Molecular Evolution of Mammalian Aquaporin-2: Further Evidence that Elephant Shrew and Aardvark Join the Paenungulate Clade

Ole Madsen,* Peter M. T. Deen,† Graziano Pesole,‡ Cecilia Saccone,§ and Wilfried W. de Jong*†

*Department of Biochemistry and †Department of Cell Physiology, University of Nijmegen, The Netherlands; ‡Department of Biology, University of Basilicata, Potenza, Italy; §Department of Biochemistry and Molecular Biology, University of Bari, Italy; and †Institute for Systematics and Population Biology, University of Amsterdam, The Netherlands

A 328-bp sequence from exon 1 of the gene for aquaporin-2 (AQP2) was compared in 12 mammalian species, representing as many eutherian orders. This sequence encodes the N-terminal half of this kidney-specific water channel protein. Most amino acid replacements, as well as an insertion, have occurred in extracellular loops connecting the transmembrane helices, in agreement with a lower functional importance of these loops. Phylogenetic analyses were performed with parsimony, distance, and maximum-likelihood methods. The AQP2 data set, alone as well as in combination with previously published αA-crystallin protein sequences, strongly supports a clade consisting of elephant, hyrax, aardvark, and elephant shrew, reaching bootstrap values of 99%. This finding fully agrees with the only other presently available sequence data sets that include these taxa, those of von Willebrand factor and interphotoreceptor retinoid-binding protein, and suggests that this extended paenungulate clade is one of the most conspicuous superordinal groupings in eutherian phylogeny. Some support was obtained for an artiodactyl/perissodactyl clade, while the grouping of pholidotes with edentates was contradicted.

Introduction

The divergence of the major groups of placental mammals occurred in the Cretaceous, about 100 million years ago (Archibald 1996; Hedges et al. 1996). Because of their relatively rapid radiation, the resolution of the phylogenetic relationships between the 18 orders of living eutherian mammals is intrinsically difficult. Indeed, comparison of recent reviews about eutherian ordinal phylogeny reveals distressingly little consensus, both within and between molecules and morphology (Benton 1988; Novacek 1992; Graur 1993; Honeycutt and Adkins 1993; Szalay, Novacek, and McKenna 1993; Al-lard, McNiff, and Miyamoto 1996). A major reason for the as yet limited progress in the resolution of mammalian ordinal phylogeny is the lack of adequate molecular data sets covering orthologous sequences from a sufficiently broad array of orders. This is especially true for such traditionally problematic orders as Pholidota ( pangolins), Hyracoidea (hyraxes), Tubulidentata (aard- varks), and Macroscelidea (elephant shrews). Pholidota have usually been considered a sister group to the Ed- entata (sloths, anteaters, and armadillos) (Patterson 1978; Novacek 1992), but certain morphological studies (McKenna 1975; Rose and Emry 1993) and the scanty protein sequence and serological data (de Jong and Goodman 1982; Shoshani 1986; Czelusniak et al. 1990; de Jong, Leunissen, and Wistow 1993; Sarich 1993) contradict this view, and rather suggest some relation with carnivores. As for Hyracoidea, morphologists are undecided (Novacek 1992) whether to place them with Perissodactyla (Fischer and Tassy 1993; Prothero 1993) or with Proboscidea and Sirenia in the Paenungulata (Simpson 1945; Shoshani 1986). The molecular data, however, unequivocally support their inclusion within the Paenungulata (de Jong, Zweers, and Goodman 1981; Czelusniak et al. 1990; Lavergne et al. 1996; Porter, Goodman, and Stanhope 1996; Stanhope et al. 1996). Morphologically, Tubulidentata is now generally given an unresolved position within the Ungulata (Artiodac-tyla, Cetacea, Perissodactyla and Paenungulata) (Patter-son 1978; Thewissen 1985; Novacek 1992), but a group-ing with Insectivora still gets support (Gaudin et al. 1996). Serological data (Shoshani 1986; Sarich 1993), but especially molecular sequences (de Jong, Zweers, and Goodman 1981; Czelusniak et al. 1990; Porter, Goodman, and Stanhope 1996; Stanhope et al. 1996), firmly group aardvark with Paenungulata. Macroscelidea, originally grouped with Insectivora, are now on a morphological basis generally placed as an outgroup of the Glires (Novacek 1992; Gaudin et al. 1996), although a derivation from Condylarthra—the ungulate ances-tors—has also been proposed (Hartenberger 1986; Si-mons, Holroyd, and Bown 1991) (for review, see Butler 1995). Consistent with the latter proposition is the growing molecular evidence that elephant shrews form a monophyletic clade with elephants, sea cows, hyraxes and aardvarks (de Jong, Leunissen, and Wistow 1993; Porter, Goodman, and Stanhope 1996; Stanhope et al. 1996).

To help resolve the disputed affinities of the above four orders, we decided to determine sequences of a suitable protein-coding gene from the nuclear DNA of relevant mammalian species. The aquaporin-2 (AQP2) gene was chosen because it is, in man, a single-copy gene—located on chromosome 12 (Deen et al. 1994a)—and demonstrates a promising degree of sequence difference between man (Uchida et al. 1994) and rat (Fu-simi et al. 1993). The encoded protein has, moreover, interesting evolutionary and biomedical features. Aquaporins are water channel proteins that facilitate selective water transport across cell membranes in a variety of tissues (Agre, Brown, and Nielsen 1995). AQP2 is ex-
clusively expressed in the kidney collecting duct, where it is responsible for the vasopressin-dependent concentration of urine (Fushimi et al. 1993; Nielsen et al. 1993; Ma et al. 1994). Mutations in this gene cause nephrogenic diabetes insipidus in man (Deen et al. 1994b; van Lieburg et al. 1994). AQP2 and the other four known mammalian AQPs belong to the ubiquitous major intrinsic protein (MIP) family, which transport small molecules across membranes. These proteins are predicted to consist of six bilayer-spanning α-helices, with the NH2- and COOH-termini located intracellularly (Gorin et al. 1984; Preston et al. 1994). The protein consists of two repeated halves, supposed to be the result of an intragenic tandem duplication (Wistow, Pisano, and Chepelnisky 1991; Reizer, Reizer, and Saier 1993). The first half is encoded by exon 1 of the AQP2 gene and consists of 121 amino acids.

It is the coding region of this exon 1 that we now have sequenced in 10 mammalian species. Together with the published rat and human sequences, these provide a data set representing 12 eutherian orders. Though relatively short, these sequences contain considerable phylogenetic signal, which has been analyzed also in combination with sequences of the eye lens protein αA-crystallin from the same taxa.

Materials and Methods

Aquaporin-2

The AQP2 sequences reported here are from the following species and orders: nine-banded armadillo (Dasypus novemcinctus; Edentata), Indian elephant (Elephas maximus; Proboscidea), aardvark (Orycteropus afer; Tubulidentata), elephant shrew (Microscelis proboscideus; Macroscelidea), Cape hyrax (Procavia capensis; Hyaenidae), pangolin (Manis sp; Pholidota), rabbit (Oryctolagus cuniculus; Lagomorpha), horse (Equus caballus; Perissodactyla), dog (Canis familiaris; Carnivora), and bovine (Bos taurus; Artiodactyla). The AQP2 sequences from rat (Fushimi et al. 1993) and human (Uchida et al. 1994) were extracted, as published, respectively, of the first exon of AQP2 were used for PCR amplification positions -5 to +14 and +343 to +364, respectively. PCR reactions were performed in 50 µl, containing 20 mM Tris-11C1, pH 8.4; 50 mM KCl; 1-2 mM MgCl2; 0.27% Tween 20; 0.27% Nonidet P40; 0.2 mg/ml BSA; 200 µM dNTP (Boehringer Mannheim and Gibco BRL); 0.4-5 µM of each primer; 1 U of Taq DNA polymerase (provided by Dr. Wiljan Hendriks, University of Nijmegen); and 50-500 ng of genomic DNA as template. Reactions were carried out in a Bio- metra (Tampa, Fla.) TRIO-thermoblock using the following program: 7 min at 95°C, 30 cycles of 2 min at 95°C, 1 min at 54-62°C, 30 s at 72°C, and, finally, 5 min at 72°C. PCR products were ligated directly into a T/A cloning vector (Invitrogen or Promega) according to the manufacturers’ instructions, followed by transformation in E. coli JM109. Clones containing AQP2 inserts were isolated by means of filter hybridization using an 800-bp fragment of rat AQP2 cDNA as probe. Sequences in both directions were determined from at least three independently obtained PCR products, using the Sequenase version 2.0 sequencing kit (USB). All sequences have been deposited in the EMBL database under the accession numbers Y10629—Y10638.

αA-Crystallin

Sequences of αA-crystallin used in this study were obtained from the Swissprot database, except the human sequence, which was obtained from the EMBL database, and the elephant shrew (Macrogalidia proboscideus) sequence, which was as reported in de Jong, Leunissen, and Wistow (1993). Species, orders, and accession numbers were: bovine (Bos taurus; Artiodactyla), P02470; three-fingered sloth (Bradypus variegatus; Edentata), P02487; dog (Canis familiaris; Canidae), P02473; horse (Equus caballus; Perissodactyla), P02478; African elephant (Loxodonta africana; Proboscidea), P02498; Malayan pangolin (Manis javanica; Hipposodactyla), P02498; aardvark (Orycteropus afer; Tubulidentata), P02501; Cape hyrax (Procavia capensis; Hyaenidae), P02499; rabbit (Oryctolagus cuniculus; Lagomorpha), P02493; rat (Rattus norvegicus; Rodentia), P02490; and human (Homo sapiens; Primates) P02489.

Phylogenetic Analyses

Sequences were aligned using the PILEUP program from the GCG package (Devereux, Haebeli, and Smithies 1984), followed by manual editing to obtain maximum-similarity alignments. DNA data were analyzed with the following methods: (1) the Stationary Markov Model (SMM) (Sacccone et al. 1990) to calculate pairwise genetic distances and the NEIGHBOR and DRAWGRAM programs to construct the phylogenetic trees (Felsenstein 1993) and (2) the maximum-likelihood method by using the DNAML program of the PHYLIP package (Felsenstein 1993) assuming transition/transversion ratios of 1.5, 2.0, 2.5, 3.0, and 5.0, and randomizing the order of input four times. Protein data were analyzed as follows: (1) with the maximum-parsimony method by using the PAUP program (Swofford 1993); (2) with the maximum-likelihood method by using the PROTML program (Adachi and Hasegawa 1992); and (3) with the Kimura (1983) method to calculate pairwise genetic distances with the PROTDIST program of the PHYLIP package and the NEIGHBOR and DRAWGRAM programs to construct the phylogenetic trees. The statistical significance of molecular phylogenies was assessed with the bootstrap simulation technique by using the SEQBOOT and CONSENSE programs of the PHYLIP package.

Results and Discussion

Molecular Evolution of Aquaporin-2

Exon 1 of the gene for AQP2 codes for the first 121 amino acids of this protein, which span the N-terminal extension, three transmembrane domains, the interdomain loops A and B, and most of the connecting
Peptide C (fig. 1). Based on the rat and human sequences, degenerated PCR primers were designed to amplify on genomic DNA a 328-bp fragment of the AQP2 exon 1, coding for residues 6–114. Unambiguous PCR products were obtained, cloned, and sequenced from 10 mammalian species representing as many orders. The obtained DNA sequences (fig. 2) and the deduced amino acid sequences (fig. 3) were aligned with their rat and human counterparts. The similarity between the different AQP2 proteins varies between 81% and 97%. From figures 1 and 3 it appears that most of the variation at the protein level is concentrated in the extracellular loops A and, as far as determined, C. In loop A, an insertion of two residues occurred in elephant shrew AQP2. Despite this variability, it seems that two prolines are required in loop A, probably to allow a correct bending of this short peptide between two transmembrane helices. As is obvious from figures 1 and 3, none of the conserved residues in the aquaporin family has been replaced in the mammalian AQP2 sequences. This is all in agreement with the structural topology as predicted by the "hourglass" model, where the Asn-Pro-Ala sequence in loop B is assumed to combine in the membrane with a similar "NPA box" in loop E to form the pore responsible for the transport of water (Jung et al. 1994). This model suggests no special function for loops A, C, and D. The hourglass model is derived from studies of AQP1, but also fits for AQP2, as based on functional analyses of AQP2 mutants involved in nephrogenic diabetes insipidus (Deen et al. 1994b; van Lieburg et al. 1994; Mulders et al. 1996). However, recently another structure has been proposed for AQP2 (Bai et al. 1996) which predicts that loops C and D are equally as important as loops B and E in the formation of the aqueous pathway. Although only seven residues of loop C have been compared, it is striking that four out of these seven positions show conspicuous variability. This would argue against an important role of loop C in AQP2, and, rather, favors the hourglass model.

Phylogenetic Analyses

Aquaporin-2

To study the relationships between eutherian orders, a marsupial would have been the outgroup of choice. However, attempts to amplify the AQP2 gene of an opossum and a kangaroo remained unsuccessful. We therefore resorted to using the armadillo as outgroup, considering that Edentata are generally hypothesized to be the most basal eutherian order (McKenna 1975; Novacek 1992; Allard, McNiff, and Miyamoto 1996; but see Gaudin et al. 1996), although molecular data to support this notion are scarce.

Visual inspection of the protein alignment in figure 3 already reveals, apart from the usual homoplasia, the most conspicuous phylogenetic signal. Fourteen phylogenetically informative positions are present. The only occasion where a combination of several unique amino acid replacements occurs in different species is seen in elephant, hyrax, aardvark, and elephant shrew. Three unequivocally apomorphous replacements (15 A→S, 58 A→T, and 107 I→L) occur solely in these four species, while the replacements at position 87 may also reflect a shared derived origin. These synapomorphic replacements are indeed the result of unique point mutations in the DNA (fig. 2).
FIG. 2.—Alignment of the nucleotide sequences from exon 1 of the AQP2 gene from 12 mammalian species. Position numbers are according to the human sequence (Uchida et al. 1994) with the first nucleotide of the start codon as 1. Hyphens denote nucleotide identities and periods indicate gaps. Arrowheads indicate the positions of three synapomorphous substitutions in aardvark, elephant shrew, and paenungulates.
FIG. 3.—Alignment of the deduced amino acid sequences of AQP2 from 12 mammalian species. Position numbers follow the human sequence, with the initiation methionine as 1. Hyphens denote amino acid identities, periods indicate gaps, asterisks show residues conserved throughout the mammalian aquaporin family, and arrowheads indicate synapomorphous replacements in aardvark, elephant shrew, and paenungulates. The putative transmembrane helices and connecting loops A-C are indicated (cf. fig. 1).

FIG. 4.—Consensus tree, based on the amino acid sequences derived from exon 1 of AQP2 (cf. fig. 3). Majority-rule consensus trees were obtained both by the maximum-parsimony method using the PAUP program (Swofford 1993) and by the neighbor-joining method (Saitou and Nei 1987) as applied on a distance matrix calculated according to Kimura (1983). In these trees, all nodes supported by bootstrap values under 50% were collapsed into polytomies, resulting in the same topology for both methods. Bootstrap percentages are based on 1,000 replications and represented for the maximum-parsimony and neighbor-joining methods (top and bottom values, respectively). Armadillo was used as outgroup for both methods.

This observation is fully corroborated by the phylogenetic analyses of the aligned amino acid and nucleotide sequences of AQP2. The amino acid sequences were analyzed by maximum-parsimony and Kimura/neighbor-joining methods (fig. 4). In both cases, the only significant phylogenetic relationship was found to be in the clade consisting of elephant shrew, aardvark, hyrax, and elephant, supported by bootstrap values of 89% and 90% for the maximum-parsimony and Kimura/neighbor-joining methods, respectively. The same clade is again the first to be recognized by the maximum-likelihood analysis as carried out by using the program PROTML (Adachi and Hasegawa, 1992):(((elephant, hyrax), aardvark), elephant shrew) (further data not shown). Within this clade, all three methods weakly supported the elephant and hyrax as sister groups. Figure 5 shows the consensus tree similarly obtained by both the SMM/neighbor-joining method and the maximum-likelihood method, on first and second codon positions of AQP2 DNA sequences. The clade including elephant shrew, aardvark, hyrax, and elephant is also supported at the DNA level by bootstrap values of 75% and 80% for the SMM/neighbor-joining and maximum-likelihood methods, respectively. No other clade obtained convincing and consistent support from the AQP2 data with all methods.
FIG. 5.—Consensus tree, based on first and second codon positions of AQP2 exon 1 DNA sequences (cf. fig. 2). Majority-rule consensus trees were obtained both by the neighbor-joining method (Saitou and Nei 1987) using the distance matrix as calculated with the Stationary Markov Model (Saccone et al. 1990) and by the maximum-likelihood method (DNAML with transition/transversion ratio 2.0; Felsenstein 1981). All nodes supported by bootstrap values under 50% in both these trees were collapsed into polytomies, resulting in the presented topology. Bootstrap percentages are based on 100 replications and represented for the neighbor-joining and maximum-likelihood methods (top and bottom values, respectively). Armadillo was used as outgroup.

Combined Aquaporin-2 and αA-Crystallin

The eye lens protein αA-crystallin is the only other protein for which sequences are known from the same 12 orders (including Pholidota) as described for AQP2. In order to see if additional phylogenetic information could be retrieved, the AQP2 and αA-crystallin amino acid sequences representing these 12 orders were tandemly combined, aligned, and subsequently subjected to the same phylogenetic analyses as the AQP2 amino acid sequence data. The combined amino acid sequences together have a length of 282 amino acids, except for the combined elephant shrew sequence (length 248 residues) and the edentate sequence (combined armadillo AQP2 and three-fingered sloth αA-crystallin, length 279 residues). Phylogenetic analyses of the tandemly aligned data set even more strongly supported a close phylogenetic relationship between elephant shrew, aardvark, hyrax, and elephant, with bootstrap values of 99% for both the maximum-parsimony method and the Kimura/neighbor-joining method (fig. 6). The parsimony analysis required 102 steps and gave only one tree. Both methods also weakly supported some other superordinal clades, notably those grouping bovine and horse, and rat and rabbit. Also, the maximum-likelihood analysis (PROTML) on this combined data set supported the elephant/hyrax/elephant shrew/aardvark clade in all of the 32 top ranking trees, while the horse/cow and rat/rabbit clades received support in 28 and 25 of the 32 trees, respectively (data not shown).

Paenungulate Phylogeny

Molecular sequence evidence showing that not only aardvark but—surprisingly—also elephant shrews should be grouped in one clade with the paenungulates was first obtained from the eye lens protein αA-crystallin (de Jong, Zweers, and Goodman 1981; de Jong, Leunissen, and Wistow 1993). This evidence is now fully corroborated by the AQP2 sequences and even more strongly supported by the recent studies of exon 28 of the von Willebrand factor (vWF) (Porter, Goodman, and Stanhope 1996) and of exon 1 of interphotoreceptor retinoid-binding protein (IRBP) (Stanhope et al. 1996). Within this clade, the AQP2, vWF, and IRBP data all indicate that hyrax is more closely related to Tethytheria (elephants and sirenians; McKenna 1975) than are aardvark and elephant shrew. αA-crystallin data, and as a consequence the combined AQP2/αA-crystallin protein analyses, place elephant as the first branch off the clade, probably as the result of a single back mutation (de Jong, Leunissen, and Wistow 1993). The branching order of elephant shrew and aardvark is less clear, although the analyses of especially vWF, but also AQP2 and—depending on the method—IRBP, suggest that aardvark is closer to elephant and hyrax than is elephant shrew.

The various paleontological and morphological evidence to support the widely diverging opinions about the relationships of hyrax, aardvark, and elephant shrew has been amply discussed by Porter, Goodman, and Stanhope (1996) and Stanhope et al. (1996). Molecular sequences had until recently contributed little to solving these phylogenetic problems. Four data sets are now available that include these orders as well as a sufficient
number of other eutherian orders. The fact that these data sets consistently and independently provide strong support for a close relationship of aardvark and elephant shrew with Paenungulata leaves little doubt that this clade represents one of the most conspicuous lineages in eutherian radiation. Evidence from a single nuclear gene can easily be dismissed by assuming hidden paralogy or invoking a molecular evolutionary mechanism like gene conversion or homoplasia due to functional constraints. Such explanations cannot apply, however, to the genes coding for proteins with such diverse structures and functions as αA-crystallin, vWF, IRBP and AQP2. It is puzzling why this pronounced phylogenetic signal at the molecular level is apparently not paralleled by similarly convincing paleontological and morphological evidence.

Pholidote Relationships

Although no other phylogenetic relationships gained consistent support from the AQP2 data by itself or in combination with αA-crystallin, some trends are noteworthy. AQP2 and αA-crystallin are the only available sequences yet for pangolins. αA-crystallin alone placed pangolin within an unresolved ensemble of artiodactyls, cetaceans, carnivores, and perissodactyls (de Jong and Goodman 1982; de Jong, Leunissen, and Wistow 1993). It actually was depicted as the sister group of bears because of a possibly homoplastic replacement. However, pangolin was clearly far removed from edentates, which branched off at the root of the eutherians. Similarly, the AQP2 data do not support a sister group relationship of pholidotes and edentates, in accordance with the considerable and growing morphological evidence that contradicts such a grouping (McKenna 1975; Rose and Emry 1993). However, the combined AQP2 and αA-crystallin evidence (fig. 6) does not significantly support clustering of pangolin with any other order either.

Other Superordinal Clades

Artiodactyls and perissodactyls are associated in the SMM/neighbor-joining and maximum-likelihood analyses of the AQP2 DNA sequences (fig. 5), as well as in maximum-parsimony and Kimura/neighbor-joining analyses of the combined AQP2/αA-crystallin protein sequences (fig. 6). Perissodactyls are also placed closest to the artiodactyl/cetacean clade in the IRBP analyses (Stanhope et al. 1996) and in earlier analyses of five to seven combined protein sequences (Czelusniak et al. 1990). Also, the maximum-parsimony analysis of the combined mitochondrially encoded proteins from species representing up to seven eutherian orders supports this grouping, but neighbor-joining analyses of these proteins and of the RNA-coding mitochondrial genes rather place perissodactyls and carnivores as sister groups (D’Erchia et al. 1996; Xu, Janke, and Arnason 1996). Neither grouping is in agreement with the prevailing morphological opinion, which places Perissodactyla closer to the Paenungulata (Novacek 1992; Prothero 1993).

Finally, it is remarkable that the cohort Glires gets some support in the analyses of the combined AQP2 and αA-crystallin sequences (fig. 6). Also, the IRBP data set gave some evidence for such a grouping of rodents and lagomorphs (Stanhope et al. 1996), which is otherwise strongly refuted by most molecular analyses (Graur, Duret, and Gouy 1996; D’Erchia et al. 1996).

This study shows that even a relatively short sequence—328 bp from the first exon of the AQP2 gene—may provide significant phylogenetic signal. It is a further indication that there is no reason anymore for Simpson’s (1945) pessimistic view that the eutherian radiation would be irresolvable. It is clear, however, that many more nuclear genes, as well as complete mitochondrial genomes, need to be analyzed in a sufficiently broad sampling of species to convincingly resolve the higher level mammalian phylogeny.

Acknowledgments

For tissue samples, we thank Drs. P. R. Klaster, Royal Tropical Institute, Amsterdam, The Netherlands (armadillo); J. Wensing, Burgers Zoo, Arnhem, The Netherlands (aardvark); G. Oilbright, Wuppertal Zoo, Germany (elephant shrew); E. Harley, South Africa (hyrax); G. Bernardi, Paris, France (pangolin); and the Noorder Zoo, Emmen, The Netherlands (elephant). We thank Rogier van Otterlo for technical assistance. This study was supported by grants from the European Commission (HCM CHR XC 930254) to W.W. de J. and C.S., and from the Dutch Kidney Foundation to PM.T.D. (93.1299).

LITERATURE CITED


DEEN, P. T. M., D. OLDE WEGHUIS, R. J. SINKE, A. GEURTS VAN KESSEL, B. WIERINGA, and C. H. VAN OS. 1994a. As-


DAN GRAUR, reviewing editor

Accepted December 20, 1996