Protein sequence signatures support the African clade of mammals

Marjon A. M. van Dijk*, Ole Madsen*, François Catzeflis†, Michael J. Stanhope‡, Wilfried W. de Jong*§¶, and Mark Pagel¶‖

*Department of Biochemistry, University of Nijmegen, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands; †Institut des Sciences de l'Évolution, Université Montpellier 2, 34095 Montpellier, France; ‡Queen's University of Belfast, Biology and Biochemistry, Belfast BT9 7BL, United Kingdom; §Institute for Systematics and Population Biology, University of Amsterdam, 1090 GT Amsterdam, The Netherlands; and ‖School of Animal and Microbial Sciences, University of Reading, Whiteknights, Reading RG6 6AJ, United Kingdom

Edited by Elwyn L. Simons, Duke University Primate Center, Durham, NC, and approved October 9, 2000 (received for review May 12, 2000)

DNA sequence evidence supports a superordinal clade of mammals that comprises elephants, sea cows, hyraxes, aardvarks, elephant shrews, golden moles, and tenrecs, which all have their origins in Africa, and therefore are dubbed Afrotheria. Morphologically, this appears an unlikely assemblage, which challenges—by including golden moles and tenrecs—the monophyly of the order Lipotyphla (Insectivora). We here identify in three proteins unique combinations of apomorphous amino acid replacements that support this clade. The statistical support for such ''sequence signatures'' as unambiguous synapomorphic evidence for the naturalness of the Afrotherian clade is reported. Using likelihood, combinatorial, and Bayesian methods we show that the posterior probability of the mammalian tree containing the Afrotherian clade is effectively 1.0, based on conservative assumptions. Presenting sequence data for another African insectivore, the otter shrew *Micropotamogale lamottei*, we demonstrate that such signatures are diagnostic for including newly investigated species in the Afrotheria. Sequence signatures provide ''protein-morphological'' synapomorphies that may aid in visualizing monophyletic groupings.

Molecular sequence data are increasingly used in mammalian phylogeny and recently have led to a number of unorthodox proposals (1–3). These proposals range from the claim that the guinea pig is not a rodent (4) to making whales and hippos sister groups (5). One of the most remarkable propositions is that of an ''African clade'' in which species as diverse as elephant shrews (Macroscelidea), golden moles (Chrysochloridae), and tenrecs (Tenrecidae) are grouped with aardvarks (Tubulidentata) and paenungulates (elephants, sea cows, and hyraxes; refs. 6 and 7). All of the African clade species find their fossil roots in Africa, and most are still confined to this continent, hence the name Afrotheria (7). The sequence evidence for Afrotheria is unanimous and strong, deriving from various nuclear and mitochondrial genes (6–10). Morphologically, however, there is no evidence whatsoever for a natural grouping of these taxa (11–14), prompting us to subject the molecular evidence to further scrutiny.

If Afrotheria is a real clade, it might be possible to find specific combinations of amino acid replacements in the proteins that support them. These replacements would represent synapomorphous character states, as remnants of mutational events during the last common ancestry of a clade. Several authors have used the concept of such "sequence signatures" qualitatively in molecular phylogeny (e.g., refs. 15–19), but thorough statistical interpretations are lacking.

We here search for the presence of unique Afrotherian sequence signatures in nine protein data sets—eight nuclear and one mitochondrial—that include at least four Afrotherian orders. Putative Afrotherian signatures were traced in αA-crystallin (CRYAA), aquaporin-2 (AQP2), and interphotoreceptor retinol-binding protein (IRBP). To demonstrate the diagnostic value of the signatures we seek their presence in CRYAA and AQP2 of other potential members of the African clade, including the otter shrew—representing the only tenrecid subfamily living outside of Madagascar. To assess the significance of the candidate signatures, we use likelihood methods (20) to reconstruct their most probable ancestral states at the basal node of the Afrotherian clade. These calculations use a phylogeny reconstructed independently of the protein under investigation. We further use likelihood and combinatorial methods to estimate the probability of the signatures on three alternative morphology-based trees that are incompatible with an African clade. We then combine the evidence from CRYAA, AQP2, and IRBP by using Bayesian techniques to yield a posterior probability for the Afrotherian clade. Demonstrating the statistical improbability of such events in the course of biological evolution (21) may help to escape from the current stalemating in the molecules-versus-morphology debate on vertebrate phylogeny (3).

Materials and Methods

Searching for Afrotherian Signatures. Databases were searched for sets of protein sequences that included representatives of at least four Afrotherian orders, i.e., Proboscidea (elephants), Sirenia (sea cows), Hyracoidea (hyraxes), Tubulidentata, Macroscelidea, and Afrosoricida (golden moles and tenrecs; ref. 7). This yielded data sets of CRYAA, AQP2, IRBP, von Willebrand factor, α-2B adrenergic receptor, γ-fibrinogen, hemoglobin-α and -β, and cytochrome *b*. The AQP2 data set was complemented with newly determined sequences of pig, fin whale, and sperm whale (see below). From these data sets, one or, if available, two representatives of all included eutherian orders were taken. When more than two species were available for an order, only the two most divergent sequences were retained. This increases the homoplastic background, and thus the significance of retrieved signatures. Retaining all sequences would make the taxon representation unbalanced and hamper the signature searches. The selected sequences were aligned, using PILEUP, and manually edited. Where available, two divergent Marsupialia were included as outgroups. For full species names and accession numbers, and for protein alignments of CRYAA, AQP2, and IRBP, see Table 4 and Figs. 3–5, which are published as supplemental data on the PNAS web site, www.pnas.org.

Candidate sequence signatures were retrieved from the align-

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: CRYAA, αA-crystallin; AQP2, aquaporin-2; IRBP, interphotoreceptor retinol-binding protein; ML, maximum likelihood.

See commentary on page 1.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AJ251100–AJ251106, AJ277647, and AJ270463–AJ270468).

¶To whom reprint requests should be addressed. E-mail: w.dejong@bioch.kun.nl or m.pagel@reading.ac.uk.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Article published online before print: *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.250216797. Article and publication date are at www.pnas.org/cgi/doi/10.1073/pnas.250216797

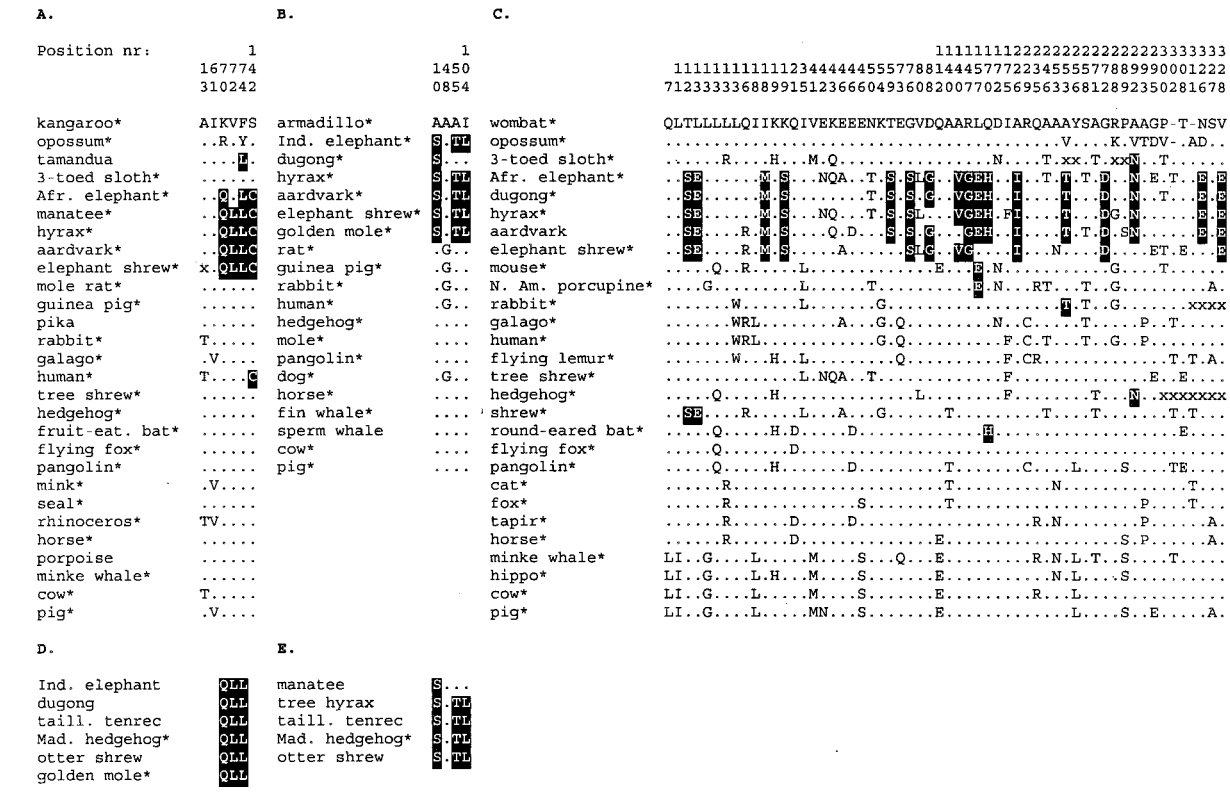

Fig. 1. Afrotherian signatures in CRYAA, AQP2, and IRBP. (*A*) Positions from an alignment of 26 eutherian and two marsupial CRYAA sequences at which the same putatively apomorphous replacement occurs in 4–6 eutherian sequences. In black are replacements that occur in at least four of the five Afrotheria. Note that ''.'' indicates all residues that did not pass the 4–6 search window; they may be identical to the two top out-group sequences or be apomorphies occurring in <4 or >6 sequences; x is unassigned residue; * denotes species that are included in the trees in Fig. 2. (*B*) Apomorphous replacements occurring in 5–7 AQP2 sequences, considering armadillo as out-group for the other Eutheria; rodents as out-group yields the same signature (shown in Fig. 4). (*C*) Positions in IRBP that pass the search for four- to six-species eutherian clades. Note that at certain positions different putatively apomorphous replacements fulfil the search criterium and may set apart different clades (e.g., 13E for Afrotheria, 13G for Cetartiodactyla). (*D* and *E*) Afrotherian signature positions in newly determined CRYAA and AQP2 sequences, respectively.

ments by using the spreadsheet SIGNWIN (available from the authors). No phylogenetic information is included in this search; SIGNWIN solely selects positions at which a designated number of in-group species have the same putatively apomorphous replacement, considering the out-group residue(s) as plesiomorphous condition. The selection window is set to be appropriate for the number of species for which the monophyly is investigated. Thus, when searching for positions that might support the monophyly of the five Afrotheria among the 26 selected eutherian CRYAA sequences (Figs. 1*A* and 3), the window is set at 5 ± 1. This allows for 20% back or otherwise superimposed replacements within a five-species clade, and the same absolute number of parallel replacements in the other 21 in-group sequences. Positions at which 4–6 species share the same apomorphy are then candidates for any Afrotherian sequence signature. Using a wider or narrower criterion would change our candidate signatures, but as seen in *Results* the candidate sites for a potential signature emerge clearly from the data.

Sequence Determination of CRYAA and AQP2. CRYAA genomic sequences, coding for amino acid residues 64–94, were determined for Indian elephant (*Elephas maximus*), dugong (*Dugong dugon*), tail-less tenrec (*Tenrec ecaudatus*), small Madagascar hedgehog (*Echinops telfairi*), otter shrew (*Micropotamogale lamottei*), and golden mole (*Amblysomus hottentotus*). Otter shrew DNA was extracted from ethanol-preserved liver (voucher specimen IZEA-7083); sources of other DNA were as before (6, 7). Amplification was performed by using a forward primer hybridizing to exon 1 and a reverse primer complementary to the 3′ end of exon 2 (22). AQP2 was sequenced (23) for pig (*Sus scrofa*), sperm whale (*Physeter macrocephalus*), manatee (*Trichechus manatus*), fin whale (*Balaenoptera physalus*), tree hyrax (*Dendrohyrax dorsalis*), tail-less tenrec, small Madagascar hedgehog, and otter shrew.

Phylogenetic Tree Reconstruction. To study the evolution of the candidate Afrotherian signatures found in CRYAA, AQP2, and IRBP (see *Results*), phylogenetic trees were constructed from a 5,708-bp data set of concatenated α-2B adrenergic receptor, von Willebrand factor, IRBP, and 12S rRNA-tRNA valine-16S rRNA sequences (10), taking those entries that corresponded most closely with the species in our CRYAA, AQP2, and IRBP data sets (see Table 5, which is published as supplemental data on the PNAS web site). In the case of the IRBP signature, phylogeny was constructed with exclusion of the IRBP sequences. Topologies and branch lengths of the obtained trees are thus independent of the protein sequences whose signatures we investigate. It also avoids the problem that covarion processes might influence our tree building (24).

We used a two-step procedure to derive the maximum likelihood (ML or maximum average likelihood *sensu* Steel and Penny, ref. 25) phylogeny from our sequence data. The size of our phylogenies precluded an exhaustive search of all possible topologies to find the global ML tree. We therefore first calculated the likelihood of the sequence data on starter topologies obtained from a simple neighbor-joining (minimum evo-

EVOLUTION

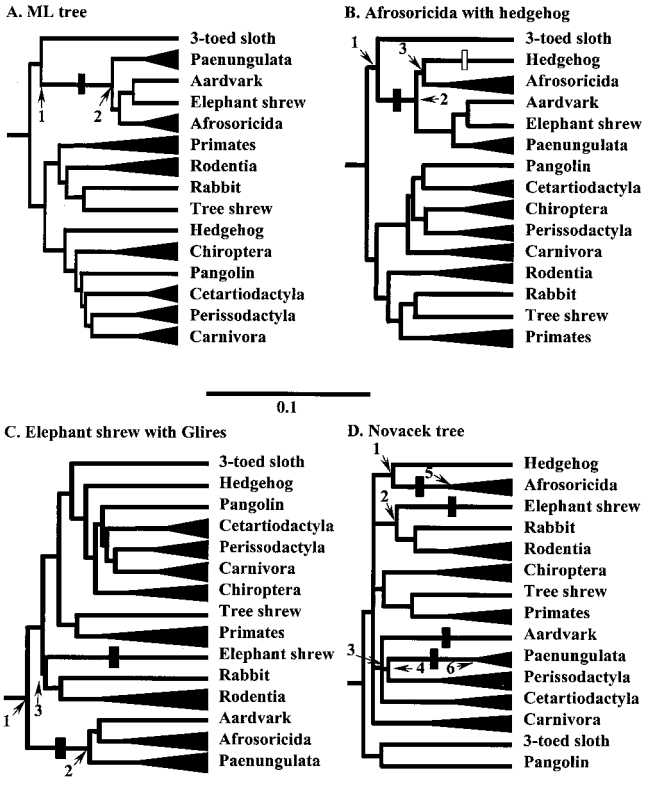

Fig. 2. Alternative topologies used to calculate the likelihood of the CRYAA signature. Trees are constructed from a 5,708-bp concatenation of six genes representing the species as indicated by * in Fig. 1 *A* and *D*, using kangaroo and opossum as out-group. (*A*) Unconstrained ML tree. (*B*) Tree enforcing the association of Afrosoricida (Madagascar hedgehog and golden mole) with hedgehog. (*C*) Tree constrained to group elephant shrew with Glires. (*D*) Tree constrained to conform with morphological relationships of eutherian orders as proposed by Novacek (11). All trees present internal branch lengths proportional to likelihood; terminal branches are shortened, and related species combined (Paenungulata: African elephant, manatee and hyrax; primates: galago and human; Rodentia: mole rat and guinea pig; Chiroptera: fruit-eating bat and flying fox; Cetartiodactyla: minke whale, cow, and pig; Perissodactyla: rhinoceros and horse; Carnivora: mink and seal). Filled and open bars indicate where the QLLC signature is assumed to have evolved and disappeared, respectively. In *B* it is equally parsimonious to have the signature evolve twice, in the ancestor of Afrosoricida and the aardvark-elephant shrew-paenungulate clade, respectively. For complete CRYAA trees and corresponding AQP2 and IRBP trees see Fig. 6. The estimated posterior probabilities of observing the signature QLLC at the numbered nodes are (*B*) ≈0.0, 0.820, and 0.796 at nodes 1, 2, and 3; (*C*) ≈0.00, 0.982, and ≈0.00 at nodes 1, 2, and 3; (*D*) ≈0.0 at nodes 1–4, and 0.923 and 0.507 at nodes 5 and 6, respectively.

lution) analysis. Likelihood calculations were done by using PAUP-ML (26) assuming the HKY85 model of evolution with gamma rate heterogeneity to allow for the possibility of unequal rates of evolution across sites. We estimated the shape parameter (α) of the gamma distribution, and the transition/transversion ratio, from the data. This yielded a candidate topology with branch lengths based on the ML distance calculation. Subsequently, we searched for better topologies in the region of the initial neighbor-joining topology by using the tree-bisection-reconnection branch swapping algorithm in PAUP-ML. We repeated this analysis procedure with random sequence input orders and always found the same ML tree. "Constrained trees" were constructed to conform with alternative morphology-based hypotheses for eutherian relationships (see Fig. 2). For these trees we supplied the topologies and reconstructed branch lengths by ML.

Computation of Ancestral States. We conducted two sorts of likelihood computation to investigate which trees best described the CRYAA, AQP2, and IRBP sequence evolution. In one we calculated the overall likelihood of observing the protein sequence signature separately for CRYAA, AQP2, and IRBP, using as our model the empirical JTT substitution rate matrix (27). A separate likelihood was calculated on the unconstrained ML tree and on the three morphology-constrained topologies. The second set of computations involved the likelihood of the most probable ancestral character states of the candidate sequence signatures. These calculations used the same model of evolution, and followed established procedures of which the details have been described (20, 28, 29). These procedures calculate the likelihood of observing the protein sequence data given a topology and a specified amino acid at some node. A likelihood is calculated for each possible amino acid, with the largest corresponding to the ML estimate. The ratio of the largest likelihood to the sum over all amino acids (the total likelihood), each weighted by the prior probabilities of occurrence, is a measure of the posterior probability of that amino acid at that node. As is customary in such analyses, we assume equal prior probabilities for each amino acid, although basing our calculations on priors equal to the proportion of a given amino acid in the sequence does not alter our conclusions. The product of the probabilities over the separate amino acids that comprise a signature measures the probability of the entire signature at that node. By comparing probabilities at a pair of ancestral and descendant nodes it can be inferred whether the signature arose in the branch leading to the descendant node.

Results

Candidate Sequence Signatures in CRYAA, AQP2, and IRBP. In the alignment of 28 mammalian CRYAA sequences, six positions were found to be relevant for distinguishing any possible five-species clade (Fig. 1*A*). The only group of five species set apart by a combination of two or more putative apomorphies, namely 70Q, 74L, and 142C, is formed by elephant, manatee, hyrax, aardvark, and elephant shrew. The combination QLC at positions 70, 74, and 142 thus is a unique feature for the Afrotheria in this CRYAA data set. All Afrotheria, apart from African elephant, share in addition the apomorphy 72L. We therefore investigated the phylogenetic value of 70Q, 72L, 74L, 142C as a putative Afrotherian signature in CRYAA. Among 20 aligned AQP2 sequences, we traced four positions at which putative apomorphies might be diagnostic for a six-species clade (Fig. 1*B*). The combination 10S, 55T and 104L perfectly set apart the Afrotheria, with exception of dugong, which only shares the 10S apomorphy. The signature STL was studied as an Afrotherian marker in AQP2. In the alignment of 28 IRBP sequences, 47 positions passed the search for a five-species grouping (Fig. 1*C*). There are 17 putative apomorphies in support of at least four of the five Afrotheria. The combination 18 M, 19S, 76S, 147G, 226I, 272D, and 328E is even perfectly unique for all five Afrotheria. At the 10 other positions the signature is affected by homoplasy, within the limits allowed by our search procedure (see *Materials and Methods*). The "degenerate" 17-residue signature is used in our further analyses. No signatures were detected in the other six proteins.

CRYAA and AQP2 Signatures in Other Afrotheria. To perform meaningful likelihood calculations for the retrieved signatures it was desirable to broaden the Afrotherian representation by sequencing CRYAA and IRBP in golden mole and tenrec, and AQP2 in tenrec. This would also be a test for the diagnostic value of these signatures; if they are genuine synapomorphies for Afrotheria, one expects to find them, completely or partially, in CRYAA, AQP2, and IRBP from other members of this clade. We sequenced exon 2 of the CRYAA gene, which encodes the

Table 1. Likelihood of the signatures in CRYAA, AQP2, and IRBP when reconstructed on alternative tree topologies

Protein	Topology*	Log-likelihood†	Log-difference from best tree†
CRYAA (n = 27)	A	−90.07 (−1540.09)	0.00 (0.00)
	B	−100.12 (−1548.75)	10.05 (8.66)
	C	−103.65 (−1563.69)	13.58 (23.60)
	D	−119.17 (−1667.25)‡	29.1 (127.16)‡
AQP2 (n = 20)	A	−51.95 (−876.97)	0.00 (0.00)
	B	−61.83 (−891.97)	9.88 (15.00)
	C	−61.52 (−896.81)	9.57 (19.84)
	D	−98.78 (−1001.13)‡	46.83 (124.16)‡
IRBP (n = 28)	A	−541.55 (−6118.73)	0.00 (0.00)
	C	−590.41 (−6217.33)	48.86 (98.27)
	D	−735.74 (−6419.23)‡	194.19 (300.46)‡

n, Number of sequences used for tree constructions.

*Topologies as explained in legends of Fig. 2; tree B is lacking for IRBP, Afrosoricida not being available.

†Likelihoods in parentheses are those calculated for the entire protein sequence and agree in every case with those calculated for the signature sequence alone.

‡Likelihood of tree D adjusted to have same number of branches as other trees.

signature positions 70, 72, and 74, in Indian elephant, dugong, golden mole, and three Tenrecidae, including the otter shrew. All new CRYAA sequences were found to code for 70Q, 72L, and 74L, including that of Indian elephant, suggesting that 72V in African elephant is a back mutation (Fig. 1*D*). For AQP2, additional sequences were obtained for manatee, tree hyrax, and three tenrecs, again including otter shrew. All of these species have the STL signature, apart from manatee, which like dugong AQP2 misses 55T and 104L (Fig. 1*E*). Unfortunately, sequences for golden mole and tenrec IRBP could not be obtained.

These new sequences illustrate that signatures, even in short proteins like CRYAA and AQP2, have the potential to identify newly investigated species as belonging to a specific clade. These data confirm that golden moles and tenrecs associate with Afrotheria, and indicate that the otter shrew joins this clade.

Likelihoods of the CRYAA, AQP2 and IRBP Signatures. To calculate the likelihood of the signatures in the Afrotherian species we needed topologies representing alternative hypotheses about their relationships. To construct these alternative topologies we used a 5,708-bp concatenation of six genes (10) that is the only extensive sequence data set available for most taxa that are relevant for our calculations (indicated by asterisks in Fig. 1). It allowed us to make trees with topologies and branch lengths independent of the particular signature under investigation. Fig. 2*A* shows the topology of the ML tree used for calculating the likelihoods of the CRYAA signature. In this tree the African clade receives bootstrap support of 100%. The principle morphologically favored alternatives are to group Afrosoricida with hedgehog in a monophyletic Lipotyphla, and elephant shrew with Glires (rabbits and rodents; refs. 11–14). The trees in Fig. 2 *B–D* are constrained to comply with these morphology-based hypotheses. Similar sets of alternative trees were constructed for the AQP2 and IRBP data sets (see Fig. 6, which is published as supplemental data on the PNAS web site).

The log-likelihoods of the CRYAA, AQP2, and IRBP signatures were separately calculated on the corresponding ML and constrained trees (Table 1). The signatures fit in every case the unconstrained ML tree substantially better than any of the constrained trees, providing independent support in three proteins for the Afrotherian clade.

Likelihoods of Ancestral State Reconstructions. If the signatures in CRYAA, AQP2, and IRBP are synapomorphies of Afrotheria they should have evolved in the branch leading to the basal node of the Afrotherian clade. The estimated posterior probabilities of observing the signature QLLC at nodes 1 and 2 in Fig. 2*A* are 3.0×10^{-9} and 0.984, respectively. For the AQP2 and IRBP ML trees the corresponding probabilities are 2.0×10^{-6} and 0.987, and 7.8×10^{-32} and 0.391, respectively. The sequence signatures of all three proteins thus have a high probability of evolving in the branch leading to the basal node of the Afrotheria. Probabilities this high for the CRYAA and AQP2 signatures imply that each amino acid replacement in the signatures has a near 1.0 probability of having evolved in that branch. Even for the IRBP signature, which requires 17 separate events in a specified branch, the combined probability is 0.391. Removing just two of the more variable sites (e.g., 59S and 326E, each of which has an approximately 0.65 probability of having evolved in the branch), the combined probability rises to 0.94.

These results confirm that the absence of 72L in African elephant CRYAA must be a loss of L at that site. Similarly, the absence of 55T and 104L in dugong AQP2 is reconstructed as a loss in the branch leading to that species. We also infer that elephant shrew IRBP has lost 59S and 326E, and other instances of homoplasy arise (compare Fig. 1*C*). However, none constitutes an alternative to the signatures we investigate here.

The morphology-constrained trees each require that the signatures evolved more than once or have evolved and been lost again. Reconstructions of ancestral states similar to those for the ML tree support this interpretation, as shown for CRYAA by the probabilities at the nodes numbered in Fig. 2 *B–D*, and given in the legends. Comparable values were found for the constrained AQP2 and IRBP trees (Fig. 6). However, to reject the constrained trees solely on the basis that they require more than one gain or loss of the signatures requires a framework within which to consider the probability of a signature event occurring more than once on a tree. If this probability is high, then the alternative topologies are not ruled out by our data.

Phylogenetic Value of the Afrotherian Signatures. Is it unlikely that the signatures we observed have evolved more than once? To answer this question we developed a methodology that takes account of all possible ways a signature could have arisen given the number of elements (i.e., amino acid replacements in the signature) and the length of the protein. This removes the possibility that we have capitalized on chance. First, we calculate the probability of a given class of signature events arising once. Let r be the number of apomorphic elements in a signature. The class of r-events (i.e., all of the possible signatures of size r) need not be unlikely itself, but for the signature to be an unambiguous marker of a clade the probability must be low that the same (identical) member of the class arises twice.

Given V variable sites in a sequence, and a signature of size r there are $\binom{V}{r}$ possible signatures of size r. Each signature has probability $p^r q^{V-r}$ of occurring in any given branch, where p is the probability of an amino acid replacement at a given site in a branch, and $q = 1 - p$. We assume that p is constant across sites. The product

$$\binom{V}{r} p^r q^{V-r} \qquad [1]$$

gives the probability of an r-event. Summing this product over r, allowing r to range from r to V, gives the probability of a signature of length r or greater. Call this probability p_s, where s denotes signature.

Table 2. Probabilities of signatures occurring in the ML trees

Protein (length)	Variable sites	Total changes on tree*	$p^{\dagger}$	$p_s^{\ddagger}$	$p_t^{\S}$	Probability p for same signature of r sites to occur		
						Twice	Three times	Four times
CRYAA (173)	57	123	0.014	0.008	0.328	2.80×10^{-7}	1.22×10^{-13}	3.43×10^{-20}
AQP2 (111)	31	78	0.018	0.019	0.523	6.76×10^{-5}	4.47×10^{-9}	1.92×10^{-13}
IRBP (334)	227	912	0.051	0.069	0.979	9.48×10^{-26}	4.51×10^{-51}	1.40×10^{-76}

*Calculated by reconstructing the most parsimonious set of amino acid replacements on the ML tree.
†Probability of a substitution per site per branch. Calculated as total changes/length of protein/no. of branches in tree. Our conclusions are unaltered if we calculate p using the number of variable sites rather than the length of the protein.
‡Probability of a signature of length r; for CRYAA, AQP2, and IRBP, r is 4, 3, and 17, respectively.
§Probability of any p_s event at least once on a tree (further defined in the text).

The probability that a signature of length r or greater will arise at least once in a given tree is calculated as follows. Let there be N_b branches in the tree. Then

$$\binom{N_b}{b} p_s^b (1 - p_s)^{(N_b - b)} \qquad [2]$$

gives the probability of observing a signature of length r or greater in b branches of the tree. Summing this product over b ranging from 1 to N gives the probability of observing on the tree at least one signature of length r or greater. Call this quantity p_t, where the t denotes the tree.

We estimated p for each protein, from the number of sites in the sequence, the total number of changes reconstructed on the ML tree, and the number of branches in the tree. We then applied this estimate of p to all sites to calculate p_s and p_t (Table 2). The results show that none of our signature classes alone is improbable. Thus, given as many variable sites as we observe in each protein, signatures of the sort we have detected or longer, are expected somewhere on each tree.

For the identical signature to arise twice in a tree of N_b branches, any of the r-length events can happen first and anywhere on the tree, but the second r-length event can only be one of the $\binom{V}{r}$ possible signatures of size r. Each of these occurs in any give branch with probability $(p^r q^{V-r})$; call this probability p_b, where b denotes branch. Then, the probability of the identical signature arising twice is given by the product of p_t and all possible ways of the second signature arising in the $N_b - 1$ remaining branches. (In fact the number of branches in which the second signature can arise typically will be less than $N_b - 1$ because the first signature will usually be present in more than one branch of the tree, owing to identity by descent. This makes our calculations conservative.) This product is written as

$$p_t \binom{N_{b-1}}{I} p_b^I (1 - p_b)^{(N_{b-1} - I)} \qquad [3]$$

and the symbol I takes the value 1 to account for one additional signature arising. Using the same logic, Eq. **3** can be used to calculate the probability of the same signature arising three or four times by allowing $I = 2$ or $I = 3$. Table 2 reports the resulting probability for two, three, and four identical r-events. In these calculations we have replaced the p_b of Eq. **3** with p_b summed over all signatures of length r or greater. The calculations reported in Table 2 reveal that, although the class of r-events (p_t) is not improbable, the probability of the identical r- or greater-length event occurring twice or more by chance is always small and often negligible.

Combining Results from the Three Proteins. How do these results alter our view about the likelihood of Afrotherian monophyly? Using Bayes' rule (30) we can combine the signature probabilities from Table 2 to arrive at a posterior probability for the Afrotherian hypothesis. From Bayes' rule

$$P(Afrotheria) = \frac{w(Afrotheria)P(signature/Afrotheria)}{P(signatures)}, \qquad [4]$$

where *P(Afrotheria)* is the posterior probability of the Afrotheria signature, *w(Afrotheria)* is our prior belief in the Afrotherian hypothesis, *P(signature/Afrotheria)* is the probability of the Afrotherian signature given the unconstrained ML tree, and *P(signatures)* is the combined probabilities of the signatures summed over all four trees, weighted by their prior probabilities. *P(signature/Afrotheria)* is obtained from the p_t column in Table 2, and *P(signatures)* by combining the Afrotherian results with those from the appropriate column of Table 2, corresponding to the number of times a signature has appeared in the three alternative trees.

Let our prior belief be skeptical to adopt a conservative view against the Afrotherian hypothesis. Let *w(Afrotheria)* be 0.001. Let our prior belief in the morphology-based hypotheses represented by the other trees be higher, at 0.333 each (0.4995 for IRBP). These weights then sum to 1.0 as they must. Table 3 reports that for all three proteins the posterior belief in Afrotheria is strong and substantially altered from the prior. Calculating the combined posterior support of the three proteins for the Afrotherian hypothesis yields $P \sim 1.0$, even when a prior weight of only 0.0001 is used. Thus, the combined data effectively rule out support for polyphyly of the Afrotherian species.

Discussion

The sequence signatures that we identified in CRYAA, AQP2, and IRBP (Fig. 1), without resorting to prior phylogenetic analyses, provide independent evidence for the Afrotherian clade. The signatures are specific to Afrotheria, they arose with high probability at the basal node of the Afrotherian clade, and it is highly improbable that they would have arisen more than once as is required by the morphologically favored tree hypotheses. We demonstrated their predictive value by finding them in several species for which sequence data on the CRYAA and AQP2 proteins did not previously exist. Notably, the finding of the Afrotherian signatures in the otter shrew—for which no

Table 3. Bayesian analysis of the evidence for the monophyly of Afrotheria (see text for explanation)

Protein	Prior weight for Afrotherian clade	Posterior probability of Afrotheria
CRYAA	0.001	0.999
AQP2	0.001	0.921
IRBP	0.001	~1.00
Combined	0.0001	1.000

other sequence data have yet been published—supports the inclusion of this African insectivore in the Afrotheria.

Can the Afrotherian signatures be dismissed as homoplasy? The parallel appearance of signatures in a data set could be caused by the admixture of paralogous sequences, convergence, covarion processes, lineage sorting, or even to bias in base composition or differences in mutational mechanisms or repair systems (31). However, it seems highly implausible that such evolutionary mechanisms would cause similarly misleading signatures in three functionally independent proteins in precisely the same set of species.

At a methodological level our assumption that sites evolve independently may be questioned. An extensive literature deals with the correlated evolution of amino acid residues in a protein (e.g., refs. 32 and 33). Such mutual dependence makes it understandable that two or more replacements can originate or disappear in concert. To the extent that the amino acid replacements we have identified do change in a correlated manner, our calculations may underestimate the true probabilities of the signatures arising twice. Similarly, we have used a single estimate of the probability of a substitution to characterize every site and every branch. To the extent that the true probability varies our estimates may be affected. However, we reiterate that we have found similar highly improbable signature patterns in three independent proteins and always in the same set of species. Even using our simplifying assumptions, the results are congruent across trees and proteins. Further, our approach uses a statistical methodology that controls for the problem of capitalizing on chance that arises when searching for signatures of unknown length and composition.

The phylogenetic signal contained in sequence signatures, if present in a protein, contributes in any conventional phylogenetic analysis to the topology that is eventually reconstructed. What then is added by identifying and analyzing signatures on their own? It appears that the quantitative approach of analyzing ever longer sequences is not in all instances the panacea of molecular phylogeny, as in the case of deeper level analyses of mitochondrial protein sequences (e.g., refs. 34–38). If one accepts that synapomorphies are the cornerstones of phylogeny reconstruction, it is logical then to additionally search for mutational events that act as qualitative sequence characteristics for a specific clade. Such can be retropositions (39), specific insertions or deletions (e.g., refs. 10, 19, and 22), and the sequence signatures as discussed here. These molecular character-state data may allow a better discrimination between homoplasy and homology, a prerequisite for finding "true" trees (31). Where conventional analyses combine all of the site-by-site information into a single result, the signature approach highlights a concrete set of events whose most plausible evolutionary explanation can help to choose among competing phylogenetic hypotheses.

The "protein morphological" evidence provided by the signatures in CRYAA, AQP2, and IRBP may give an impetus to reevaluate the apparent absence of any morphological synapomorphies for the African clade against the backdrop of the various scenarios for lipotyphlan phylogeny (14, 40).

We thank Collin van Asten for writing SIGNWIN, Marcel Sweers for technical help, and Peter Vogel for otter shrew tissue. This work was supported by grants from the Netherlands Foundation for Life Sciences (to W.W.d.J.), the European Commission (to W.W.d.J., F.C. and M.J.S.), and the Leverhulme Trust and the Natural Environment Research Council (to M.P.).

1. de Jong, W. W. (1998) *Trends Ecol. Evol.* 13, 270–275.
2. Waddel, P. J., Okada, N. & Hasegawa, M. (1999) *Syst. Biol.* **48,** 1–5.
3. Allard, M. W., Honeycutt, R. L. & Novacek, M. J. (1999) *Cladistics* **15,** 213–219.
4. D'Erchia, A. M., Gissi, C., Pesole, G., Saccone, C. & Arnason, U. (1996) *Nature (London)* **381,** 597–600.
5. Gatesy, J. (1997) *Mol. Biol. Evol.* **14,** 537–543.
6. Springer, M. S., Cleven, G. C., Madsen, O., de Jong, W. W., Waddell, V. G., Amrine, H. M. & Stanhope, M. J. (1997) *Nature (London)* **388,** 61–64.
7. Stanhope, M. J., Wadell, V. G., Madsen, O., de Jong, W. W., Hedges, S. B., Cleven, G. C., Kao, D. & Springer, M. S. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 9967–9972.
8. Gatesy, J., Milinkovitch, M., Waddell, V. & Stanhope, M. J. (1999) *Syst. Biol.* **48,** 6–20.
9. Springer, M. S., Amrine, H. M., Burk, A. & Stanhope, M. J. (1999) *Syst. Biol.* **48,** 65–75.
10. Madsen, O., Scally, M., Kao, D. J., DeBry, R. W., Douady, C. J., Adkins, R., Amrine, H., Stanhope, M. J., de Jong, W. W. & Springer, M. S. (2001) *Nature (London)*, in press.
11. Novacek, M. J. (1992) *Nature (London)* **356,** 121–125.
12. McKenna, M. C. & Bell, S. K. (1997) *Classification of Mammals Above the Species Level* (Columbia Univ. Press, New York).
13. Liu, F. R. & Miyamoto, M. M. (1999) *Syst. Biol.* **48,** 54–64.
14. Asher, R. J. (1999) *Cladistics* **15,** 231–252.
15. Woese, C. R., Kandler, O. & Wheelis, M. L. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 4576–4579.
16. Germot, A., Philippe, H. & Le Guyader, H. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 14614–14617.
17. Gupta, R. S. & Johari, V. (1998) *J. Mol. Evol.* **46,** 716–720.
18. Luckett, W. P. & Hong, J. (1998) *J. Mamm. Evol.* **5,** 127–182.
19. Huchon, D., Catzeflis, F. M. & Douzery, E. J. (2000) *Proc. R. Soc. London Ser. B* **267,** 393–402.
20. Pagel, M. (1999) *Syst. Biol.* **48,** 612–622.
21. Pagel, M. (1999) *Nature (London)* **401,** 877–884.
22. van Dijk, M. A. M., Paradis, E., Catzeflis, F. & de Jong, W. W. (1999) *Syst. Biol.* **48,** 94–106.
23. Madsen, O., Deen, P. M. T., Pesole, G., Saccone, C. & de Jong, W. W. (1997) *Mol. Biol. Evol.* **14,** 363–371.
24. Lockhart, P. J., Steel, M. A., Barbrook, A. C., Huson, D. H., Charleston, M. A. & Howe, C. J. (1998) *Mol. Biol. Evol.* **15,** 1183–1188.
25. Steel, M. & Penny, D. (2000) *Mol. Biol. Evol.* **17,** 839–850.
26. Swofford, D. L. (1998) PAUP*: *Phylogenetic Analysis using Parsimony (* and other methods)* (Sinauer, Sunderland, MA), Version 4.
27. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992) *Comput. Appl. Biosci.* **8,** 275–282.
28. Yang, Z., Kumar, S. & Nei, M. (1995) *Genetics* **141,** 1641–1650.
29. Koshi, J. M. & Goldstein, R. A. (1996) *J. Mol. Evol.* **42,** 313–320.
30. Bain, L. J. & Engelhardt, M. (1987) *Introduction to Probability and Mathematical Statistics* (Duxbury, Boston).
31. Doyle, J. J. (1996) in *Homoplasy: The Recurrence of Similarity in Evolution,* eds. Sanderson, M. J. & Hufford, L. (Academic, San Diego), pp. 37–66.
32. Mateu, M. G. & Fersht, A. R. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 3595–3599.
33. Pollock, D. D., Taylor, W. R. & Goldman, N. (1999) *J. Mol. Biol.* **287,** 187–198.
34. Philippe, H. (1997) *J. Mol. Evol.* **45,** 712–715.
35. Naylor, G. J. P. & Brown, W. M. (1998) *Syst. Biol.* **47,** 61–76.
36. Curole, A. P. & Kocher, T. D. (1999) *Trends Ecol. Evol.* **14,** 394–398.
37. Takezaki, N. & Gojobori, T. (1999) *Mol. Biol. Evol.* **16,** 590–601.
38. Waddell, P. J., Cao, Y., Hauf, J. & Hasegawa, M. (1999) *Syst. Biol.* **48,** 31–53.
39. Hillis, D. M. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 9979–9981.
40. Seiffert, E. R. & Simons, E. L. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 2646–2651. (First Published February 29, 2000; 10.1073/pnas.040549797)