Mammalian phylogeny studied by sequence analysis of the eye lens protein α-crystallin

W. W. de Jong and M. Goodman

Receipt of Ms. 3. 2. 1982

Abstract

Analyzed the amino acid sequences of the eye lens protein α-crystallin in 41 mammalian species with the aim to resolve phylogenetic relationships among mammalian orders. The species represented 17 orders of mammals. Chicken and frog *(Rana esculenta)* were included as outgroups. The observed amino acid differences were used to construct cladograms, either solely on the basis of the lowest numbers of required nucleotide replacements in the DNA, or also taking into account certain well-established phylogenetic relationships. The α-crystallin A sequences indicate that: the paenungulate orders Proboscidea, Hyracoidea and Sirenia are a monophyletic group to which the Tubulidentata (aardvarks) also belong; the paenungulates are not related to the ungulates, but together with the edentates represent the oldest offshoots of the eutherian stem; the pangolins show no relationship with edentates and are most parsimoniously placed close to the carnivores; the ungulates, whales, and carnivores form a monophyletic grouping; among the carnivores the seals and sea lions are monophyletic; the investigated bat (Microchiroptera) appeared not to be related to insectivores or primates. The α-crystallin A sequences left the rodents, lagomorphs, insectivores, primates and *Tupaia* as an unresolved cluster of orders, but within the primates the prosimians are clearly set apart from the Anthropoidea. The results are compared with current opinions about mammalian phylogeny and related to other comparative protein sequence data.

Introduction

Evolutionary trees can be constructed from genealogical analyses carried out by the parsimony method on the amino acid sequences of homologous proteins of different organisms. Such trees are capable of offering important insights on phylogeny despite the limitations of current algorithms (Dayhoff 1972; Peacock and Boulter 1975; Fitch 1977; Hendy et al. 1978; Goodman et al. 1979; Fitch 1979; Goodman 1981). Previously, an analysis of the amino acid sequences of the A chain of the eye lens protein α-crystallin from 17 mammalian species showed its usefulness in the study of mammalian phylogeny, especially at the ordinal level (De Jong et al. 1977). We now have extended the number of mammalian species from which α-crystallin A sequences have been determined to 41. The aim was to obtain information about the phylogenetic affinities among mammalian orders and about certain phylogenetic relationships within orders. Representatives from all eutherian orders, except Dermoptera, and from many of the major subordinal groups have therefore been studied. A considerable number of species whose phylogenetic affinities are not at stake, has been included to provide a framework in which to place the phylogenetic problem cases.

Evolutionary trees have been constructed from the αA sequence data using the chicken and frog *Rana esculenta* as outgroups. Our major conclusions are based on the most parsimonious trees found in a computer search which was not constrained in any way by evidence from other sources on the phylogenetic relationships of the 41 species represented by the sequences. This search was guided solely by those changes in branching arrangement of the trees which most lowered the number of nucleotide replacements (i.e. point mutations) required for the descent of the contemporary sequences from a common ancestor.

U.S. Copyright Clearance Center Code Statement: 0044-3468/82/4705-0257 $ 02.50/0
Z. Säugetierkunde 47 (1982) 257–276

ISSN 0044-3468 / InterCode: ZSAEA 7

Certain peculiar branching arrangements in the lowest nucleotide replacement (NR) length trees for the αA sequences, such as the grouping of marsupials with chicken rather than other mammals, do not agree with the results obtained in lowest NR length trees constructed for other proteins or with classical zoological evidence. To resolve this dilemma, we have used a newly developed parsimony procedure (Goodman et al. 1979), which incorporates phylogenetic evidence from other proteins or from classical taxonomic sources to decide whether the branching arrangement of a set of homologous sequences (in this case, αA sequences) should be made concordant with other phylogenetic evidence on the species represented by the sequences or should show gene duplications in some regions of the phylogeny. Obviously we only incorporated in the decision making process concepts on phylogenetic relationships which we considered firmly established such as the monophyly of therian mammals (in the species phylogeny marsupials should group with other mammals not with a chicken). Since almost nothing was assumed about mammalian relationships at such higher taxonomic levels as the interordinal, the parsimonious trees found by constraining the computer search with a limited number of a priori assumptions, lead to phylogenetic conclusions which were relatively independent of any prior bias. Our findings indicate that such a synthesis of classical taxonomy and amino acid sequence data may contribute to the eventual elucidation of the most probable course of events in mammalian evolution.

Material and method

α-crystallin A sequences

α-crystallin is a water-soluble structural protein which occurs exclusively in the epithelial and fiber cells of the vertebrate eye lens (for a review see Bloemendal 1981). It makes up a variable proportion of the total lens protein; the amount depends on the species and age of the animal. In many species α-crystallin constitutes 25 to 50 % of the total lens protein. α-crystallin forms large aggregates, of average molecular weight from 400,000 to 800,000, and is composed of two types of chains, αA and αB. The ratio of αA to αB chains varies between species, ranging from 9 : 1 in the kangaroo to 1 : 4 in the spiny dogfish. The αA and αB chains of the ox show 58 % homology between their amino acid sequences, thus reflecting an ancient common ancestry of their genes.

α-crystallin can easily be obtained in considerable quantities from most vertebrate species, and the amino acid sequence of the αA chain, 173 residues long in most species, is relatively simple to establish. The procedures involved in the isolation and sequence determination of the αA chains are described in de Jong and Terwindt (1976). In most cases αA chains were isolated from pooled lenses from different specimens of the same species. Apart from hyrax αA chain (where both alanine and threonine were found at position 55), no polymorphisms were ever detected in the amino acid sequences of mammalian αA chains. The amino acid sequences of bovine, kangaroo, chicken and frog αA chains have been completely established by identifying all residues in these chains by the Edman-degradation method. Because only 10 % of the residues in the sequences of bovine and kangaroo αA chains are different, a simplified procedure has been used to study other mammalian αA chains. Their sequences have largely been deduced from the amino acid composition of small peptides obtained by enzymatic and chemical cleavage of the chains. When the amino acid composition of such peptides were found to be the same as those from the corresponding peptides of bovine or kangaroo αA chain, it was assumed that their sequences were also identical. When a difference in composition was found, such a peptide was usually subjected to Edman-degradation in order to confirm the position and type of the underlaying substitution.

Using this approach there is a risk of overlooking double substitutions which change the sequence but not the composition of a peptide. It has been established, however, that the risk of overlooking such reciprocal substitutions is extremely small if the analyzed peptides are small (van Druten et al. 1978).

The choice of species to investigate was directed by their phylogenetic relevance; species were selected either because they pose particular taxonomic problems or to increase the denseness of the phylogenetic tree in appropriate places. An important limiting factor was obviously the availability of the desired lens material. For this reason several interesting taxa have not yet been studied. The names and sources of all mammalian species of which the αA chain sequences have now been determined are given in Table 1. As already mentioned, 17 of these sequences had been employed in a previous study (de Jong et al. 1977). Details of the chemical determination of the 24 new mammalian αA sequences added to the present study are being described elsewhere (de Jong et al. in preparation). Certain

findings concerning trends in amino acid substitutions in the reconstructed phylogeny of mammalian αA sequences have already been reported (DE JONG et al. 1980).

The observed differences between αA chain sequences are summarized in Table 2. From inspection of the sequence differences it is easy to identify at certain positions residues which apparently are ancestral, primitive ones, and others which are derived ones. For example, at position 3 Ile occurs both in the outgroups (frog and chicken) and in the marsupials, whereas Val only occurs in several eutherian orders. It thus seems likely that at this position Ile is the primitive and Val the derived character. Similarly, at position 4 Thr seems to be primitive and Ala derived. Also some shared derived (synapomorphous) amino acid substitutions in certain species can easily be recognized, as for instance 13 Pro and 61 Val in the prosimians lemur, potto and galago. Table 2 also shows, however, that certain substitutions, such as 55 Ser and 61 Val, must have occurred more than once in entirely unrelated taxa, and that back substitutions, for instance 101 Asn → Ser, may complicate the interpretation. Because convergent substitutions and back substitutions occur frequently, it is obvious that rigourous computer handling of the data is required to assess objectively the numerous possible branching patterns.

Construction of cladograms

A maximum parsimony approach (MOORE et al. 1973; GOODMAN et al. 1979) was employed to construct cladograms for the 43 α-crystallin A chain sequences. In this approach the contemporary amino acid sequences, i.e. the operational taxonomic units (OTUs), are mapped, through the genetic code, into corresponding messenger ribonucleic acid (mRNA) sequences. The object then is to find an ancestral order of branching and ancestral mRNA sequences which account for the descent of the OTUs by the fewest possible nucleotide replacements (NRs). Such a parsimony tree maximizes the number of nucleotide identities among descendant sequences ascribable to shared common ancestry rather than to convergence or parallelisms and back mutations.

As indicated in the Introduction, it has been found (e.g. GOODMAN et al. 1979; MAEDA and FITCH 1981) that the trees with the fewest NRs constructed from different proteins can yield non-concordant branching arrangements for the same animal species and thus violate in each tree some of the features of the animal phylogeny strongly supported by the evidence from other proteins. Such violations could be indicative of incorrect groupings of sequences that happen to have an excess of convergent residues. Alternatively, such violations could be due to real differences in the branching arrangement between the gene phylogeny and the species phylogeny. This latter possibility indeed opens the way for the construction of more accurate genealogical trees by an extension of the parsimony criterion. Not only are base replacements counted, but also the additional cost in gene duplication (GD) and gene expression (GE) events that must be assumed to fit the putative gene phylogeny into well established features of the species phylogeny. The object is to minimize the total NR + GD + GE length. (Computer algorithms for counting the number of GDs and GEs are described in Appendix A-2 and A-3 of GOODMAN et al. 1979.)

Lowest NR length trees

For a set of contemporary amino acid sequences the only sure way to find the tree or trees of lowest NR length is by examining all unrooted trees that the OTUs can possibly form and then choosing the tree or trees with the lowest score. Unfortunately this method is prohibitive in computer time when there are more than 8 or 9 OTUs in the data set. For larger sets of data, such as that employed in the present study, heuristic approaches can be used which limit the search procedure to practical dimensions.

We started the search by calculating a matrix of minimum mutation distances for the 43 α-crystallin A sequences by the method of FITCH and MARGOLIASH (1967). Then two initial dendrograms were constructed from this matrix, the distance Wagner tree by the method of FARRIS (1972) and the unweighted pair group tree by the clustering algorithm of SOKAL and MICHENER (1958). With the initial dendrograms and with the computer file of α-crystallin A sequences, we employed the maximum parsimony branch swapping algorithm described in Appendix A-1 of GOODMAN et al. (1979) to search for the lowest NR length trees. In order to test a wide range of phylogenetic possibilities the search was continued using 33 further starting dendrograms, from which literally thousands of alternative dendrograms were examined in the progression of branch swaps. The trees found with the lowest NR score each required 152 NRs. Representative examples of the major different branching arrangements exhibited by these trees are shown in Figs 1–3. In a few local regions of each of these three trees minor changes in the branching arrangement also yield trees of the lowest NR length, 152 NRs.

The lowest NR length tree shown in Fig. 1 differs from the distance Wagner tree only in the position of the bear (with the whale-porpoise branch in the Wagner tree and with the pangolin in Fig. 1) and in containing one less NR. The distance Wagner tree itself costs 153 NRs. The NR score of this Wagner tree could not be lowered when the Wagner tree itself was the starting dendrogram, i.e.

Table 1

Names of species from which the α-crystallin A sequences have been studied, and sources of the investigated lens material

Scientific name	Common name	Number of lenses	Source
Cetacea			
Balaenoptera acutorostrata	Minke whale	5	Mr. I. Christensen, Institute of Marine Research, Bergen, Norway
Phocaena phocaena	Common porpoise	5	Dr. C. Smeenk, Rijksmuseum van Natuurlijke Historie, Leiden, Neth., and Dr. P. van Bree, Inst. Taxon. Zoology, Univ. of Amsterdam
Perissodactyla			
Equus caballus	Horse	19	Municipal slaughterhouse, Utrecht, Neth.
Tapirus indicus	Malayan tapir	2	Dr. L. de Boer, Blijdorp Zoo, Rotterdam
Ceratotherium simum	White rhinoceros	6	Mr. M. Keep, Hluhluwe Game Reserve, South Africa
Artiodactyla			
Sus scrofa	Pig	50	Central Animal Facilities, Univ. of Nijmegen
Giraffa camelopardalis	Giraffe	2	Safaripark Beekse Bergen, Hilvarenbeek, Neth.
Hippopotamus amphibius	Hippopotamus	2	Dr. P. van Bree, Inst. Taxon. Zool., Univ. of Amsterdam
Bos taurus	Ox	many	Central Animal Facilities, Univ. of Nijmegen
Camelus dromedarius	Dromedary	2	Dr. R. Yagil, Univ. of the Negev, Beer-Sheva, Israel
Carnivora			
Canis familiaris	Dog	10	Central Animal Facilities, Univ. of Nijmegen
Felis catus	Cat	8	Central Animal Facilities, Univ. of Nijmegen
Melursus ursinus	Sloth bear	2	Dr. P. van Bree, Inst. Taxon. Zool., Univ. of Amsterdam
Mustela vison	American mink	50	Central Animal Facilities, Univ. of Nijmegen
Halichoerus grypus	Gray seal	3	Dr. J. van Haaften, Rijksinstituut voor Natuurbeheer, Arnhem, Neth.
Zalophus californianus	California sea lion	2	Dr. L. Cornell, Seaworld, San Diego, California
Pholidota			
Manis javanica	Malayan pangolin	4	Dr. P. van Bree, Inst. Taxon. Zool., Univ. of Amsterdam
Manis (Phataginus) tricuspis	Tree pangolin[1]	4	Dr. K. Joysey, Univ. Museum of Zoology, Cambridge, U.K.
Chiroptera			
Artibeus jamaicensis	Jamaican fruit-eating bat	20	collected by W. W. de J. in Panama Canal Zone
Insectivora			
Erinaceus europaeus	European hedgehog	42	Dr. W. Peteks, traffic casualties around Nijmegen
Scandentia			
Tupaia belangeri	Treeshrew	20	Dr. A. Schwaier, Battelle Institut, Frankfurt am Main
Rodentia			
Rattus norvegicus	Rat	100	Central Animal Facilities, Univ. of Nijmegen
Mesocricetus auratus	Golden hamster	100	Central Animal Facilities, Univ. of Nijmegen

Primates			
Lemur fulvus	Brown lemur	6	Mr. D. Anderson, Duke Primate Facility, Durham, N.C.
Galago crassicaudatus	Galago	4	Mr. D. Anderson, Duke Primate Facility, Durham, N.C.
Perodicticus potto	Potto	4	Prof. M. Goffart, Univ. of Liège, Belgium
Macaca mulatta	Rhesus monkey	10	Central Animal Facilities, Univ. of Nijmegen
Homo sapiens	Human	14	Dept. of Anatomy, Univ. of Nijmegen School of Medicine
Proboscidea			
Loxodonta africana	African elephant	11	Dr. U. de V. Pienaar, Kruger National Park, South Africa
Hyracoidea			
Procavia capensis	Cape hyrax	30	Dr. V. de Vos, Kruger National Park, South Africa
Sirenia			
Trichechus inunguis	Brazilian manatee	6	Dr. R. Best, Inst. Nac. Pesq. da Amazonia, Manaus, Brazil
Tubulidentata			
Orycteropus afer	Aardvark	3	Dr. P. van Bree, Inst. Taxon. Zool., Univ. of Amsterdam, and Mr. J. Shoshani (Detroit) collected in South Africa
Edentata			
Choloepus hoffmanni	Two-toed sloth	12	collected by W. W. de J. in Panama Canal Zone
Bradypus variegatus	Three-toed sloth	16	collected by W. W. de J. in Panama Canal Zone
Tamandua mexicana	Ant bear	2	Dr. G. Montgomery, Smithsonian Trop. Res. Inst., Panama
Marsupialia			
Macropus rufus	Red kangaroo	120	commercial hunter, Australia
Didelphis marsupialis	North American opossum	36	Dr. P. Stenzel, Dept. of Biochemistry, Univ. of Oregon, Portland, Oregon
Aves			
Gallus gallus	Chicken	200	Central Animal Facilities, Univ. of Nijmegen
Amphibia			
Rana esculenta[2]	Frog	150	Central Animal Facilities, Univ. of Nijmegen

[1] The amino acid sequence of the αA chain of the tree pangolin has only partially been determined. – [2] At the time of collection of these frogs the distinction between the *Rana esculenta* complex and *R. lessonae* had not yet been made.

Table 1

Names of species from which the α-crystallin A sequences have been studied, and sources of the investigated lens material

Scientific name	Common name	Number of lenses	Source
Cetacea			
Balaenoptera acutorostrata	Minke whale	5	Mr. I. Christensen, Institute of Marine Research, Bergen, Norway
Phocaena phocaena	Common porpoise	5	Dr. C. Smeenk, Rijksmuseum van Natuurlijke Historie, Leiden, Neth., and Dr. P. van Bree, Inst. Taxon. Zoology, Univ. of Amsterdam
Perissodactyla			
Equus caballus	Horse	19	Municipal slaughterhouse, Utrecht, Neth.
Tapirus indicus	Malayan tapir	2	Dr. L. de Boer, Blijdorp Zoo, Rotterdam
Ceratotherium simum	White rhinoceros	6	Mr. M. Keep, Hluhluwe Game Reserve, South Africa
Artiodactyla			
Sus scrofa	Pig	50	Central Animal Facilities, Univ. of Nijmegen
Giraffa camelopardalis	Giraffe	2	Safaripark Beekse Bergen, Hilvarenbeek, Neth.
Hippopotamus amphibius	Hippopotamus	2	Dr. P. van Bree, Inst. Taxon. Zool., Univ. of Amsterdam
Bos taurus	Ox	many	Central Animal Facilities, Univ. of Nijmegen
Camelus dromedarius	Dromedary	2	Dr. R. Yagil, Univ. of the Negev, Beer-Sheva, Israel
Carnivora			
Canis familiaris	Dog	10	Central Animal Facilities, Univ. of Nijmegen
Felis catus	Cat	8	Central Animal Facilities, Univ. of Nijmegen
Melursus ursinus	Sloth bear	2	Dr. P. van Bree, Inst. Taxon. Zool., Univ. of Amsterdam
Mustela vison	American mink	50	Central Animal Facilities, Univ. of Nijmegen
Halichoerus grypus	Gray seal	3	Dr. J. van Haaften, Rijksinstituut voor Natuurbeheer, Arnhem, Neth.
Zalophus californianus	California sea lion	2	Dr. L. Cornell, Seaworld, San Diego, California
Pholidota			
Manis javanica	Malayan pangolin	4	Dr. P. van Bree, Inst. Taxon. Zool., Univ. of Amsterdam
Manis (Phataginus) tricuspis	Tree pangolin[1]	4	Dr. K. Joysey, Univ. Museum of Zoology, Cambridge, U.K.
Chiroptera			
Artibeus jamaicensis	Jamaican fruit-eating bat	20	collected by W. W. de J. in Panama Canal Zone
Insectivora			
Erinaceus europaeus	European hedgehog	42	Dr. W. Peters, traffic casualties around Nijmegen
Seandentia			
Tupaia belangeri	Treeshrew	20	Dr. A. Schwaier, Battelle Institut, Frankfurt am Main
Rodentia			
Rattus norvegicus	Rat	100	Central Animal Facilities, Univ. of Nijmegen
Mesocricetus auratus	Golden hamster	100	Central Animal Facilities, Univ. of Nijmegen
Meriones unguiculatus	Mongolian gerbil	27	Central Animal Facilities, Univ. of Nijmegen
Ochotona princeps	American pika	28	collected by Dr. M. McKenna in Colorado
Oryctolagus cuniculus	Rabbit	12	local butcher, Nijmegen
Primates			
Lemur fulvus	Brown lemur	6	Mr. D. Anderson, Duke Primate Facility, Durham, N.C.
Galago crassicaudatus	Galago	4	Mr. D. Anderson, Duke Primate Facility, Durham, N.C.
Perodicticus potto	Potto	4	Prof. M. Goffart, Univ. of Liège, Belgium
Macaca mulatta	Rhesus monkey	10	Central Animal Facilities, Univ. of Nijmegen
Homo sapiens	Human	14	Dept. of Anatomy, Univ. of Nijmegen School of Medicine
Proboscidea			
Loxodonta africana	African elephant	11	Dr. U. de V. Pienaar, Kruger National Park, South Africa
Hyraeoidea			
Procavia capensis	Cape hyrax	30	Dr. V. de Vos, Kruger National Park, South Africa
Sirenia			
Trichechus unguis	Brazilian manatee	6	Dr. R. Best, Inst. Nac. Pesq. da Amazonia, Manaus, Brazil
Tubulidentata			
Orycteropus afer	Aardvark	3	Dr. P. van Bree, Inst. Taxon. Zool., Univ. of Amsterdam, and Mr. J. Shoshani (Detroit) collected in South Africa
Edentata			
Choloepus hoffmanni	Two-toed sloth	12	collected by W. W. de J. in Panama Canal Zone
Bradypus tarigatus	Three-toed sloth	16	collected by W. W. de J. in Panama Canal Zone
Tamandua mexicana	Ant bear	2	Dr. G. Montgomery, Smithsonian Trop. Res. Inst., Panama
Marsupialia			
Macropus rufus	Red kangaroo	120	commercial hunter, Australia
Didelphis marsupialis	North American opossum	36	Dr. P. Stenzel, Dept. of Biochemistry, Univ. of Oregon, Portland, Oregon
Aves			
Gallus gallus	Chicken	200	Central Animal Facilities, Univ. of Nijmegen
Amphibia			
Rana esculenta[2]	Frog	150	Central Animal Facilities, Univ. of Nijmegen

[1] The amino acid sequence of the αA chain of the tree pangolin has only partially been determined. – [2] At the time of collection of these frogs the distinction between the *Rana esculenta* complex and *R. lessonae* had not yet been made.

Table 2

Substitutions among the α-crystallin A chains of 41 mammalian species, chicken *(Gallus gallus)* and frog *(Rana esculenta)*

Only those positions in the αA chains are shown at which substitutions have been found in at least two species. References to the sequence analyses of the αA chains of whale, horse, rhinoceros, pig, ox, dog, cat, hedgehog, treeshrew, rat, rabbit, monkey, human, elephant, hyrax, kangaroo and opossum are given in DE JONG et al. (1977). The other sequence determinations will be described elsewhere (DE JONG et al., in prep.).

POSITION NR	3	4	7	13	16	17	20	37	51	52	55	56	58	61	70	72	74	89	90	91	93	101	122	123	126	127	129	133	142	146	147	148	149	150	152	153	155	156	158	168	169	172	a)
MINKE WHALE	I	A	H	A	P	F	S	L	S	L	T	V	D	I	K	V	F	V	Q	E	F	N	S	N	Q	S	L	L	S	V	P	S	G	M	A	G	S	E	A	S	S	S	
PORPOISE											S																				T												
HORSE																										T		V		I										G			
TAPIR				T																						T								L									
RHINOCEROS				T							S			V												T				I													
PIG														V																				V								T	
GIRAFFE, HIPPOPOTAMUS																														I				V									
OX				T																										I				V									
CAMEL																			L											I				V									
DOG, CAT																			L															V									
BEAR																	Y		L																								
MINK			Q					M						V					L	Q														V									
SEAL, SEA LION									P	V	S								L															V									
PANGOLIN																	Y		L		S	S												V	T	S							
BAT		T																												L	Q			V	G								
HEDGEHOG	V	T																	L			S	P	S		A					A			L									2
TUPAIA	V	T																	L												Q			L									
RAT, HAMSTER, GERBIL	V	T																	L												Q			L									
GUINEA PIG, SPRINGHAAS	V	T																													Q			L									
PIKA	V	T																				S									Q			L									
RABBIT	V	T		T																											Q			L									
LEMUR	V	T		P										V																	Q			L									
GALAGO, POTTO	V	T		P										V														V			Q			L									
RHESUS MONKEY	V	T		T																D										I	Q	T		L		-	T						1
HUMAN	V	T		T																D									C	I	Q	T		L		-	T			T			
ELEPHANT	V	T													Q		L			D									C	I	Q					S							
HYRAX	V	T													Q	L	L		L	D									C		Q					S							
MANATEE	V	T					N								Q	L	L		L	D					K				C		Q					S						N	2
AARDVARK	V	T													Q	L	L		L	D									C		Q		S		D								2
2-TOED SLOTH	V	T																	L	D			T	A						I	-	-	-	V	P	S			T				1
3-TOED SLOTH	V	T	Q									A							L	G			T	A						I	-	-	-	V	P	S			P				1
TAMANDUA	V	T										A			R		L		L	D			T	A				.	.	L	-	-	-	V	P	S	G		T				
KANGAROO		T			S	L							E						L	D		S					I			I	H		D			S		D	S	T	L		
OPOSSUM		T			S	L							E		R		Y		L	D	Y	S					I			I	H	T	N		S	S		D	S	T	L		1
CHICKEN		T				L					S		E			.	.	I	I	D		S	A				I						D		P	S			P	T			7
FROG		T					N	F		F	G	F			R	.	.	I	L	D		S			E		I			L	M		S	L	S	S	G		P	T			7

The one-letter notation for amino acids has been used: A = alanine; C = cysteine; D = aspartic acid; E = glutamic acid; F = phenylalanine; G = glycine; H = histidine; I = isoleucine; K = lysine; L = leucine; M = methionine; N = asparagine; P = proline; Q = glutamine; R = arginine; S = serine; T = threonine; V = valine; Y = tyrosine; (–) means deletion and (·) not determined. – a) Substitutions observed in only a single species: hedgehog 23 F → L, 154 H → P; rhesus monkey 162 S → A; manatee 19 P → H, 167 P → A; aardvark 84 D → E, 170 A → V; two-toed sloth 11 K → R; three-toed sloth 34 Y → S; opossum 151 D → E; chicken 18 Y → I, 32 F → L, 39 F → L, 40 L → F, 86 T → S, 130 S → T, 135 A → S; frog 26 F → V, 31 L → M, 33 E → D, 124 V → L, 125 D → N, 128 A → S, 138 M → I.

the single step changes in branch arrangement at adjacent nodes made by the branch swapping algorithm failed to find a tree of lower NR score. We noticed, however, that bear and pangolin αA chains are unique among eutherian mammals in having 74 Tyr (Table 2). Thus by joining bear to pangolin we found the tree (Fig. 1) at 152 NRs, which may be the lowest NR score for this set of sequences since further searching has not resulted in any lower score.

Lowest NR+GD+GE length trees

In order to find the trees of lowest NR+GD+GE length we first specified the a priori cladistic relationships for which there is strong classical zoological evidence or sizeable evidence from the sequences of other molecules such as α-hemoglobin, β-hemoglobin, and myoglobin. This search was aided by a modification of the branch swapping algorithm described in section E of Appendix A-1 in

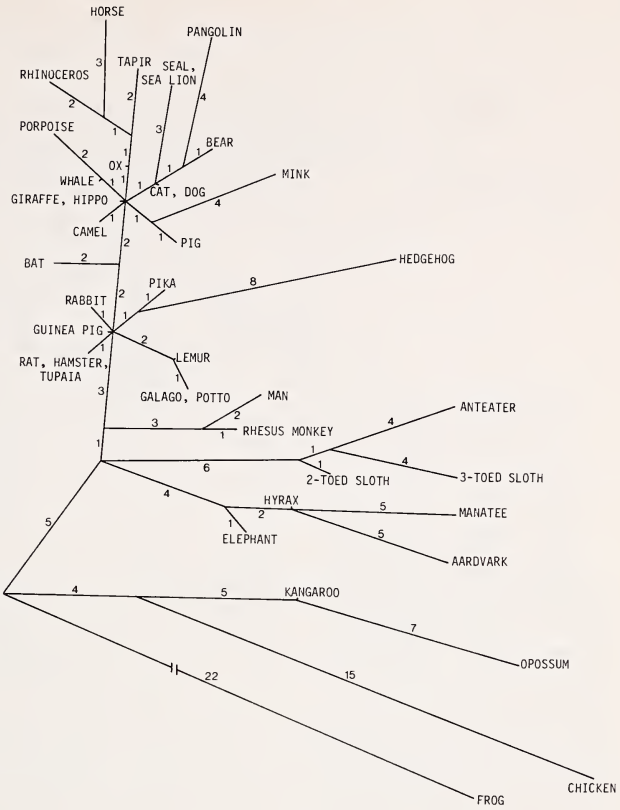

Fig. 1. Lowest NR length tree (152 NRs) of α-crystallin A sequences obtained on starting the search from the distance Wagner tree. Unaugmented NR values are given for all branches in this tree and the following ones. A computer algorithm was used which selected a particular set of most parsimonious ancestral residues, called the A-solution set (GOODMAN et al. 1974). Wherever possible, these A-solution residues had the nucleotide replacements fall on terminal links to contemporary species rather than between ancestral nodes. By this means, the synapomorphous residues characterizing each monophyletic grouping of the proposed cladograms are kept to a minimum. Springhaas and gerbil are not shown in Figs. 1–3, but are identical in position to guinea pig and rat, respectively

GOODMAN et al. (1979). The modification allows one to designate nodes within the tree as though they were the roots of monophyletic subtrees; it then prevents branch swaps across each particular edge (the link connecting two adjacent nodes) which joins such an hypothesized monophyletic subtree to the rest of the tree. This modification helped us find trees of lowest NR score that did not require hypothetical GDs and GEs. We could then evaluate whether the NR score of these trees was less than the NR + GD + GE scores of the lowest NR length trees (those in Figs 1–3), or conversely whether it would be more parsimonious to have some paralogous gene lineages rather than have only orthologous lineages.

Our a priori cladistic assumptions were:

- all mammals are more closely related to each other than to chicken or frog;
- the grouping of the investigated species into their respective traditional orders (e.g. rabbit and pika in Lagomorpha) is accepted, with tree shrew in its own higher taxonomic group (order Scandentia) rather than assigned, as it sometimes has been, to either Primates or Insectivora;
- the classical intra-order relationship within the Edentata is adhered to (i.e. Bradypodidae and Myrmecophagidae are each considered monophyletic groups).

The main freedom then allowed in the search for parsimonious trees that did not require hypothetical GDs and GEs was the reshuffling of the mammalian orders in relation to each other as well as appreciable reshuffling of species within their orders. The two major trees of lowest NR score that required no GDs and GEs are shown in Figs 4 and 5. Each had a score of 157 NRs. Equally parsimonious variants placed rhinoceros first with horse rather than with tapir. These alternative positions of rhinoceros were also found for the 152 NR length trees, and are caused by the sharing of residue 13 Thr by tapir and rhinoceros, and residue 146 Ile by horse and rhinoceros (Table 2).

Before we could conclude that the trees with score 157 NR + 0 GD + 0 GE were the lowest NR + GD + GE length trees found we had to determine the numbers of GDs and GEs in the 152 NR length trees, i.e. we had to evaluate the score in total genetic events of these trees. By our criteria, the trees shown in Figs 2 and 3 are clearly more parsimonious than the tree shown in Fig. 1 because they violate fewer of the a priori assumed relationships. Nevertheless, these Figs 2 and 3 trees still each require, in addition to the 152 NRs, 5 GDs and 15 GEs. The extra genetic events are needed to account

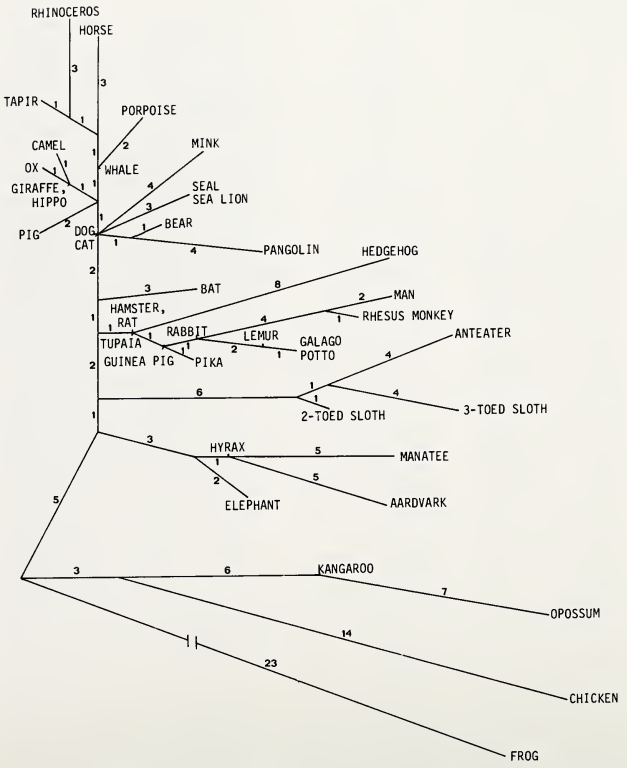

Fig. 2. Lowest NR length tree (152 NRs) of α-crystallin A sequences obtained on starting with phylogenetically plausible dendrograms and carrying out the search by maximum parsimony branch swapping, not constrained by a priori cladistic assumptions

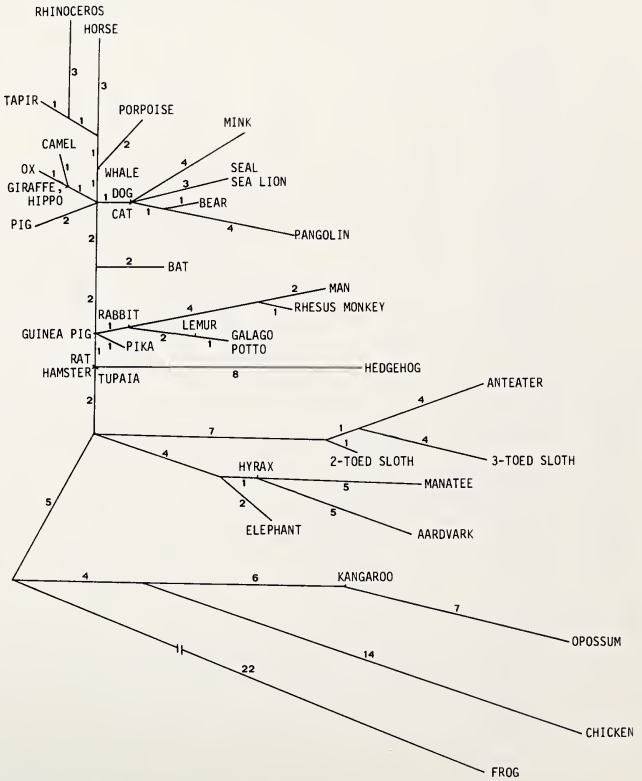

Fig. 3. Alternative 152 NR tree of α-crystallin A sequences

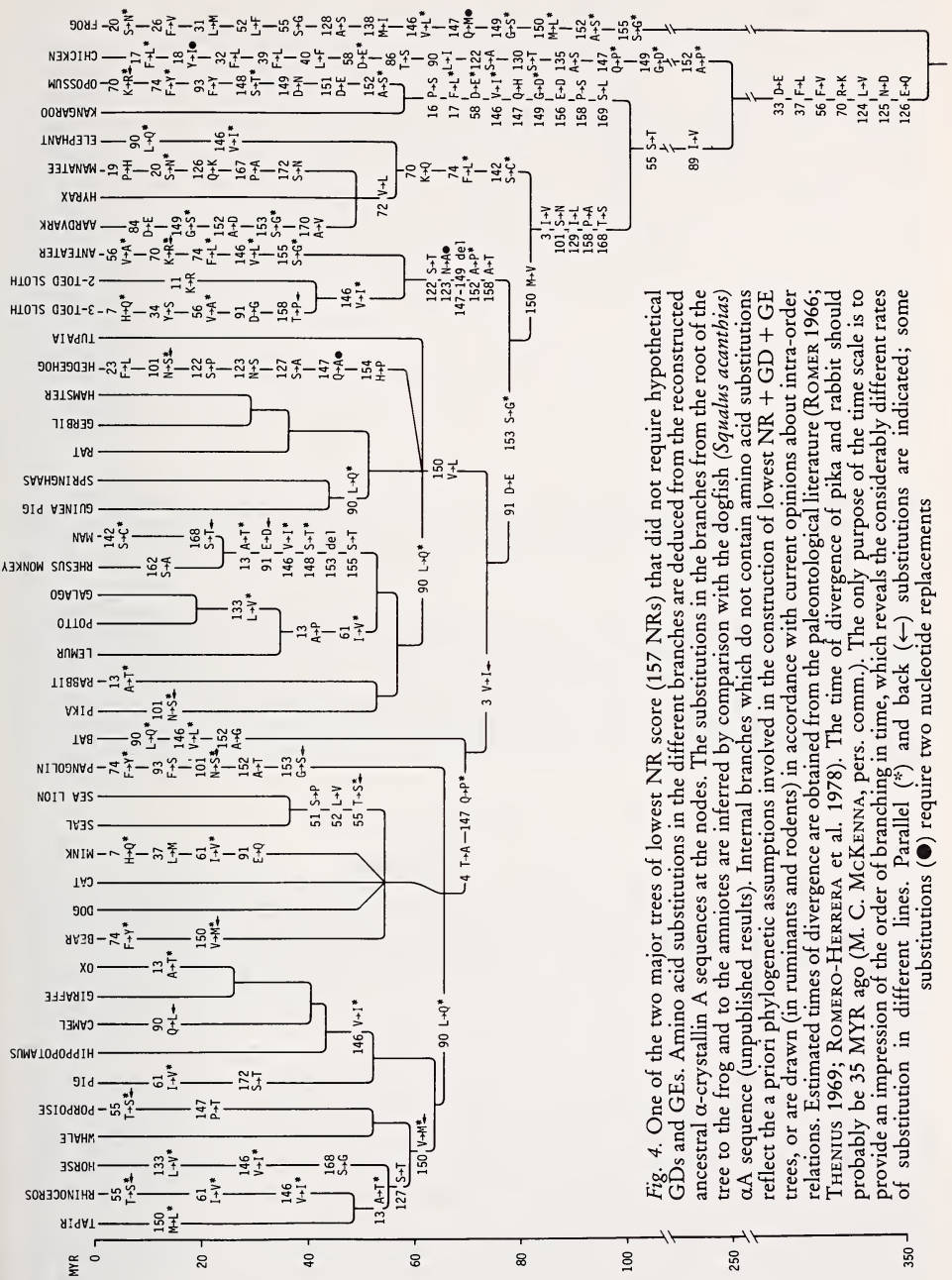

Fig. 4. One of the two major trees of lowest NR score (157 NRs) that did not require hypothetical GDs and GEs. Amino acid substitutions in the different branches are deduced from the reconstructed ancestral α-crystallin A sequences at the nodes. The substitutions in the branches from the root of the tree to the frog and to the amniotes are inferred by comparison with the dogfish (*Squalus acanthias*) αA sequence (unpublished results). Internal branches which do not contain amino acid substitutions reflect the a priori phylogenetic assumptions involved in the construction of lowest NR + GD + GE trees, or are drawn (in ruminants and rodents) in accordance with current opinions about intra-order relations. Estimated times of divergence are obtained from the paleontological literature (Romer 1966; Thenius 1969; Romero-Herrera et al. 1978). The time of divergence of pika and rabbit should probably be 35 MYR ago (M. C. McKenna, pers. comm.). The only purpose of the time scale is to provide an impression of the order of branching in time, which reveals the considerably different rates of substitution in different lines. Parallel (*) and back (←) substitutions are indicated; some substitutions (●) require two nucleotide replacements

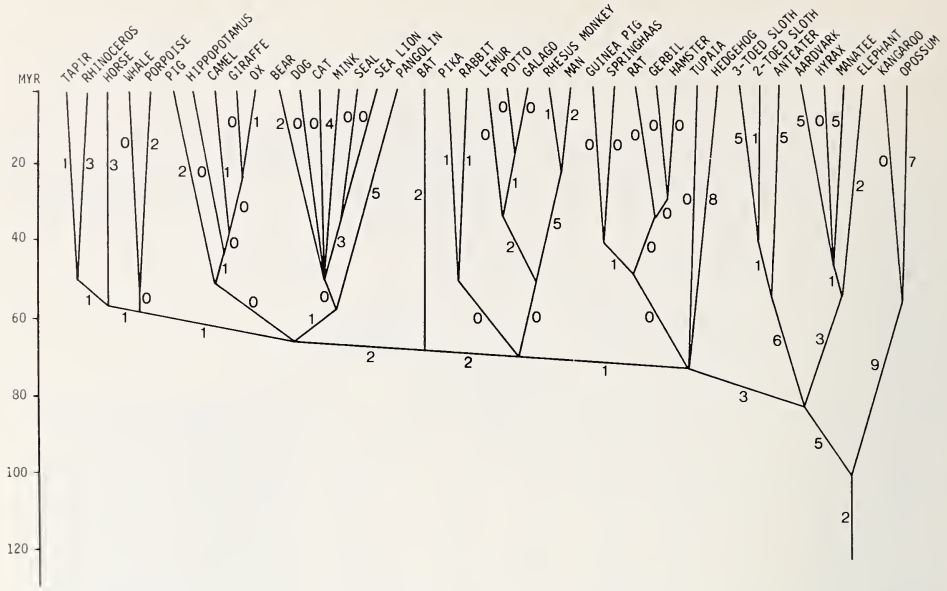

Fig. 5. Alternative major tree requiring 157 NRs and no GDs and GEs. Only the numbers of NRs are indicated for the different branches; internal 0-links are due to a priori phylogenetic assumptions. The chicken and frog branches (not shown) are as in Fig. 4

for: 1) the marsupial branch being closer to chicken than to the eutherian branch; 2) the three-toed sloth being closer to anteater than to the two-toed sloth; 3) rabbit being closer to Primates than to pika; 4) guinea pig and springhaas being closer to Lagomorpha and Primates than to Muroidea; 5) bear being closer to pangolin than to other carnivores. These violations of well accepted cladistic relationships each cost 1 GD + 3GEs. In contrast the orthologous arrangement can be had for only 1 NR more in each case. Thus it is indeed more parsimonious (total genetic score 157 compared to 172) to have the αA sequences in the orthologous branching arrangement to one another, as shown in Figs 4 and 5.

Results

Phylogenetic inferences

The most unbiased and therefore strongest inferences obtainable from the α-crystallin A sequences come from the lowest NR length trees (Figs 1–3). The constant features in these trees provide important indications of possible phylogenetic relationships. There are, however, several conspicuous deviations from generally accepted opinions about mammalian phylogeny. Most disturbing is the joining of chicken to the marsupial branch; in addition several species belonging to the same orders fail to group together, albeit often separated by the presence of only a single amino acid difference.

These shortcomings are overcome by the introduction of some a priori phylogenetic assumptions, resulting in the lowest NR+GD+GE trees shown in Figs 4 and 5. The introduction of such reasonable constraints may result in a more realistic branching pattern of those OTUs of which we would like to study the phylogenetic relationships and which have been left free in their positioning.

The differences in the branching patterns of equally parsimonious trees (Figs 1–3 and Figs 4–5) reveal the limitations of the method and indicate those regions of the tree where no firm phylogenetic conclusions can be made. The relative significance of a branching arrangement increases with the length (in NRs) of the branches involved. A 1 NR connection between two OTUs is more likely to be due to a parallel or back mutation than

a longer branch. Also the position and type of a substitution has some importance, for instance: the substitutions 13 Ala → Thr or 61 Ile → Val occur scattered over four orders (Table 2), and thus have less diagnostic value than 13 Pro, which only occurs in prosimians, or 127 Thr, which only occurs in Perissodactyla. We therefore have included in Fig. 4 the positions and directions of all inferred amino acid substitutions, and indicated all instances of supposed parallel or back mutations.

In the following sections we will successively discuss the major branching characteristics of the α-crystallin A cladograms in relation to prevailing ideas on mammalian phylogeny; i.e., the phylogenetic information which can be deduced from these cladograms is judged and weighed in relation to the prevailing opinions about mammalian phylogeny. For this purpose we have relied mainly upon some recent authorative works (Simpson 1945; Romer 1966; Thenius 1969; van Valen 1971; Hoffstetter 1973; McKenna 1975; Szalay 1977). We have also taken into account evidence from other comparative biochemical studies of myoglobin (Romero-Herrera et al. 1978), hemoglobin (Goodman et al. 1979), pancreatic ribonuclease (Beintema and Lenstra 1982), cytochrome *c* (Foulds et al. 1979; Baba et al. 1981) and fibrinopeptides (O'Neil and Doolittle 1974; Dayhoff 1972; Goodman 1981). The immunological comparisons of mammalian albumins are also relevant (Sarich 1976, 1982).

Monophyly of therian mammals

There seems to be little reason to doubt the monophyletic origin of the eutherian and metatherian mammals from an advanced synapsid reptilian stem in the Upper Triassic (Marshall 1979). Nevertheless our lowest NR trees (Figs 1–3) depict marsupials and birds as sister groups. A single additional NR restores the therian monophyly, but it is still amazing that no greater number of synapomorphies should have accumulated in the estimated 200 million years of supposed evolution of the common therian ancestor. A marked reduction in the number of amino acid substitutions in the mammalian ancestral stem has also been observed in myoglobin (Romero-Herrera et al. 1978; Goodman 1981), hemoglobin (Goodman 1981) and cytochrome *c* (Baba et al. 1981). However, the lowest NR length trees for globin sequence data support the monophyly of therian mammals.

Metatherian-eutherian divergence

Both in the lowest NR and lowest NR + GD + GE trees the marsupials are well separated from the placental mammals, each being characterized by a convincing number of apomorphous (unique) substitutions. This agrees well with traditional taxonomic evidence that Metatheria and Eutheria are each monophyletic. However, this therian dichotomy is not always reflected in the lowest NR length trees constructed for globins and cytochrome *c* sequences.

Edentate relationships

The three investigated edentate species (2-toed sloth, 3-toed sloth and tamandua) share 5 or 6 substitutions and a rare deletion of 3 residues, clearly reflecting a monophyletic origin. The phylogenetic unity of the South American edentates (= Xenarthra) has indeed never seriously been questioned (Engelmann 1982) The αA trees depict the edentates as one of the oldest offshoots of the eutherian stem. The trichotomy in Fig. 5 indicates that the αA sequences cannot discriminate between three possibilities: the edentates are the oldest eutherian branch and the paenungulates the next oldest; the paenungulates are the oldest eutherian branch and the edentates the next oldest; the edentates and paenungulates are sister groups and together constitute the oldest eutherian branch.

The relationship of the edentates to the other mammals is obscure, although pangolins and aardvarks have been considered as relatives (reviews by Glass 1982; Engelmann

1982). Placement of the Edentata within the Eutheria appears to be preferred. It has recently been hypothesized that the Edentata are a sister group to all other eutherians (McKenna 1975; Engelmann 1982), which is compatible with the α-crystallin data.

The two investigated sloth genera *Bradypus* and *Choloepus* are grouped together in the Bradypodidae, although it recently has been recognized that many similarities may be regarded as convergent features and important differences have been noted (Engelmann 1982). It is therefore not surprising that the two sloth αA chains show no synapomorphies in the lowest NR trees, and that *Bradypus* αA chain joins the anteater branch on the basis of the unique substitution 56 Val → Ala. There is, however, little need to conjecture that the sloths represent a grade rather than a clade, because one additional NR restores the accepted division between Bradypodidae and Myrmecophagidae.

The only other edentate protein sequence data are of sloth ribonuclease (Beintema and Lenstra 1982) and armadillo β-hemoglobin (de Jong et al. 1982). Sloth ribonuclease, as compared with rodent, whale, ungulate and kangaroo ribonucleases, indicates that sloths are on the oldest branch among the eutherians. Armadillo Hbβ joins in the most parsimonious solutions the branch leading to elephant Hbβ, but does not specifically appear as the oldest eutherian branch. Immunological comparisons of albumin supported the traditional grouping of three-toed and two-toed sloths in Bradypodidae and of anteater in Myrmecophagidae, but did not reveal the position of edentates among the eutherians, although they were especially distant from ungulates and rabbits (Sarich 1982). The immunological distances between sloths, anteaters and armadillos suggests that the period of edentate monophyly has been rather brief, and that they have diverged from each other at about the same time as bats, primates and carnivores diverged. Preliminary results did not indicate a special relationship of any edentate to either pangolin or aardvark.

Aardvark-paenungulate relationships

From both Table 2 and the computer-constructed cladograms it can be seen that elephant, hyrax, manatee and aardvark, as compared to all other investigated mammals, share 3 to 4 apparently derived substitutions, of which 70 Lys → Gln and 72 Val → Leu are unique for this group, while 74 Phe → Leu and 142 Ser → Cys also occur as supposedly parallel substitutions in *Tamandua* and *Homo*, respectively.

Most authors agree on the grouping together of Proboscidea, Sirenia and Hyracoidea in the superorder Paenungulata, although Simpson (1945) as he proposed this grouping, clearly stated that superorders are theoretical constructions, not necessarily reflecting common origins. In fact Sirenia and Proboscidea are usually considered to be most closely related among the three orders. To bring elephant and manatee together on the same branch in Figs. 1–5 would require one additional NR, most likely a back substitution 72 Leu → Val in elephant. It has been proposed that Hyracoidea might be more closely related to Perissodactyla than to Proboscidea and Sirenia (McKenna 1975). Removing hyrax from the paenungulates (Fig. 4) and placing it with Perissodactyla costs at least 10 additional NRs, and therefore seems unjustified. Similarly, bringing all paenungulates together with Perissodactyla (van Valen 1971; Szalay 1977) costs 5 additional NRs and therefore also seems improbable. Paenungulate monophyly is further supported by immunological crossreactivity between hyrax and elephant (Weitz 1953).

Most significant is the strong connection of the aardvark αA sequence to those of the paenungulates. The aardvark is clearly a phylogenetic problem case. The hypothesis that aardvark is related to edentates and pangolins seems to have been largely abandoned. Indeed, to separate the aardvark αA chain from the paenungulates and connect it, without GDs and GEs, to the edentate branch would cost an additional 4 NRs. It is now usually suggested that the Tubulidentata originated from a condylarthran stem, and thus should be most closely related to the ungulates (Patterson 1978). Because the paenungulate orders

Proboscidea, Sirenia, and Hyracoidea are generally thought to be descended like the true ungulates from a common condylarth stock, the joining of aardvark and paenungulates is not incompatible with the current vague opinion that aardvark ancestry also traces back to a common condylarth stock. Morphological resemblances between aardvarks, hyraxes and elephants have been noted (LE GROS CLARK and SONNTAG 1926), and some recent immunological and osteological findings further support such a relationship (SHOSHANI et al. 1978). Furthermore, both Tubulidentata and the paenungulates apparently originated in Africa. Unfortunately no other protein sequence or immunological data from the aardvark are as yet available to help further elucidate its phylogenetic relationships.

Position of the Paenungulata

In all the most parsimonious trees of the α-crystallin A chains the enlarged Paenungulata (including Tubulidentata) are far apart from the true ungulates, in disagreement with the view of a common condylarth ancestry. Placing the paenungulates with Perissodactyla, which is its most parsimonious position if constrained to be in the ungulate-cetacean region of the tree, costs, as already mentioned, 5 NRs more than when it is separated from other eutherians as one of the most ancient branches. The early divergence of the paenungulates in these trees is mainly caused by the lack of some derived substitutions shared by most of the other eutherians: 91 Glu, 150 Leu or Val, and 153 Gly. The ancient branches of paenungulates and edentates can be changed to a more recent and simultaneous radiation of eutherian orders, at a cost of 2 NRs, by placing an edentate-paenungulate branch next to the branch of primates, insectivores, rodents and lagomorphs (cf. Fig. 4).

The only other paenungulate protein sequences known are elephant β hemoglobin, myoglobin and fibrinopeptides. In a recent analysis of the combined sequences of up to 7 proteins in 49 vertebrate taxa, the most parsimonious trees had the Paenungulata (represented by Proboscidea) originate as a separate branch in the earliest Eutheria (GOODMAN 1981). MCKENNA and MANNING (1977) have provided paleontological evidence for a very early origin of Proboscidea. Such evidence greatly reduces the likelihood of a common ungulate-paenungulate origin.

The insectivore-primate-rodent-lagomorph cluster

The lowest NR trees (Figs 2 and 3) fail to resolve the clustering of the αA sequences from these orders. This is mainly due to the paucity of substitutions in rodents, lagomorphs and early primates. To place the investigated rodents and lagomorphs in their respective orders, as has been done in Figs 4 and 5, requires 2 more NRs than the shortest trees. The need to do so, however, is supported by the combined sequence analysis (GOODMAN 1981) in which cytochrome *c*, hemoglobin, and fibrinopeptide A sequences override the αA sequence and cause guinea pig to group with Myomorpha. Similarly the recently completed pika myoglobin sequence clearly groups pika and rabbit together (DENE et al. unpublished data).

The few α-crystallin A substitutions which are present in rodents and lagomorphs and which might be used for tree-construction, are unreliable indicators of relationship, because the same substitutions are found frequently in other mammalian orders (see positions 13, 90, 101 and 150 in Table 2).

The frequency of substitutions is greater in later primate evolution and in the hedgehog line than in most other mammalian lines and brings to light the undisputed dichotomy between prosimians and Anthropoidea, and agrees with a lorisoid monophyly. The considerable differences between prosimians and Anthropoidea are also reflected in the myoglobin sequences (ROMERO-HERRERA et al. 1978). The common substitution 90 Leu → Gln in springhaas and guinea pig would be compatible with a grouping of the Pedetidae (springhaas) in the African rodent suborder Hystricomorpha, and the grouping

of this suborder with that of the guinea pig, i.e. with the South American Caviomorpha (Hoffstetter 1973; Lavocat 1978). Considerable support for a caviomorph-hystricomorph monophyly stems from ribonuclease sequence data (Beintema and Lenstra 1982).

Despite the inability to resolve fully the relationships of the lineages among the orders Insectivora, Primates, Rodentia, and Lagomorpha, the αA sequences are nevertheless useful in indicating that these four orders may have descended either from a common ancestor shortly after it separated from the major branch leading to carnivores and ungulates (Fig. 4) or separately from the stem of this carnivore-ungulate branch (Fig. 5), and after the earlier separation of edentates and paenungulates. As far as primates and insectivores are concerned either pattern of descent is not seriously at odds with the prevailing views on eutherian phylogeny. However, the analysis of up to seven combined polypeptide chains and of just combined α and β hemoglobin chains (see respectively Figs. 4 and 8 of Goodman 1981) supports the more distant ancestral separation of Insectivora (hedgehog) from Primates depicted in Fig. 5 rather than the closer one depicted in Fig. 4.

The prevailing views on eutherian phylogeny do not allow us to choose between the positioning of Rodentia near a primate-lagomorph branch depicted in Fig. 4 or the somewhat more ancient origin of Rodentia depicted in Fig. 5. In fact both lagomorphs and rodents appear as isolated groups of largely obscure origin, and their grouping together in the cohort Glires (Simpson 1945) still finds opponents as well as proponents (Szalay 1977).

A considerable body of protein data in addition to that provided by αA crystallin is available for investigating the phylogenetic relationships of Primates, Lagomorpha, Rodentia, and Insectivora. As yet, however, no consistent picture of their relationships emerges from it. It may well be that the periods of common ancestry between these orders and other eutherian branches were too short to have left their traces in the DNA and protein sequences so far available.

The position of Tupaia

The relationship of the tree shrews to primates or insectivores is a much discussed issue (Luckett 1980), and there are reasons to place them in the separate order Scandentia (Butler 1972). The *Tupaia* α-crystallin A sequence is identical to that of the investigated muroid rodents. This does not reflect a special relationship, but just the complete lack of fixed substitutions in both evolutionary lines. Because of this lack of change the *Tupaia* αA chain can equally parsimoniously be connected to the base of the rodent or insectivore lineage as to a common lagomorph-primate stem. Connecting it at the base of the primate line adds only 1 NR. The previously investigated *Tupaia* myoglobin and hemoglobin sequences certainly do not indicate a relationship with Primates, but rather in the most recent combined sequence analysis (Goodman 1981) cause *Tupaia* to group with Lagomorpha.

Chiroptera

The αA chain of the bat *Artibeus jamaicensis* is most parsimoniously connected to the branch leading to ungulates and carnivores, mainly due to the presence of residues 3 Ile and 150 Val. On the other hand the plesiomorphous residues 4 Thr and 147 Gln prevent it from being separated too far from the rodent-primate-insectivore-lagomorph cluster.

On the basis of morphological evidence the bats have been proposed to be derived from early insectivores (Romer 1966), or to be closely related to primates (Gregory 1910; Szalay 1977). To get a monophyletic bat-hedgehog or bat-primate branch in the α-crystallin tree would require an additional 2 NRs. The sequence of bat myoglobin shares some derived substitutions with that of the hedgehog (Castillo and Lehman 1977) and in the most parsimonious phylogenetic trees of myoglobin these species are shown as sister groups (Goodman et al. 1979). The minimal tree for vertebrate cytochrome *c* joins bat to

the carnivore branch, not to primates (an insectivore cytochrome *c* sequence is not yet available) (Foulds et al. 1979; Baba et al. 1981). Actually the myoglobin was obtained from a representative of the suborder Megachiroptera and cytochrome *c* and α-crystallin from microchiropterans. Although a monophyletic origin of these two suborders seems likely (Thenius 1969), the varying placement of the two suborders in the cytochrome *c* and α-crystallin trees may be explained if these two suborders prove to be of biphyletic origin.

Pholidotes not close to Edentates

Apart from the complete sequence of the Malayan pangolin αA chain we have obtained the largest part of the αA sequence of the tree pangolin *Manis (Phataginus) tricuspis* (de Jong et al. 1982). The sequences differ from each other at least at four positions, indicating a considerable time of evolutionary divergence. The Malayan pangolin αA sequence is most parsimoniously connected to the carnivore-ungulate region of the tree, on the basis of the shared derived substitutions 4 Ala and 147 Pro. On the other hand 3 Ile, 101 Ser (which, however, is Asn in the tree pangolin) and 153 Ser could be considered as primitive characters, placing the pholidotes at the base of the eutherian radiation. Such a separation of the pholidotes as the earliest eutherian offshoot would cost an additional 2 NRs, or 3 NRs in the case of the tree pangolin.

The pangolin αA sequence certainly shows no synapomorphies with the edentates, and it actually adds 3 NRs to group the pangolins with the edentates. This finding has relevance for the continuing discussion about possible pholidote-edentate relationships. Some authors consider the pholidotes as the closest relatives of the edentates (van Valen 1971; Szalay 1977; Patterson 1978), or do not exclude this possibility (Simpson 1945; Engelmann 1982), but others have entirely abandoned this idea (Romer 1966; Thenius 1969; McKenna 1975) and leave their origins completely open. The possibility that the Pholidota might be placed within the Edentata, derived from a myrmecophagid-like species (Szalay 1977; Engelmann 1982) seems to be excluded by the αA sequence data.

The palaeanodonts, which may be ancestral to or relatives of early pangolins, have been placed by Emry (1970) in the order Pholidota. Rose (1978) has suggested that palaeanodonts are possibly close to the Pantolestoidea, which McKenna (1975) grouped with the carnivores in the grandorder Ferae. The similarities between pholidote and carnivore αA sequences thus support the relationships based on this paleontological evidence. Although no other protein sequence data are yet available for the pangolins, immunological findings with chicken antisera support the grouping of Pholidota with Carnivora (Shoshani, J. unpublished data).

Carnivore-ungulate relationship

In all parsimony solutions carnivores, ungulates, cetaceans and pangolins are grouped on the same branch. This is due to the shared derived substitutions 4 Thr→ Ala and 147 Gln→ Pro, which are present in all investigated species from these groups (apart from porpoise, which has the autapomorphous substitution 147 Pro→ Thr). These residues at these positions do not occur in any other mammalian order (Table 2). Furthermore the back substitution 3 Val→ Ile contributes to the separation of a carnivore-ungulate branch.

Apart from the uncertainties concerning the position of the pangolins, the joining of ungulates, whale and carnivore αA sequences on the same branch may well reflect a monophyletic origin. Van Valen (1966) and McKenna (1969) argue that the cetaceans arose within the mesonychid condylarthrans. The placement of the whales among the ungulates, strongly supported by neontological studies (Thenius 1969), is now widely accepted. Simpson (1945) brought together the ungulates (but not the whales) and the carnivores in the cohort Ferungulata. Both Lillegraven (1969) and McKenna (1969) considered the possibility that late Cretaceous palaeoryctids gave rise to creodonts, carnivores and ungulates. According to Szalay (1977) no substantive evidence exists to

contradict the concept Ferungulata, although no undisputed shared derived characters have clearly been brought forward. A common origin of ungulates and carnivores is not widely accepted, and most authors prefer to remain uncommitted on this issue.

The sequences of cytochrome *c*, myoglobin and pancreatic ribonuclease of different whales are known. They tend to join the Cetacea to the ungulates, although the most parsimonious solutions in these cases are not similar to those based on traditional anatomical evidence. Carnivores and ungulates can be compared by known sequences of myoglobin, cytochrome *c*, α- and β-hemoglobins, and fibrinopeptides. Cytochrome *c* seems to support a carnivore-ungulate relationship, whereas myoglobin tends to place Carnivora closer to Lagomorpha and *Tupaia* than to either Primates or ungulates and cetaceans while fibrinopeptides and hemoglobin tend to place carnivora on a branch containing Primates, Lagomorpha, *Tupaia*, and Rodentia.

Monophyly of Pinnipeda

The αA sequences show no shared derived substitutions for all six investigated carnivores, and are unable to resolve the relationships between dog, cat, bear and mink, which represent four different families, and the pinnipeds. The identical sequences of seal and sea lion αA chains, however, contain 3 synapomorphous substitutions. Two of these (51 Pro and 52 Val) are unique among all investigated αA chains, and strongly indicate a monophyletic pinniped origin. Such a pinniped monphyly is indeed the classical phylogenetic opinion. It has, however, been proposed that the Otariidae (sea lions) might be related to the families Canidae or Ursidae, and the Phocidae (seals) to the family Mustelidae (Savage 1957; McLaren 1960; Tedford 1976). The only other protein sequenced both in seal and sea lion is myoglobin which also supports, albeit weakly, a monophyletic origin. Albumin immunological evidence likewise indicated a pinniped monophyly (Prager and Wilson 1978). Preliminary comparative data on hemoglobin chains from six fissiped carnivore families revealed close relationships between badger (Mustelidae) and raccoon (Procyonidae) chains (Hombrados et al. 1978; Brimhall et al. 1979).

Ungulate interrelationships

The order of branching of the orders Carnivora, Pholidota, Cetacea, Artiodactyla and Perissodactyla is not really resolved by the αA sequences. The ungulates and whales appear to be monophyletic in Figs 2 and 4, but just on the basis of a single substitution (90 Leu → Gln), which has occurred repeatedly, back and forth, in different taxa. The same is true for the preferred position of the whales as a sister group to the Perissodactyla, based on substitution 150 Val → Met. There actually is some immunological and karyological evidence to consider the whales as most closely related to Artiodactyla among the mammalian orders (Thenius 1969), but in recent classifications of the mammals this issue is left undecided. The improbable, but sometimes discussed possibility of a biphyletic origin of Odontoceti and Mysticeti is refuted nor supported by the sequences of the αA chains of their respective representatives porpoise and minke whale.

Among ungulates only the three perissodactyls are joined together by a unique synapomorphous substitution 127 Ser → Thr. Within the Perissodactyla it is equally parsimonious to show tapir closer to rhinoceros than to horse, or alternatively to bring rhinoceros and horse together. However, in a recent analysis involving fibrinopeptides A and B of 47 mammals tapir and rhinoceros grouped first before joining Equidae (Goodman 1981), in agreement with the taxonomically preferred position.

Although the Artiodactyla are classically divided in the suborders Suiformes and Ruminantia, the hippopotamus which belongs to the first group is placed on the ruminant branch of all αA trees, due to the shared derived substitution 146 Val → Ile. As can be seen in Fig. 4, however, this 146 Val → Ile substitution occurs frequently throughout the tree.

Nevertheless the separation of the pig and hippopotamus αA sequences deserves some attention since the ribonuclease sequences also join hippopotamus to the ruminants rather than to pig (Beintema and Lenstra 1982). Pig and hippopotamus cytochrome *c* differ at three positions (Thompson et al. 1978), while porcine cytochrome *c* is identical to the bovine and ovine sequence, and hippopotamus is slightly more similar to camel and guanaco than to pig.

Discussion and prospects

In using protein sequences for the elucidation of phylogenetic problems, one is dependent on the numbers and kinds of amino acid substitutions which occurred during descent of the proteins under investigation. The α-crystallin A chain has undergone just the "right" degree of change in certain evolutionary lines to be very informative. For example, the fact that in our most parsimonious trees there are three to four amino acid substitutions on the paenungulate stem which are then retained by all four paenungulate orders, Proboscidea, Sirenia, Hyracoidea, and Tubulidentata, provides evidence for the monophyletic origin of these orders. Similarly a sufficient number of synapomorphous or shared derived substitutions occur on the stem to sloths and anteaters as to provide evidence of the monophyly of these edentates. This is also the case on the ancestral line to the catarrhine primates man and rhesus monkey. In other lines, the rate of change of the αA gene has been so slow as to be uninformative, as in the rodents and lagomorphs. In these instances proteins such as hemoglobin, myoglobin and pancreatic ribonuclease, which evolve faster than α-crystallin, may be able to help unravel the relationships. However, patterns of relationship between taxa involving relatively short periods of common ancestry in the distant past, may turn out to be unresolvable by any macromolecular data.

Holmquist (1978) has shown that the "denseness" of a phylogenetic network bears importantly on the accuracy of evolutionary reconstructions from protein sequence data. The quality of the evolutionary information derived from many closely related sequences is higher than that derived from a few distantly related sequences. A tree is maximally dense when the link lengths between nodes correspond to one nucleotide replacement or zero replacement. The α-crystallin A tree approaches this maximal density fairly well in certain regions. It can indeed be seen that the increased denseness of this tree as compared with the one in de Jong et al. (1977) (in which few regions have link lengths of 1 or 0) allows more precision in the assignment of substitutions to certain branches. For instance, from the present data set the substitution 13 Ala → Thr is seen to be an autapomorphy of the ox, which probably occurs independently in tapir and rhinoceros. The previous limited data set showed horse and pig to have 13 Ala, and ox and rhinoceros to have 13 Thr, but this information did not allow a decision about location and direction of the substitution.

Taken as a whole, the present results demonstrate the general usefulness of protein sequence data in the study of mammalian phylogeny, provided that an appropriate choice of taxa is made, and that the investigated protein shows an appropriate degree of change. The suggestion (de Jong et al. 1977) that the α-crystallin A chain would probably be a suitable tool to study the phylogenetic relationships of Edentata, Pholidota, Tubulidentata, and Chiroptera has been justified by the present study. Similarly, it can now be suggested from the results in Figs 4 and 5 which remaining taxa are likely to further elucidate mammalian phylogeny if added to the αA tree. The occurrence of 5 synapomorphous substitutions and a deletion in the αA chain of *Macaca mulatta* and *Homo sapiens* make it very promising to study *Tarsius*, in view of its disputed relationship with the Anthropoidea. The inclusion of a new-world monkey, preferrably the large-eyed *Aotes*, of which a few specimens should yield ample material, would obviously make the outcome more significant. The considerable numbers of autapomorphies in hedgehog and primate αA chains would make it worthwhile to study the Dermoptera and possible relatives of the

insectivores such as the elephant shrew. Since only 2 or 3 autapomorphous substitutions were detected in the microchiropteran bat *Artibeus* it is unlikely that αA sequences can reveal whether the origin of Microchiroptera and Megachiroptera was monophyletic or biphyletic. Nevertheless, it would still be desirable to study a representative of the latter suborder because it may strengthen the positioning of the Chiroptera in the αA tree. The presence of a number of autapomorphies in several carnivores indicates that useful studies could be conducted on the remaining carnivore superfamilies. Finally, the finding of 9 synapomorphies in the marsupial αA chains and 7 autapomorphies in the opossum make it likely that the αA chain sequences could help unravel the phylogenetic relationships among marsupial higher taxa.

Acknowledgements

We thank all persons who have made available eye lenses for our investigations (Table 1), Eugénie Nuy-Terwindt, Anneke Zweers and Marlies Versteeg for technical assistance with the sequence determinations, and Dr. M. C. McKenna for valuable discussions. We also thank Dr. Gabriel Lasker for critically reading an earlier draft of this paper and offering helpful suggestions. This work was supported in parts by the Netherlands Foundation for the Advancement of Tropical Research (WOTRO) and by the Netherlands Foundation for Chemical Research (SON) with financial aid from the Netherlands Organization for Pure Research (ZWO), and by grant DEB 78-10717 from the National Science Foundation.

Zusammenfassung

Untersuchung der Stammesgeschichte der Säugetiere durch Sequenzanalysen des Augenlinsenproteins α-Kristallin

Die Aminosäuresequenzen des Augenlinsenproteins α-Kristallin A von 41 Säugetierspecies (17 Ordnungen vertretend) wurden untersucht mit dem Ziel, die Beziehungen zwischen den Säugerordnungen zu klären. Die beobachteten Aminosäuredifferenzen wurden zur Konstruktion von Kladogrammen verwendet. Die α-Kristallin-A-Sequenzen weisen darauf hin, daß:

- die Paenungulatenordnungen Proboscidea, Hyracoidea und Sirenia eine monophyletische Gruppe bilden, zu der auch die Tubulidentata (Erdferkel) gehören
- die Paenungulaten nicht mit den Ungulaten verwandt sind, sondern gemeinschaftlich mit den Zahnarmen die ältesten Abzweigungen des Stammes der Placentalia darstellen
- die Schuppentiere keine Beziehungen zu den Zahnarmen aufweisen und am besten in die Nähe der Raubtiere gestellt werden
- die Ungulaten zusammen mit den Walartigen und den Raubtieren eine monophyletische Gruppierung bilden
- unter den Raubtieren die Robben und Seelöwen monophyletisch sind
- die untersuchten Fledermäuse (Microchiroptera) sich nicht als verwandt mit den Insektenfressern oder Primaten herausstellten.

Die α-Kristallin-A-Sequenzen ließen die Nagetiere, Hasenartige, Insektenfresser, Primaten und Spitzhörnchen als einen unaufgelösten Cluster von Ordnungen erscheinen, aber innerhalb der Primaten konnten die Prosimiae eindeutig von den Anthropoidea unterschieden werden. Die Ergebnisse werden mit gängigen Auffassungen über die Säugetierphylogenie verglichen und mit anderen vergleichenden Proteinsequenzdaten in Beziehung gebracht.

References

Baba, M.; Darga, L.; Goodman, M.; Czelusniak, J. (1981): Evolution of cytochrome *c* investigated by the maximum parsimony method. J. Mol. Evol. **17**, 197–213.

Beintema, J. J.; Lenstra, J. A. (1982): Evolution of mammalian pancreatic ribonucleases. In: Macromolecular sequences in systematics and evolutionary biology. Ed. by M. Goodman, New York: Plenum Press (in press).

Bloemendal, H. (ed.) (1981): Molecular and cellular biology of the eye lens. New York: Wiley-Interscience.

Brimhall, B.; Stangland, K.; Jones, R. T.; Becker, R. R.; Bailey, T. J. (1979): Tryptic peptide compositions of hemoglobins from mink *(Mustela vison)*, and hyena *(Hyaenae hyaenae)* and Hb α-chain of coatimundi *(Nasua nasua)*. Hemoglobin **3**, 271–292.

Butler, P. M. (1972): The problem of insectivore classification. In: Studies in vertebrate evolution. Ed. by K. A. Joysey and F. R. Kemp. Edinburgh: Oliver and Boyd. pp. 253–265.

Castillo, O.; Lehmann, H. (1977): The myoglobin of the fruit-bat *(Rousettus aegypticus)*. Biochim. Biophys. Acta **492**, 232–236.

Dayhoff, M. O. (ed.) (1972): Atlas of protein sequence and structure. Vol. 5. Silver Spring, Maryland: National Biomedical Research Foundation. pp. 1–124.
De Jong, W. W.; Terwindt, E. C. (1976): The amino acid sequences of the α-crystallin A chains of red kangaroo and virginia opossum. Eur. J. Biochem. **67**, 503–510.
De Jong, W. W.; Gleaves, J. T.; Boulter, D. (1977): Evolutionary changes of α-crystallin and the phylogeny of mammalian orders. J. Mol. Evol. **10**, 123–135.
De Jong, W. W.; Zweers, A.; Goodman, M. (1980): Trends in the molecular evolution of α-crystallin. Prot. Biol. Fluids, 28th Colloq. Oxford: Pergamon Press. pp. 161–164.
De Jong, W. W.; Zweers, A.; Joysey, K. A.; Gleaves, J. T.; Boulter, D. (1982): Protein sequence analysis applied to xenarthran and pholidote phylogeny. In: The evolution and ecology of sloths, anteaters, and armadillos. Ed. by G. G. Montgomery. Washington, D. C.: Smithsonian Institution Press (in press).
Emry, R. J. (1970): A North American Oligocene pangolin and other additions to the Pholidota. Bull. Amer. Mus. Nat. Hist. **142**, 459–510.
Engelmann, G. F. (1982): The phylogeny of the Xenarthra. In: The evolution and ecology of sloths, anteaters, and armadillos. Ed. by G. G. Montgomery. Washington, D. C.: Smithsonian Institution Press (in press).
Farris, J. S. (1972): Estimating phylogenetic trees from distance matrices. Amer. Nat. **106**, 645–668.
Fitch, W. M. (1977): On the problem of discovering the most parsimonious tree. Amer. Nat. **111**, 223–257.
— W. M. (1979): Cautionary remarks on using gene expression events in parsimony procedures. Syst. Zool. **28**, 375–379.
Fitch, W. M.; Margoliash, E. (1967): The construction of phylogenetic trees. Science **155**, 279–284.
Foulds, L. R.; Penny, D.; Hendy, M. D. (1979): A general approach to proving the minimality of phylogenetic trees illustrated by an example with a set of 23 vertebrates. J. Mol. Evol. **13**, 151–166.
Glass, B. P. (1982): History of classification and nomenclature in Edentata (= Xenarthra). In: The evolution and ecology of sloths, anteaters, and armadillos. Ed. by G. G. Montgomery. Washington, D. C.: Smithsonian Institution Press (in press).
Goodman, M. (1981): Decoding the pattern of protein evolution. Progress in Biophysics and Molecular Biology **37**, 105–164.
Goodman, M.; Moore, G. W.; Barnabas, J.; Matsuda, G. (1974): The phylogeny of human globin genes investigated by the maximum parsimony method. J. Mol. Evol. **3**, 1–48.
Goodman, M.; Czelusniak, J.; Moore, G. W.; Romero-Herrera, A. E.; Matsuda, G. (1979): Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Zool. **28**, 132–163.
Gregory, W. K. (1910): The orders of mammals. Bull. Amer. Mus. Nat. Hist. **27**, 1–524.
Hendy, M. D.; Penny, D.; Foulds, L. R. (1978): Identification of phylogenetic trees of minimal length. J. Theor. Biol. **71**, 441–452.
Hoffstetter, R. (1973): Origine, compréhension et signification des taxons de rang supérieur: quelques enseignements tirés de l'histoire des mammifères. Ann. Paléont. **59**, 137–169.
Holmquist, R. (1978): A measure of the denseness of a phylogenetic network. J. Mol. Evol. **11**, 225–231.
Hombrados, I.; Neuzil, L.; Debuire, B.; Han, K. (1978): The amino acid sequence of the α chain of badger *(Meles meles)* haemoglobin. Biochim. Biophys. Acta **535**, 1–10.
Lavocat, R. (1978): Rodentia and Lagomorpha. In: Evolution of African mammals. Ed. by V. J. Maglio and H. B. S. Cooke, Cambridge, Mass.; Harvard Univ. Press. pp. 69–89.
Le Gros Clark, W. E.; Sonntag, Ch. F. (1926): A monograph of *Orycteropus afer*. III. The skull, skeleton of the trunk and limbs. Proc. Zool. Soc. Lond. 1926: I–II, 445–485.
Lillegraven, J. A. (1969): Latest Cretaceous mammals of upper part of Edmonton Formation of Alberta, Canada, and review of marsupial-placental dichotomy in mammalian evolution. Univ. Kansas Paleont. Contr., art. **50** (vert. 12), 1–122.
Luckett, W. P. (ed.) (1980): Comparative biology and evolutionary relationships of tree shrews. New York; Plenum Press.
Maeda, N.; Fitch, W. M. (1981): Amino acid sequence of a myoglobin from lace monitor lizard, *Varanus varius*, and its evolutionary implications. J. Biol. Chem. **256**, 4301–4309.
Marshall, L. G. (1979): Evolution of metatherian and eutherian mammalian characters: a review based on cladistic methodology. Zool. J. Linn. Soc. **66**, 369–410.
McKenna, M. C. (1969): The origin and early differentiation of therian mammals. Ann. N. Y. Acad. Sci. **167**, 217–240.
— (1975): Toward a phylogenetic classification of the Mammalia. In: Phylogeny of the primates: a multidisciplinary approach. Ed. by W. P. Luckett and F. S. Szalay. New York; Plenum Press. pp. 21–46.
McKenna, M. C.; Manning, E. (1977): Affinities and palaeobiogeographic significance of the Mongolian paleocene genus *Phenacolophus*. Geobios Mem. spec. **1**, 61–85.
McLaren, I. A. (1960): Are the Pinnipedia biphyletic? Syst. Zool. **9**, 18–28.

Moore, G. W.; Barnabas, J.; Goodman, M. (1973): A method for constructing maximum parsimony ancestral amino acid sequences on a given network. J. Theor. Biol. **38**, 459–485.
O'Neil, P. B.; Doolittle, R. F. (1974): Mammalian phylogeny based on fibrinopeptide amino acid sequences. Syst. Zool. **22**, 590–595.
Patterson, B. (1978): Pholidota and Tubulidentata. In: Evolution of African mammals. Ed. by N. J. Maglio and H. B. S. Cooke. Cambridge, Mass.: Harvard Univ. Press. pp. 268–278.
Peacock, D.; Boulter, D. (1975): Use of amino acid sequence data in phylogeny and evaluation of methods using computer simulation. J. Mol. Biol. **95**, 513–527.
Prager, E. M.; Wilson, A. C. (1978): Construction of phylogenetic trees for proteins and nucleic acids: empirical evaluation of alternative matrix methods. J. Mol. Evol. **11**, 129–142.
Romer, A. S. (1966): Vertebrate paleontology. 3rd ed. University of Chicago Press.
Romero-Herrera, A. E.; Lehmann, H.; Joysey, K. A.; Friday, A. E. (1978): On the evolution of myoglobin. Phil. Trans. Roy. Soc. B **283**, 61–163.
Rose, K. D. (1978): A new paleocene epoicotheriid (Mammalia), with comments on the Palaeanodonta. J. Paleont. **52**, 658–674.
Sarich, V. M. (1982): Edentate systematics: albumin immunological evidence. In: The evolution and ecology of sloths, anteaters, and armadillos. Ed. by G. G. Montgomery. Washington, D. C.; Smithsonian Institution Press (in press).
Sarich, V. M.; Cronin, J. E. (1976): Molecular systematics of the primates. In: Molecular anthropology. Ed. by M. Goodman and R. Tashian. New York: Plenum Press. pp. 141–170.
Savage, R. J. G. (1957): The anatomy of *Potamotherium*, an Oligocene lutrine. Proc. Zool. Soc. Lond. **129**, 151–244.
Shoshani, J.; Goodman, M.; Prychodko, W. (1978): Clasdistic analysis of Paenungulata by computer (Abstract). Amer. Zool. **18**, 601.
Simpson, G. G. (1945): Principles of classification and a classification of mammals. Bull. Amer. Mus. Nat. Hist. **85**, 1–350.
Sokal, R. R.; Michener, C. D. (1958): A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. **38**, 1409–1438.
Szalay, F. S. (1977): Phylogenetic relationships and a classification of eutherian mammals. In: Major patterns in vertebrate evolution. Ed. by. M. K. Hecht, P. C. Goody and B. Hecht. New York: NATO ASI Series, Plenum Press. pp. 315–374.
Tedford, R. H. (1976): Relationship of pinnipeds to other carnivores (Mammalia). Syst. Zool. **25**, 363–374.
Thenius, E. (1969): Phylogenie der Mammalia. Berlin; Walter de Gruyter.
Thompson, R. B.; Borden, D.; Tarr, G. E.; Margoliash, E. (1978): Heterogeneity of amino acid sequence in hippopotamus cytochrome *c*. J. Biol. Chem. **253**, 8957–8961.
Van Druten, H. A. M.; Peer, N. G. M.; Bos, F. A. B. H.; De Jong, W. W. (1978): Reciprocal amino acid substitutions in the evolution of homologous peptides. J. Theor. Biol. **73**, 549–561.
Van Valen, L. (1971): Adaptive zones and the orders of mammals. Evolution **25**, 420–428.
Weitz, B. (1953): Serological relationships of hyrax and elephant. Nature 171, 261.

Authors' adresses: Dr. Wilfried W. de Jong, Laboratorium voor Biochemie, Geert Grooteplein Noord 21, Universiteit van Nijmegen, 6525 EZ Nijmegen, The Netherlands, and Prof. Morris Goodman, Department of Anatomy, Wayne State University School of Medicine, 540 East Canfield Avenue, Detroit, Michigan 48201, USA