferring one tree over another.

Faith might not calculate $S_K(f,h,Y)$ that way. Because he treated Popper's $p(e,hb)$ as if it were fixed at 1, he would presumably treat $K(f,h\ Y)$ likewise, and this would mask the symptom that $S_K(f,h,Y) \equiv 0$. But there is no justification for treating $K(f,h\ Y)$ as if it were fixed at 1, and in any case, shifting the value of $S_K(f,h,Y)$ would not dispose of the underlying fault of Faith's position, that

with Y used in place of b, the data convey no information about h.

Like the other aspects of "Popper*" that conflict with likelihood, the choice of Y as "background" is Faith's, not Popper's, but this point requires no further attention here, for it has been discussed in detail by Farris (1995, 2000). The behavior of $S_K(f,h,Y)$ is of further interest, however, because it provides an informative perspective on Faith's (1999:678) concluding remarks:

> Whereas a sloganized Popper has provided an exclusive philosophy, twisting and turning to uniquely justify cladistic parsimony, Popper* is relevant to phylogenetic analyses using parsimony and other methods.

By that he meant, in present notation, that fit criteria used in other methods—clique size, cophenetic correlation, even the likelihood score under (say) a Jukes–Cantor model—could be used as the f in $K(f,Y)$. Faith believed that such calculations would provide a useful assessment of corroboration, but in that, his reasoning consisted simply of ignoring difficulties. He departed from Popper both in using probabilities of fit and in using cumulative probabilities, but even if those discrepancies could be overlooked, the conclusion that $S_K(f,h,Y) \equiv 0$ for any tree and data holds for any of those fit measures. If Y were used in place of b, no tree could ever be "corroborated," no matter what method or data used.

Parsimony

Faith's (1999:678) concluding remarks also involved twisting and turning. Although he did not expand on this, he seems to have felt that there was some difficulty in relating parsimony to corroboration. Apparently he was unaware of Tuffley and Steel's (1997:599) Theorem 5:

> Theorem 5. *Maximum parsimony and maximum likelihood with no common mechanism are equivalent in the sense that both choose the same tree or trees.*

Because the maximum likelihood trees are also the best-corroborated trees for the given evidence and background, one can immediately see that parsimony maximizes corroboration when no common mechanism is included in the background knowledge. Inasmuch as Popper's explanatory power E has the same formula as S, this also means that the most-parsimonious trees have the greatest explanatory power, in agreement with Farris' (1983) conclusion that most-parsimonious trees can best explain observed similarities as the result of inheritance and common ancestry.

Of course any maximum likehood method would maximize corroboration for the given evidence, *if* the stochastic model used were included in the background knowledge. But only well-corroborated theories can legitimately be included in the background, and the importance of Tuffley and Steel's result is that it relies on a defensible model. Nearly all the systematic methods now called maximum likelihood (for a review see Siddall and Whiting, 1999) are based on restrictive homogeneity assumptions: that all sites have substitution rates drawn from the same distribution, and that the ratio in rates between any two sites is the same in all parts of the tree (see Farris, 1999). Because those assumptions are known to be unrealistic, they are definitely not well-corroborated theories. The no common mechanism avoids that difficulty because it allows (but does not require) sites to vary independently in their rates.

Parsimony thus has a clear and eminently useful connection to corroboration. In that, if differs markedly from "Popper*."

References

Archie, J. W. 1989. A randomization test for phylogenetic information in systematic data. Syst. Zool. 38:219–252.

Carpenter, J. M., P. A. Goloboff, and J. S. Farris. 1998. PTP is meaningless, T-PTP is contradictory: A reply to Trueman. Cladistics 14:105–116.

Faith, D. P. 1999. Error and the growth of experimental knowledge. Syst. Biol. 48:675–679.

Farris, J. S. 1983. The logical basis of phylogenetic analysis. Pages 7–36 *in* Advances in cladistics II (N. I. Platnick and V. A. Funk, eds.). Columbia Univ. Press, New York.

Farris, J. S. 1995. Conjectures and refutations. Cladistics 11:105–118.

Farris, J. S. 2000. Corroboration versus "strongest evidence." Cladistics 16:385–393.

Farris, J. S., M. Källersjö, A. G. Kluge, and C. Bult. 1994. Permutations. Cladistics 10:65–76.

Källersjö, M., J. S. Farris, A. G. Kluge, and C. Bult. 1992. Skewness and permutation. Cladistics 8:275–287.

Kluge, A. G. 1997. Testability and the refutation and corroboration of cladistic hypotheses. Cladistics 13: 81–96.

Legendre, P. 1986. Report on nineteenth international numerical taxonomy conference. Syst. Zool. 35:135–139.

Lindgren, B. W. 1962. Statistical theory. Macmillan, New York.

MAYO, D. G. 1996. Error and the growth of experimental knowledge. Univ. of Chicago Press, Chicago.
POPPER, K. 1968. The logic of scientific discovery. Harper and Row, New York.
POPPER, K. 1972. Conjectures and refutations: The growth of scientific knowledge. Routledge and Keegan Paul, New York.
POPPER, K. 1992. Realism and the aim of science. Routledge, London.
SIDDALL, M. E., AND M. F. WHITING. 1999. Long branch abstractions. Cladistics 15:9–24.

Received 19 October 2000; accepeted 28 December 2000
Associate Editor: R. Olmstead

Syst. Biol. 50(3):444–453, 2001

Are the Fossil Data Really at Odds with the Molecular Data? Morphological Evidence for Cetartiodactyla Phylogeny Reexamined

GAVIN J. P. NAYLOR AND DEAN C. ADAMS
Department of Zoology and Genetics, Iowa State University, Ames, Iowa 50010, USA; E-mail: gnaylor@iastate.edu

The phylogenetic position of Cetacea within the mammalian tree has long been a subject of debate. The traditional paleontological view is that an extinct order of mammals, the Mesonychia, is the sister taxon to Cetacea (e.g., Van Valen, 1966; Prothero et al., 1988). This view has recently been supported by morphological studies that examined both fossil and extant material (Geisler and Luo, 1998; O'Leary and Geisler, 1999). The molecular evidence, by contrast, supports a phylogenetic hypothesis in which Cetacea is nested deeply within the Artiodactyla, implying that Artiodactyla is paraphyletic with respect to Cetacea (Sarich, 1985; Milinkovitch et al., 1993; Gatesy et al., 1999, and references therein). Furthermore, several molecular studies have suggested that hippopotamids are the sister taxon to Cetacea (e.g., Irwin and Arnason, 1994; Gatesy et al., 1996; Gatesy, 1997, 1998; Montgelard et al., 1997; Nikaido et al., 1999). Although the "return to water" aspect of this phylogenetic hypothesis has a certain intuitive appeal, it has met with resistance from those who work primarily with morphology (e.g., Geisler and Luo, 1998; O'Leary and Geisler, 1999; O'Leary, 1999). Despite the resurgent interest in the problem, no consensus reconciling the different signals has yet been reached.

Obviously, a serious limitation of molecular data is that this information cannot be gathered from the fossil remains of extinct taxa. In contrast, fossils can play an important role in recovering phylogeny in morphological studies, often providing information about character states of stem lineages that are not present in extant taxa. In some cases, the inclusion of fossils can even overturn inferences based solely on extant character distributions (e.g., Gauthier et al., 1988; Eernisse and Kluge, 1993). Because molecular data sets, by their very nature, can include only extant taxa, biased taxon sampling could possibly lead to incorrect phylogenetic inferences.

Recently, O'Leary and Geisler (1999) presented a phylogenetic analysis of morphological data from both fossil and extant mammals. Using a combined data set of characters from basicranial, cranial, dental, postcranial, and soft tissue regions, they found evidence supporting the monophyly of Cetacea, Mesonychia, Artiodactyla, and Perissodactyla. Their data also supported a sister-group relationship between Mesonychia and Cetacea, affirming the traditional paleontological view (Fig. 1). Interestingly, when fossils were excluded from the data, O'Leary and Geisler found that the phylogenetic signal present in the extant taxa was similar to that seen for the molecular data, that is, with Cetacea deeply nested within Artiodactyla. This pattern lead them to propose that phylogenies based on extant data alone might be predisposed to recover inconsistent branches as a result of sparse taxon sampling. If true, this could explain why phylogenetic inferences based on molecular data (derived exclusively from extant taxa) are frequently at odds with inferences from skeletal and dental data derived from both fossil and extant forms. This is an interesting hypothesis and one that, if borne out, would have sobering implications for molecular systematics—a discipline almost