Qualitative molecular markers
for evolutionary analyses
in vertebrates

een wetenschappelijke proeve op het gebied van de
Natuurwetenschappen, Wiskunde en Informatica.

Proefschrift
ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Nijmegen,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op
donderdag 7 maart 2002,
des namiddags om 1.30 uur precies

doors
Miriam Afra van Dijk
geboren op 10 juli 1970
te Kampen
Promotor:
Prof. Dr. W.W. de Jong (UvA)

Manuscriptcommissie:
Prof. Dr. J.J. Beintema (RUG)
Dr. J.H.P. Hackstein

The investigations described in this thesis were carried out under the auspices of the Research Council for Earth and Life Sciences (ALW) with financial aid from the Netherlands Organization for Scientific Research (NWO) and performed at the Department of Biochemistry, Faculty of Science, University of Nijmegen, The Netherlands.

Printed by PrintPartners Ipskamp, Enschede
## Contents

<table>
<thead>
<tr>
<th>Chapter 1</th>
<th>General introduction: aspects of molecular evolution and phylogeny</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 2</td>
<td>The virtues of gaps: xenarthran (edentate) monophyly supported by a unique deletion in αA-crystallin</td>
<td>27</td>
</tr>
<tr>
<td>Chapter 3</td>
<td>Indels in tumor necrosis factor α and islet amyloid precursor protein indicate that rodents are monophyletic and lagomorphs are their sister group</td>
<td>45</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>Protein sequence signatures support the African clade of mammals</td>
<td>61</td>
</tr>
<tr>
<td>Chapter 5</td>
<td>αA-crystallin in avian phylogeny: the hoatzin groups with Cuculiformes</td>
<td>79</td>
</tr>
<tr>
<td>Chapter 6</td>
<td>The evolution of an alternatively spliced exon in the αA-crystallin gene</td>
<td>91</td>
</tr>
<tr>
<td>Chapter 7</td>
<td>General discussion</td>
<td>103</td>
</tr>
<tr>
<td>Chapter 8</td>
<td>Summary &amp; Samenvatting</td>
<td>115</td>
</tr>
<tr>
<td></td>
<td>List of publications</td>
<td>122</td>
</tr>
<tr>
<td></td>
<td>Dankwoord</td>
<td>124</td>
</tr>
<tr>
<td></td>
<td>Curriculum vitae</td>
<td>125</td>
</tr>
</tbody>
</table>
CHAPTER 1

General introduction:
aspects of molecular evolution and phylogeny
1. Why the interest in molecular evolution?

1.1 Research fields of molecular evolution

The evolutionary history of organisms can be studied in two independent ways. Traditionally, paleontologists unravel the evolutionary history of living or extinct organisms by using characters like anatomy, teeth, cranioskeleton, etc., or fossil remnants. In addition, morphologists apply physiological and developmental characters. Since fossil evidence and morphological data are often difficult to interpret and inconclusive (reviewed in Patterson et al., 1993), molecular data has been very useful to determine phylogenies in these situations. The purpose of molecular phylogeny is to identify the successive steps in the evolution of living organisms. A robust phylogenetic tree of the major groups of organisms is necessary to reconstruct the biogeographical, morphological, developmental or behavioral pathways of their evolution. This broad field of molecular evolution is illustrated by the following recent examples.

The historical biogeography of placental mammals has been studied by analyzing a concatenation of gene sequences (5708 bp) (Madsen et al., 2001). This indicated a basal separation of placental orders into two groups, with southern and northern hemispheric origins, respectively. In addition, the combined data of phylogenetic topologies, molecular clock calculations and plate tectonic reconstructions suggested that placental mammals originated in Gondwana, and subsequently dispersed via microplates to the Northern Hemisphere. A second example concerning the origin of morphological characters - echolocation and flight in bats (order Chiroptera) - was examined by the analysis of bat relationships. This order is traditionally subdivided into Microchiroptera and Megachiroptera. A feature of microbats is the presence of complex laryngeal echolocation systems, while the megabats have an enhanced visual acuity. Molecular evidence supported paraphyly of the microbats: two microbat species from different families within the superfamily Rhinolophoidea cluster together with the megabats. This indicates that the echolocation systems either evolved independently in Rhinolophoidea and other microbats or were lost in the evolution of megabats (Teeling et al., 2000). Another example shows how the phylogenetic tree of geckos helped to understand the developmental pathway by which the eye lens protein iota-crystallin of the gecko Lygodactylus picturatus originated (Werten et al., 2000). This diurnal gecko, living in African savannas, has evolved from nocturnal ancestors. It is unable to close its eyelids or diminish the aperture of its iris to reduce the intensity of incoming light. Instead, the eye lenses of this gecko are yellow due to the presence of 3-dehydroretinol (vitamin A2) which forms a complex with iota-crystallin. This complex absorbs short-wave radiation, effectively protecting the retina against ultraviolet damage. Iota-crystallin turned out to be identical to cellular retinol-binding protein type I, overexpressed in the eye lens of L. picturatus to fulfil this protective function.

 Besides allowing to reconstruct the evolutionary history of genes and organisms, investigations at the molecular level give additional information about the rates and patterns of change occurring in DNA sequences and proteins during evolutionary time, as well as about the mechanisms responsible for such changes. Selective constraints controlling a protein function can explain part of the changes in rates and patterns of DNA sequence evolution. For instance, decreased selective constraints act upon the eye of the blind mole rat (Spalax ehrenbergi) due to a change of habitat. After its ancestors adopted a subterranean way of life, some 25 million years ago, only remnants of eyes were left. One would expect that such decreased functional constraints might affect those mole rat proteins that are related to visual
abilities, such as the highly conserved eye lens protein $\alpha$A-crystallin. Indeed, an increased number of substitutions was detected in $\alpha$A-crystallin (Hendriks et al., 1987; Smulders et al., 2002). However, the sequence of mole rat opsin coding for a functional photopigment in the eye shows substantial similarity to both mouse (91%) and human (87%) long-wave-length-sensitive cone opsins. The evolutionary conservation indicates that the function of this photopigment is possibly associated with a light detection task to entrain circadian rhythms of locomotor behavior (David-Gray et al., 1998).

2. Molecular phylogeny

2.1 Vertebrate phylogeny

In this thesis the primary interest is the molecular evolution of mammals and birds. Together with fish, amphibians and reptiles, these groups form the vertebrates. The phylogenetic relationships within the vertebrates are currently heavily debated (Fig. 1). Based on morphological (Forey and Janvier, 1993) and mitochondrial data (Rasmussen et al., 1998) the Agnatha (hagfishes and lampreys), the first to diverge within this group, are considered to be paraphyletic, in contrast to inferences from nuclear ribosomal genes (Stock and Whitt, 1992; Mallatt and Sullivan, 1998). Then, trees inferred from complete mitochondrial genomes place sharks and rays (chondrichthyians) within the teleost fish (Rasmussen and Arnason, 1999). However, the most widely accepted opinion is that teleosts are more closely related to lungfish and tetrapods (amphibians, reptiles, birds and mammals), to the exclusion of chondrichthyians. The terrestrial vertebrates (Tetrapoda) form a clearly distinguished monophyletic group with as closest relative the lungfish. The extant tetrapods can again be subdivided into amphibians and amniotes (Caroll, 1987; Benton, 1990). This latter group comprises reptiles, mammals and birds, which rapidly diverged at the end of the Paleocene, some 300-250 million years ago. Phylogenetic relationships among the major groups of amniotes therefore remain a matter of controversy. Most studies are in favor of mammals being the first to split off from the ancestral amniote lineage (Laurin & Reisz, 1995; Benton, 1997). Under intensive debate remains the phylogenetic position of the turtles. Some authors proposed a sister group relationship of turtles with lepidosaurs (lizards and snakes) (deBraga and Rieppel, 1997) or archosaurs (birds and crocodilians) (Kumazawa and Nishida, 1999), while others suggested the traditional placement of turtles as the sister clade to diapsids (archosaurs and lepidosaurs) (Laurin & Reisz, 1995; Benton, 1997).

Based on fossil records, the radiation of placental mammals (the Eutheria) is estimated to have taken place 50 to 70 million years ago. The morphological transition from only a few to many different forms of placental mammals could occur - according to paleontologists - after the extinction of the dinosaurs as a result of an asteroid colliding with Earth (Alvarez et al., 1980; Kerr, 1994). After the demise of dinosaurs, all ecological niches that they had occupied became available for the mammals to diversify. Consequently, the various mammalian clades have diverged in a rather short time scale, which greatly complicates solving higher level relationships among mammals. Based on a combination of the morphological and molecular evidence available at the time, Novacek proposed in 1992 a reasonably well-resolved tree of the placental mammalian orders (Novacek, 1992) (Fig. 2). In this tree, the Monotremata (egg-laying Platypus and Echidna) and Marsupialia (opossums, kangaroos and other pouched mammals) successively branch off from the main mammalian stem. The remaining mammals,
General introduction

Figure 1. A phylogenetic tree showing the ‘classic’ topology of vertebrate relationships.

the Eutheria, are subdivided in about 18 orders. The sister group to all other placental orders are the Edentata (or Xenarthra: anteaters, armadillos and sloths), possibly together with the Pholidota (pangolins). Some widely accepted superordinal groupings were distinguished in this tree: Glires, clustering Rodentia and Lagomorpha (rabbits, pikas, etc.); Archonta, combining Primates, Scandentia (tree shrews), Dermoptera (flying lemurs) and Chiroptera (bats); an ungulate clade grouping Artiodactyla (even-toed ungulates), Cetacea (dolphins, whales), Perissodactyla (odd-toed ungulates), the paenungulate orders Proboscidea (elephants), Sirenia
Figure 2. A phylogenetic tree of the placental mammalian orders, based on a combination of morphological and molecular evidence, as proposed by Novacek (1992). Dashed lines are relatively more ambiguous relationships.
(sea cows) and Hyracoidea (hyraxes) and finally the monospecific order Tubulidentata (aardvark). The relationships of Insectivora - or Lipotyphla - and Carnivora with other extant orders remained unsolved, and the Macroscelidea (elephant shrews), Pholidota, Artiodactyla and Hyracoidea were considered to have relatively uncertain relationships.

Contrary to expectation, the addition of further molecular data did not provide increased support or resolution of the Novacek tree, but rather introduced more controversies between morphological and molecular data. First of all, comparative molecular sequence analyses suggest that the mammalian radiation began already in the Early Cretaceous, more than 100 million years ago, about twice the time as indicated by fossil evidence (Easteal et al., 1995; Hedges et al., 1996; Kumar and Hedges, 1998). Also the topologies of trees as proposed by morphological and molecular data raised many discrepancies and controversies. We just mention here those cases that are most relevant to the work described in this thesis. Morphological data classically support, be it weakly, the monophyly of the order Lipotyphla. This order comprises the familiar insectivore families Soricidae (shrews), Talpidae (moles) and Erinaceidae (hedgehogs), but also the less known Solenodontidae (solenodons, endemics of Cuba and Haiti) and the African families Chrysochloridae (golden moles) and Tenrecidae (tenrecs) (Macphee and Novacek, 1993). Recently, extensive sequence data have suggested the placement of golden moles and tenrecs in a separate order, Afrotheria, that forms a strongly supported superordinal clade with other African orders, hence named the Afrotheria (Springer et al., 1997; Stanhope et al., 1998a,b). In addition to Afrotheria, the Afrotheria include the orders Sirenia, Proboscidea, Hyracoidea, Macroscelidea, and Tubulidentata. Another important issue is the traditionally accepted monophyly of the order Rodentia that is challenged by several molecular studies (Graur et al., 1991; D’Erchia et al., 1996; Janke et al., 1997). Finally, the morphologically supported sister grouping of Rodentia and Lagomorpha, within the cohort Glires, is rejected on basis of both mitochondrial and nuclear sequence data (D’Erchia et al., 1996; Graur et al., 1996; Janke et al., 1997; Arnason et al., 1997; Cao et al., 1997; Penny and Hasegawa, 1997; Sullivan and Swofford, 1997).

2.2 Methods and sources for the study of molecular phylogeny

The study of molecular phylogeny started using immunological approaches; the intensity of cross-reactivity of an antibody was determined between corresponding proteins from different species (Ridley, 1996, p. 499). An antibody recognizes the corresponding protein of the same species better than a similar protein of another species. Thus, the obtained immunological distances are taken to reflect the degree of relationship between the species. Then, also changes in DNA sequences resulting from nucleotide substitutions were indirectly measured by restriction enzyme mapping and DNA-DNA hybridization (Li and Graur, 1991). A restriction-fragment pattern shows the number of fragments and their size resulting from the digestion with endonucleases of genomic or mitochondrial DNA. Therefore, by sequentially and reciprocally using several restriction enzymes the location of restriction sites in a DNA sequence can be represented as a restriction map. Obviously, the greater the similarity of two DNA sequences, the more resemblance will be found in their restriction-fragment pattern. The DNA-DNA hybridization technique is based on the thermal stability of double-stranded DNA molecules. As the proportion of nucleotide matches between two strands decreases, the duplex becomes less thermally stable; two strands from different species will denature into single strands at lower temperatures than two strands from the same species. Both methods cannot provide the detailed character state information of DNA sequences; they only provide overall
distance information that has a predictive value. Since the introduction of the polymerase chain reaction (PCR) technique, these distance methods have largely become obsolete, and molecular phylogeny based on DNA sequences has rapidly gained momentum.

Several types of molecular sequences can be used to infer phylogenies. An increasingly important source to study molecular phylogeny of species as well as genes is the endless number of nuclear protein-coding genes. In the first case, genes must be orthologous, which means that sequence similarity of such genes is a consequence of a speciation event. However, paralogous sequences, descendants of genes that have diverged after gene duplication, are necessary to study the phylogeny of genes. Not very well explored yet is the possibility to help resolving phylogenetic problems with intron sequences. It is generally thought that their rate of evolution might be too fast and too erratic to be suitable for phylogenetic studies. Therefore, intron regions are mainly useful for phylogenetic analysis at lower taxonomic level. However, recently the usefulness of intronic sequences was tested on mammalian major histocompatibility complex (Mhc) genes that are known to recombine and rearrange frequently. The results showed that the rate of evolution in these introns is comparable to that of the synonymous sites in the protein-coding sequences of other genes, like von Willebrand factor (vWF) exon 28 and interphoto-receptor retinoid binding protein (IRBP) exon 1 (Kupfermann et al., 1999). The more slowly evolving nuclear ribosomal RNA sequences are often chosen at the higher taxonomic level.

Besides nuclear also mitochondrial sequences are frequently used for phylogenetic studies, mainly at lower taxonomic levels. An advantage of mitochondrial sequences, which are maternally inherited, is the small size of the genome and the lack of introns and recombination. A practical advantage is the fact that mitochondrial DNA can easily be isolated and analyzed. On the other hand, factors like heteroplasmy (the coexistence of more than one mitochondrial genome within one individual) can seriously complicate their analyses (Petri et al., 1996).

2.3 Rates and patterns of nucleotide substitutions

Molecular evolution can basically be defined as the process of change within allele frequencies in a population. Therefore, evolution can only proceed when genes are polymorphic, i.e. when two or more alleles coexist at a given locus. In general, one of these alleles will eventually become fixed and the others become extinct. One of the evolutionary forces that drive this process is natural selection, better known as the survival of the fittest. It can only take place when genetic variation among individuals in a population exists in characters that are related to reproductive success. Another evolutionary force that can cause changes in allele frequencies from generation to generation is genetic drift. This process causes random fluctuations in allele frequencies. Besides random sampling of gametes in the process of reproduction, genetic drift can also be caused by stochastic changes in selection intensity. The effects of random sampling on the frequencies of alleles in populations are much more pronounced in small populations than in larger ones (Li and Graur, 1991, pp. 22-29).

The molecular evolution of protein-coding regions is studied by the rates and patterns of nucleotide substitutions, divided in synonymous (no amino acid change) and nonsynonymous (amino-acid altering) substitutions. The synonymous rate normally exceeds the nonsynonymous rate. This difference is caused by the effect of natural selection at the protein level; synonymous mutations do not influence the structure of the protein and therefore its function. Nonsynonymous mutations may affect the function of a protein to different extent; the stronger the functional constraints working on nonsynonymous substitutions, the slower the
rate of evolution. However, this does not mean that synonymous mutations are selectively neutral. In both prokaryotes and eukaryotes, nonrandom usage of synonymous codons has been found. The choice of synonymous codons is mainly determined by the preference for the most abundant tRNAs, which is species-specific. It has been suggested that this will increase the efficiency and accuracy of translation (Li and Graur, 1991, pp. 91-93).

Variation in rates of nonsynonymous substitutions among genes can be explained by a difference in mutation rates among different genomic regions. However, more important is the intensity of selection, which is determined by functional constraints. The rates of synonymous substitutions varying from gene to gene are less understood. One possible cause is a difference in rate of mutation among the regions of the genome, and this would reflect the chromosomal localization of the gene (Wolfe et al., 1989). This is supported by the fact that each genome is made up of segments of distinct GC content called isochores. These elements may replicate independently, and consequently demonstrate different rates of mutation. Another possibility is that the fitness of synonymous codons is not similar. Varying substitution rates are also shown at the level of species. For example, the substitution rates in rodents are much higher than in primates (Wu and Li, 1985). This observation can in part be explained by rates of molecular evolution being correlated to generation times, because rodents have a much shorter generation time than primates. This means more cycles of replication and consequently more mutational errors. Furthermore, differences in metabolic rates, like the efficiency of the DNA repair system, may play a role. Finally, paleodemographic characters like changes in numbers of births, deaths, mating and cases of disease in a population over a period of time can have affected this variation.

In animals, the synonymous substitution rate in mitochondrial DNA is clearly higher than that in nuclear genes. This may be associated with oxidative damage produced by free radicals, generated during the transport of electrons to oxygen in the inner mitochondrial membrane. Other explanations are the lower proofreading activity of γ DNA polymerase during replication, and the lack of DNA repair systems. Another feature observed in mitochondrial genomes is the strong asymmetric distribution of the four bases between its L and H strand, which is reflected in the H-strand protein-coding genes. This asymmetry could be explained by the slow replication of mtDNA that is only completed after 2 hours. Therefore, the longer the parental H strand remains single stranded, the higher the probability of erroneous mutations. The spontaneous hydrolytic deamination of both cytosine and adenine on the H strand is probably responsible for the directional changes of the base composition as a nonselective process (Reyes et al., 1998). Such a process results in decreased percentages of guanine and thymidine on the L strand. Furthermore, a positive correlation is found between the percentage of nucleotide variable sites, and the duration of single-strandedness for each gene on the H-strand, with the exception of ATP8 and cytochrome b. Thus, the longer the genes remain single-stranded, the higher the number of variable sites. This indicates that the location of a gene on the H-strand is positively correlated with its degree of functional constraint (Reyes et al., 1998).

2.4 Molecular markers in phylogeny

The information for phylogenetic analyses is generally derived from base substitutions in long DNA sequences. Unfortunately, DNA sequences are prone to parallel, convergent and back mutations, jointly called homoplasy. Therefore, in addition to these quantitative sequence data, more qualitative markers like insertions and deletions (indels, or gaps), sequence
signatures, alternatively spliced exons, and short or long interspersed elements (SINEs and LINEs) are desired. These characters are more unique than base substitutions.

In spite of the fact that indels are rarer than base substitutions, and thus a potentially valuable source of phylogenetic information, they are generally removed before analysis (Swofford et al., 1996). This is mostly because of their frequent occurrence in regions that show the greatest sequence variability, corresponding with surface loops in the encoded proteins (Pascarella and Argos, 1992; Benner et al., 1993). Consequently, their positioning within a multiple alignment is often ambiguous. Also the various mutational processes and constraints involved in the creation of indels are not completely understood. Several of these mechanisms that mainly concern long indels are unequal crossing over and DNA transposition, while slipped-strand mispairing and intron-exon boundary sliding account for indels of up to 20-30 nucleotides in length (Li and Graur, 1991, p. 17). We explored the use of indels in protein-coding genes, which can only be accepted in multiples of three nucleotides, because the reading frame would otherwise be disrupted. Therefore, these indels are more constrained than in noncoding DNA.

Another molecular marker examined to determine monophyletic groups of organisms is the occurrence of characteristic ‘signatures’ in orthologous protein sequences from different species. These signatures are defined as unique combinations of amino acid replacements that according to cladistic reasoning represent synapomorphous character states. In other words, they are remnants of mutational events that occurred during the period of last common ancestry of the species in which a particular signature is found. The conservation of such sequence signatures would provide visible ‘protein morphological’ evidence for phylogenetic relationships. Finally, we analyzed the possibility of using the occurrence and sequence of an alternatively spliced exon as a phylogenetic marker. Alternative splicing is one of the evolutionary mechanisms to increase the diversity of gene products.

In the literature several SINEs and LINEs have also been described as molecular markers for phylogenetic purposes. Very well known examples of SINEs are the Alu family in man, and its rodent equivalent the B1 family (Serdobova and Kramerov, 1998; Hamdi et al., 1999). These are retrosequences that have lost their function, also called processed pseudogenes. Alu’s are around 300 bp in length, and member of the repeated sequence family. Around 5-6% of the human genome consists of Alu’s. Originally, they are derived from the 7SL functional gene. This conserved 7SL gene is essential in the cutting of signal sequences from secreted proteins. Furthermore, both Alu and B1 sequences are species-specific and therefore mainly useful for molecular phylogeny at lower taxonomic levels.

3. Phylogenetic analyses

3.1 Sequence alignment

Before reconstructing a phylogenetic tree based on sequence data, a multiple alignment of these sequences has to be made. Besides DNA sequences also amino acid sequences can be used. The latter source usually provides more reliable results at deeper levels of phylogeny. Theoretically, the most accurate way to align sequences is the ‘global optimization’ strategy (Swofford et al., 1996, p. 375). The method finds the smallest number of substitutions and gaps by comparing simultaneously several adjacent nucleotides of two sequences. However, it is very slow and has not yet been applied in computer programs. This is mainly caused by the exhaustive search required to find the global optimal alignment. Therefore, multiple alignment
programs implement the 'progressive alignment' strategy (Feng and Doolittle, 1987) as in the program CLUSTALW (Higgens et al., 1992). First, all sequences in the data set are compared to each other to identify the most similar ones. A simple phylogenetic tree is constructed using the distances across all pairwise alignments, referred to as the distance index. Then the sequences are progressively aligned together, working through the guide tree to align the most similar sequences first. Deletions and insertions can be penalized by gap creation and extension penalties (Swofford et al., 1996, p. 375). The number of gaps can be penalized with the gap creation penalty, while the gap extension penalty limits the size of the gaps. Substitutions can be assigned the same penalty or a matrix of substitution probabilities can be specified (see below in 3.2.1 Distance-based methods). A disadvantage of such an alignment strategy is that errors introduced at an early stage are carried through and can propagate themselves because of the order-dependence of this method. TreeAlign (Hein, 1989, 1990) is a program that continuously repeats this cycle until a stable alignment is obtained. In an alternative strategy the sequence alignment is part of the phylogenetic inference, rather than preceding it. MALIGN (Wheeler and Gladstein, 1994) has implemented this strategy that optimizes a multiple alignment by searching for the alignment that minimizes differences among the sequences. By using these algorithms the 'mathematically’ optimal alignment of two sequences can be found. These alignment programs will not always provide the biologically correct alignment because the given parameters do not necessarily reflect the behavior of biological sequences. Therefore, it is always required to verify the alignment by eye. Obtaining the 'correct’ alignment is essential for a reliable reconstruction of trees, because misaligned sequences will result in unrealistically and erroneously long branches, which are likely to have incorrect branch points with a tendency to migrate to the root of the tree.

3.2 Phylogenetic tree reconstructions

Several methods of tree reconstruction are available. They can be divided into two categories: distance-based and character-based (Swofford et al., 1996, pp. 407-514). Distance-based methods use the overall numbers of nucleotide or amino acid substitutions between sequences. Character-based methods consider each site in an alignment as a separate character, where the types of nucleotides or amino acids at that site are the character states. Phylogenetic trees can either be rooted or unrooted. When the direction of the evolutionary pathways is of interest, then a tree must be polarized by rooting, for which the use of an unambiguous outgroup is required. The outgroup should preferably be as closely related as possible to the studied monophyletic ingroup. An unrooted tree can only detect the branching topology of species under study but does not show the position of the deepest ancestor (Ridley, 1996, pp. 463-466).

3.2.1 Distance-based methods

To reconstruct phylogenetic trees on the basis of distances, it is necessary to determine the number of nucleotide substitutions - the genetic distance - separating two DNA sequences. The number of observed substitutions between each pair of sequences can easily be counted using the program DISTANCES (GCG package) (Devereux et al., 1984). However, this method is not very accurate because it does not take into account multiple mutations at single nucleotide sites. Therefore, corrections are needed to account for these hidden superimposed mutations. The two simplest distance methods are the Jukes-Cantor (JC) (Jukes and Cantor, 1969) and the Kimura (Kimura, 1980) method. Both methods assume that all nucleotides evolve
independently, but JC considers all types of nucleotide substitutions to occur with equal probability, while Kimura allows transitions (between purines or pyrimidines), and transversions (between purine and pyrimidine) to occur at different rates. More sophisticated is the 'Stationary Markov Model' that takes into account the base composition of the analyzed sequences. This model is able to distinguish the much faster evolving silent codon positions from the other codon positions. Another frequently used model is LogDet, which is robust to differences in base compositions among the taxa being studied (Steel, 1994; Lockhart et al., 1994).

Matrices of replacement probabilities also exist for protein sequence data. An equal-change-probability model for amino acids, comparable to JC for DNA, is the percent accepted mutation (PAM) matrix (Dayhoff, 1978). However, the replacement of an amino acid by a physicochemically similar residue is more likely. Therefore, more advanced models have been derived from Dayhoff’s original matrix. These Dayhoff PAM series are all based on estimated mutation rates, using substitutions observed in closely related proteins, and those rates are extrapolated to matrices used for distant relationships. More recently, an amino acid substitution matrix called JTT has been designed, an updated version of the substitution matrix of Jones et al. (1992). This model is preferred when the evolution of a broad diversity of species is analyzed. This is also true for the blocks substitution matrix (BLOSUM 62) (Henikoff and Henikoff, 1992) that obtains mutation frequencies directly from conserved regions of a protein family. In comparison with the frequently used PAM 160, BLOSUM 62 is less tolerant to substitutions involving hydrophilic amino acids, while it is more tolerant to hydrophobic substitutions. For rare amino acids, especially cysteine and tryptophan, BLOSUM 62 is typically more tolerant to mismatches than is PAM 160. Possibly, many of the differences between BLOSUM and PAM matrices arise from different constraints on conserved regions in general. Furthermore, the BLOSUM substitution matrix is derived from a much larger sequence data set than are the PAM substitution matrices, which makes the former more reliable.

The most frequently used distance-based method, not assuming a molecular clock, is neighbor-joining (Saitou and Nei, 1987). It sequentially identifies neighbor pairs that minimize the total length of the tree, starting with a star-like phylogeny. The correction for superimposed substitutions, as by the Jukes-Cantor and Kimura models mentioned above, is included. This is especially important when different lineages evolve at different rates. Another commonly used method related to neighbor-joining is minimum evolution (Rzhetsky and Nei, 1992).

### 3.2.2 Character-based methods

The simplest character-based method available is maximum parsimony. It selects the tree that requires the smallest number of evolutionary steps to explain the differences among the sequences under study. All phylogenetically non-informative sites are excluded first. This means that only a part of the DNA sequences is eventually used in the analysis. Till now, it is the only method that can give weights to indels (in the 4.0b1 PAUP version (Swofford, 1998)). A disadvantage of this method is its sensitivity for long branch attraction (Hendy and Penny, 1989). This phenomenon occurs when an increased number of substitutions in a lineage, as compared to the other lineages, causes convergence.

Another character-based method is maximum likelihood (ML), which determines the likelihood of changes required at each site, given a particular topology and evolutionary model. This method is, even for very short sequences, the best one available because of its superior compensation for superimposed changes and sampling variance. When a data set is too large to
find the optimal tree by an exhaustive search, a heuristic approach must be applied. Starting from an initial tree, the global optimum is sought. An initial tree can be obtained by stepwise addition of taxa to a growing tree, beginning with three taxa. It is also possible to start with a ‘star tree’ containing a single internal node, known as the star decomposition method. Then, one can sequentially add taxa in the same order as presented in the alignment. Unless the number of taxa is small or the data is very clean, these heuristic searches are unlikely to find the optimal trees. More accurate is to check all initial trees of three taxa and start with the one that yields the shortest tree. All remaining taxa are then successively connected to every branch of the tree to find the topology that requires the smallest increase in tree length. Two of the most frequently used rearrangements are nearest-neighbor interchanges (NNI), and tree bisection and reconnection (TBR). These methods are commonly referred to as branch swapping. The quartet puzzling method is another computer program to reconstruct phylogenetic trees from sequence data by maximum likelihood (Strimmer and von Haeseler, 1996). However, it is much faster because it restricts itself to analyzing all possible clusters of four sequences.

In conclusion, the main difference between the two character-based methods is that maximum parsimony ignores information contained in branch lengths while maximum likelihood does consider the probability of changes along branches.

3.3 Statistical methods

The statistical support for internal branches can be determined by a resampling method like bootstrapping (Efron, 1979). From the original alignment, new data sets of the same size are put together by random sampling of sites with replacement. Therefore, some sites from the original alignment will not be included while others are included once or more. A tree is reconstructed for each newly built data set. The proportion of resamplings supporting an internal branch determines the probability for this branch. Support for a monophyletic group can also be estimated with the decay index (Donoghue et al., 1992). This is calculated as the difference in total branch lengths between the shortest tree that contains a monophyletic group and the shortest tree that disrupts that group. Finally, it is also possible to apply tests for comparing trees. One is the ‘winning sites’ test (Prager and Wilson, 1988) that compares the number of characters that favor each of the respective trees. The test will indicate whether a tree is supported significantly better than expected when random variation among characters is assumed. Another way is to measure the difference in the minimum number of nucleotide substitutions between two trees at each informative site by using the Kishino-Hasegawa test (1989). In this method all nucleotides are assumed to be independent and distributed in the same way.

In this thesis we are especially interested in the significance of indels and protein sequence signatures used for phylogenetic purposes. Therefore, a method has been developed to calculate the likelihood for the observed distribution of deletions, based on the method of Kishino et al. (1990). This method will be described in more detail in chapter 2 (van Dijk et al., 1999). The probability that a sequence signature supports a clade of interest is examined statistically by calculating the likelihood of their most probable ancestral character state in a given topology, as further outlined in chapter 4 (van Dijk et al., 2001a).
3.4 Molecular clock

Apart from the reconstruction of phylogenetic relationships among organisms, it is also interesting to date the divergence of species. Paleontologists use the earliest fossil appearance of representatives from two lineages to determine the minimum time of divergence between these lineages. Molecular biologists estimate the mean divergence time by assuming a clock-like accumulation of sequence differences in genes since their divergence. Both morphological and molecular data have specific disadvantages and limitations. On one hand, morphological characters are difficult to code and interpret, and some of the data is inconclusive. On the other hand, molecular data must cope with variable rates of substitutions at different sites, different rates of transitions and transversions, and multiple substitutions at sites.

Finally, they both must deal with comparable methodological problems, like the choice of an appropriate outgroup, but also the choice of an algorithm in molecular phylogeny, while morphologists have to deal with the choice and definition of characters. As a result, the obtained trees may be affected by problems of convergence, long branch attraction, and rapid splitting of lineages. Therefore, divergence times based on molecular data are not always in agreement with the fossil record, especially when the radiations of modern birds and mammals are concerned (Benton, 1999). Figure 3 shows the estimated divergence times for mammalian orders and major lineages of vertebrates using an extensive data set of nuclear genes (Kumar and Hedges, 1998). The molecular clock must be calibrated with a known time of divergence obtained from the fossil record. Overall, fossil-based and molecular times are in relatively close agreement, except for the origin of placental orders and the early history of rodents. In conclusion, it is important to acknowledge that both methods have their own unique properties,
and both are equally valid in determining the pattern of phylogeny. So actually we have at our disposal two independent ways of reconstructing history.

3.5 The protein-coding genes used for our analysis

In the next chapters results are presented on molecular phylogeny using the following protein-coding genes: αA-crystallin, tumor necrosis factor alpha (TNFα) and islet amyloid polypeptide precursor (IAPP). αA-crystallin is related with the small heat shock proteins (shps) that are induced by heat and other types of stress (Ingolia and Craig, 1982). They prevent aggregation of proteins by interacting with denaturing proteins under these stressful conditions (Horwitz, 1992; Jakob et al., 1993; Merck et al., 1993). The 173-residue αA-crystallin is a structural protein in the eye lens, an organ that requires transparency for its function. Cell organelles are therefore lacking, which causes a low turn over of this protein. This means that αA-crystallin needs to be long living. Together with the related αB-crystallin it forms large aggregates ranging in size from 300 to 1,000 kDa (Groenen et al., 1994). In certain mammals, a portion of the expressed αA-crystallin possesses an insertion of 23 amino acids. This is encoded in the single-copy αA-crystallin gene by an optional exon that is usually skipped by alternative splicing (King and Piatigorsky, 1983; Hendriks et al., 1988). The function of this alternative splicing product is still unknown, but it has slightly different properties as compared to the normal αA-crystallin (Smulders et al., 1995).

For the study of phylogenetic problems at the higher level of vertebrate relationships, α-crystallin sequences have turned out to be useful because of their slow rate of evolutionary change. Of the two subunits, αA and αB, especially αA-crystallin has extensively been studied, and provided phylogenies that agreed well with commonly accepted relationships. For instance, αA-crystallin sequences of alligator and lizard showed them to be more closely related to avian than to mammalian sequences (de Jong et al., 1985). The position of the red-eared slider turtle (Trachemys scripta elegans) within amniotes was analyzed with the use of both αA- and αB-crystallin sequences. Comparing the turtle sequences with those of birds and mammals provided good evidence that birds, rather than mammals, are the sister group of the turtles (Caspers et al., 1996). In birds, the αA-crystallin protein sequences supported that Ratites (ostrich and related flightless birds) and tinamous (pigeon-like South-American birds) form a monophyletic clade, the Palaeognathae, as the first offshoot of the avian stem (Stapel et al., 1984; Caspers et al., 1994). All other birds, the Neognathae, were set apart from the palaeognaths on basis of two amino acid replacements in αA-crystallin: 122 S→A and 147 Q→P. Amongst the neognathous birds, the Galliformes (fowl-like birds) and Anseriformes (duck-like birds) are the first to branch off, as an unresolved trichotomy. Three synapomorphous amino acid replacements, 127 S→A, 135 S→N and 152 P→A, distinguish the other investigated neognathous orders from the Galliformes and Anseriformes. These five phylogenetically diagnostic replacements are all encoded in the third exon of the αA-crystallin gene. Determining the sequence of the third exon of αA-crystallin, only 207 bp in length, might therefore be suitable for resolving certain questions in avian phylogeny. This is illustrated - in chapter 5 of this thesis - for the enigmatic South American hoatzin, of which it has long been uncertain whether it is more closely related to Galliformes or to the cuckoos (Cuculiformes) (Hedges et al., 1995).

Also phylogenetic problems concerning placental mammals have been analyzed with the aid of αA-crystallin sequences. In chapters 2, 4 and 6, αA-crystallin provides molecular evidence that is pertinent to the following phylogenetic problems: the monophyly of Edentata examined
by a unique deletion of three amino acids (van Dijk et al., 1999), the ‘African clade’ of mammals as revealed by protein sequence ‘signatures’, and the distribution of the optional exon in the different mammalian lineages.

In chapter 3, TNFα and IAPP data sets are used to examine the phylogenetic positions of the guinea pig and rabbit, in relation with the controversy about the monophyly of rodents and Glires. These two genes have not been used earlier for phylogenetic purposes. Functionally, TNFα plays an important role in the defense mechanism against viral, bacterial and parasitic infections, as well as in autoimmune responses. IAPP, a hormone, is the constituent protein of amyloid deposits found in the islets of Langerhans of patients with non-insulin-dependent diabetic mellitus (NIDM). Formation of islet amyloid is associated with progressive destruction of insulin-producing beta cells. Factors responsible for the conversion of IAPP into insoluble fibrils are still unknown.

4. Outline of this thesis

The purpose of the work described in this thesis was to explore the phylogenetic potential of mutationally and evolutionary more complex features, notably indels, protein sequence signatures and alternative splicing. This also required a deeper understanding of the evolutionary process underlying these features.

First, we demonstrate the phylogenetic usefulness of indels using a unique deletion of three amino acid residues present in the αA-crystallin sequences of all investigated edentate species (chapter 2) (van Dijk et al., 1999). The complexity of interpreting indels in phylogeny is illustrated in chapter 3, where an extended data set of TNFα and IAPP is analyzed to explore the wide range and variability of indels.

Then we examine the occurrence of characteristic ‘signatures’ in orthologous protein sequences from different species to determine monophyletic groups of organisms. This type of phylogenetic marker is explored as support for the clade of African mammals, and for the phylogenetic position of the hoatzin within birds as explained in chapter 4 (van Dijk et al., 2001a) and 5 (Hedges et al., 1995), respectively.

In chapter 6 (van Dijk et al., 2001b) we study the molecular evolution of the alternative splicing product αAIns-crystallin originated by exon skipping in the single copy αA-crystallin gene. It is expressed in all studied rodents and in some unrelated mammalian species like hedgehog, pika and bat, while it is lacking in other mammals (Hendriks et al., 1988). This implies that the alternatively spliced exon 2 has independently disappeared or been silenced in many lineages. By studying the presence of exon 2 in species representing all 18 placental mammalian orders, light is shed on the phylogenetic distribution of αAIns-crystallin in mammals.

Finally, in chapter 7 the results presented in this thesis are summarized and discussed.
References


22
Chapter 1

Abbreviations

BLOSUM 62  blocks substitution matrix  
DNA     deoxyribonucleic acid  
IAPP    islet amyloid polypeptide precursor  
Indels  insertions and deletions  
IRBP    interphoto-receptor retinoid binding protein  
JC      Jukes and Cantor nucleotide substitution matrix  
JTT     Jones, Taylor and Thornton amino acid substitution matrix  
KH-test Kishino and Hasegawa test for comparing two trees  
LINEs  long interspersed elements  
Mhc     major histocompatibility complex  
ML      maximum likelihood  
NIDM    non-insulin-dependent diabetic mellitus  
NNI     nearest-neighbor interchanges method for branch swapping  
PAM     percent accepted mutation amino acid substitution matrix  
PAUP    Phylogenetic Analysis Using Parsimony  
PCR     polymerase chain reaction  
RNA     ribonucleic acid  
shps    small heat shock proteins  
SINEs   short interspersed elements  
TBR     tree bisection and reconnection method for branch swapping  
TNFα    tumor necrosis factor alpha  
vWF     von Willebrand factor
The virtues of gaps:
xenarhthran (edentate) monophyly supported
by a unique deletion in αA-crystallin
Abstract

Shared insertions or deletions (indels) in protein-coding DNA can be strong indicators of the monophyly of a taxon. A three-amino acid deletion had previously been noted in the eye lens protein αA-crystallin of two species of sloths and two species of anteaters, which represent the Pilosa, one of the two infraorders of Xenarthra (Edentata). This deletion has not been observed in 55 species from 16 other eutherian orders, or in 2 species of marsupials, or in 34 nonmammalian vertebrates, from birds to shark. At the genomic level, we have now detected this deletion in two species of armadillos of the second xenarthran infraorder, Cingulata, as well as in an additional species of anteater. Phylogenetic trees were constructed from a 145-bp sequence of the αA-crystallin gene of 39 tetrapod species, supporting xenarthran monophyly with values from 76% to 90%. To quantify the additional support for xenarthran monophyly, as given by the three-residue deletion, we computed the probabilities for the occurrence of this deletion per evolutionary time unit for alternative hypothetical tree topologies. In the estimates obtained, the six trees in which the xenarthran subgroups are unresolved or paraphyletic give an increasingly lower likelihood than do the two trees that assume xenarthran monophyly. For the monophyletic trees, the probability that the deletion observed in the xenarthrans is due to a single event is >0.99. Thus, this deletion in αA-crystallin gives strong molecular support for the monophyly of this old and diverse order.
Chapter 2

Introduction

Optimal alignment of homologous DNA or protein sequences often requires the introduction of gaps, or indels, the collective name for insertions and deletions. Such indels tend to occur in variable regions of alignments, which makes their proper positioning and comparison often subjective. To avoid this source of bias, regions containing such sequence gaps are generally removed before further phylogenetic analysis (Swofford et al., 1996). Yet, it is obvious that they are rare and thus constitute a potentially valuable source of phylogenetic information. Because indels are caused by more complex mutational mechanisms than base substitutions, homoplasy by parallel and back mutations - a plague of molecular phylogeny - is less likely to occur. The potential virtues of indels in phylogenetic studies have long been recognized (e.g., Hixson and Brown, 1986; Meyer et al., 1986; Hasegawa et al., 1987; Williams and Goodman, 1989), although the parallel appearance of deletions, as observed in bacteriophage T7, also warrants caution (Cunningham et al., 1997). Various studies have addressed the phylogenetic implications of differential weighting of indels in sequence alignments (e.g., Feng and Doolittle, 1987; Barriell, 1994; Gu and Li, 1995; Wheeler, 1995). However, fewer attempts have been made to more quantitatively express the phylogenetic significance of synapomorphic indels (e.g., Tajima and Nei, 1984; Kishino et al., 1990; Lloyd and Calder, 1991).

We here focus on the phylogenetic potential of a specific deletion occurring in the eye lens protein αA-crystallin. This is a relatively slowly evolving protein for which the 173-amino acid residue sequence is known, completely or partially, in 95 vertebrate species. These include 59 eutherians (representing all orders except Dermoptera), two marsupials, 27 species of birds, a crocodile, a lizard and a turtle, three frog species and a shark (de Jong et al., 1984, 1993; Stapel et al., 1984; Hedges et al., 1995; Jaworski, 1995; Caspers et al., 1997). Only two deletions and no insertions have been found among the 94 tetrapods in this data set. A one-residue deletion is present in Old World monkeys (Macaca mulatta and Colobus guereza; Jaworski, 1995), and a three-residue deletion in the four investigated species of xenarthran mammals, the two- and three-toed sloths (Choloepus didactylus and Bradypus tridactylus), the tamaranda (Tamandua tetradactyla) and the giant anteater (Myrmecophaga tridactyla) (de Jong et al., 1984). These represent the families Megalonychidae (two-toed sloths), Bradypodidae (three-toed sloths) and Myrmecophagidae (anteaters or vermilinguas), together forming the Pilosa, one of the two infraorders of xenarthrans (Simpson, 1945; Wilson and Reeder, 1993; McKenna and Bell, 1997). The second infraorder, Cingulata, has Dasypodidae (armadillos) as the only surviving family. The order Xenarthra (or Edentata) appears to constitute a morphologically well-defined natural grouping (Engelmann, 1985; Rose and Emry, 1993), of which the monophyly is undisputed. However, compelling molecular support for traditionally assumed clades is always desirable (de Jong, 1998). With respect to Xenarthra, molecular studies that include both infraorders and sufficient outgroup taxa are scarce and not always conclusive (Sarich, 1985; Höss et al., 1996; Cao et al., 1998; Stanhope et al., 1998). Confirming by sequence analyses the monophyly of xenarthrans, and resolving their intraordinal relationships, remains therefore an important task. It is of interest to assess whether this deletion also occurs in armadillos, representing Cingulata. If present, it would support xenarthran monophyly; if not, it would support the monophyly of Pilosa.

When using indels in phylogeny reconstruction, it is essential to understand the various underlying mutational processes and constraints. Deletions and insertions often occur at short tandem repeats in the DNA, probably as a result of slipped-strand mispairing, but in other instances the pertinent mutational mechanisms are more complicated (Kunkel, 1990; Krawczak and Cooper, 1991). For noncoding DNA in primates and rodents, it is estimated that indels occur about 10 to 30 times less frequently than nucleotide substitutions (Saitou and Ueda, 1994; Ophir and Graur, 1997). Half of these indels are single-nucleotide gap events (Saitou and
In noncoding as well as coding DNA, the frequency of deletions is always higher than that of insertions, the estimated ratio varying from 1.3 to 4.0 (de Jong and Rydén, 1981; Graur et al., 1989; Golenberg et al., 1993; Saitou and Ueda, 1994; Gu and Li, 1995; Ophir and Graur, 1997).

In proteins, more than half of the indels have a length of one or two amino acid residues (Pascarella and Argos, 1992). Obviously, in protein-coding DNA the fixation of indels is much more constrained than in noncoding DNA. As a consequence, indels in proteins occur about 50 times less frequently than do amino acid replacements (de Jong and Rydén, 1981). Indels can generally be accepted in protein-coding DNA only when they occur as multiples of three nucleotides; otherwise, they would disrupt the reading frame. But even when indels are in frame, they are more likely to disturb the structural integrity of a protein, and hence its evolutionary viability, than are amino acid replacements. Indels therefore rarely occur in regularly structured α-helical or β-sheet regions and are most easily accommodated in less constrained surface loops (Pascarella and Argos, 1992; Benner et al., 1993). Unfortunately for molecular phylogeneticists, these regions generally show the greatest sequence variability, too, making it difficult to conclude whether indels in such regions really constitute genuine homologous character states. Moreover, indels in proteins and mRNAs may result from base substitutions in the splice site consensus sequence, producing intron-exon boundary sliding. Such indels are in principle as prone to homoplasy as other non-synonymous base substitutions, and therefore much less suitable for phylogenetic reconstruction than are the indels in protein-coding genomic DNA sequences.

Thus indels in protein-coding DNA hold more promise for phylogenetic purposes than those in noncoding DNA. Especially a multiresidue indel that can be positioned unambiguously in a multiple alignment, and that does not occur at a splice junction, might be a very reliable phylogenetic marker. This is the case with the three-amino acid residue deletion at positions 147-149 in sloth and anteater αA-crystallin. This amino acid deletion corresponds with a nine-nucleotide deletion located in the middle of exon 3 of the αA-crystallin gene (van den Heuvel et al., 1985) and thus is not the result of exon sliding. Also, around this deletion there are no obvious tandem repeats in the DNA that might have triggered a repeated occurrence of this deletion by slippage. A pitfall to be excluded is the possibility that αA-crystallin molecules with and without the deletions are encoded by paralogous genes. αA-crystallin is a member of the small heat-shock protein superfamily, of which five other, quite distantly related representatives are known to occur in amniotes (de Jong et al., 1998). All evidence indicates that αA-crystallin is encoded by a single-copy gene in amniotes, making paralogy unlikely. Applying then Occam's razor, or any parsimony reasoning, it is most likely that the deletion 147-149 in αA-crystallin of sloths and anteaters is the result of a one-time mutational event in mammalian evolution.

To explore the usefulness of this event in xenarthran phylogeny, we determined its presence in the αA-crystallin genes of two species of armadillos, two species of sloths, and two species of anteaters. The results were added to the existing αA-crystallin DNA data set to model the probability of this deletion under different scenarios, with use of the basic method of Kishino et al. (1990).

Materials and Methods

DNA Amplification and Sequence Determination

Genomic DNA was isolated from frozen or ethanol-preserved liver of six species of xenarthrans (Choloepus didactylus, Bradypus tridactylus, Cyclopes didactylus, Tamandua tetradactyla, Dasypus novemcinctus, Cabassous unicinctus) and two species of marsupials (Macropus rufus, Didelphis marsupialis) obtained from the Collection of Mammal Tissue of
the Université Montpellier. Degenerated primers were designed, using the program Generunner, to amplify 145 bp from the third exon of the αA-crystallin gene, which encode amino acid residues 116-163 in the corresponding protein. Primers with the following sequences were synthesized: 5’-GAC GAC CAT GGC TAC ATT TCC CGN GAR TTY CA-3’ (acry3for) and 5’-TTA GGA CGA GGG TGC YGA GGW NGG CTT YTC YTC-3’ (acry3rev). Polymerase chain reactions (PCRs) were performed in a volume of 50 µl, containing 20 mM Tris-HCl (pH 8.0), 50 mM KCl, 1.5-2.5 mM MgCl₂, 2.5 mM dNTPs (Boehringer Mannheim and Gibco BRL), 100 pmol of each primer, 1 U of Taq DNA polymerase, and 200 ng of genomic DNA as template. The following program gave optimal results: 5 min at 95°C, 30 cycles of 1.5 min at 95°C, 2 min at 54-60°C, and 3 min at 72°C, and a final extension step of 30 min at 72°C. PCR products were ligated directly into a T/A cloning vector (Promega) according to the manufacturer's instructions, followed by transformation into Escherichia coli DH5α. Clones containing αA-crystallin inserts were detected using a 207-bp fragment of the kangaroo αA-crystallin gene as a probe for the filter hybridization. The amplified exon 3 of the αA-crystallin gene was sequenced from at least three independently obtained PCR products, except for D. novemcinctus where only one clone could be obtained. DNA sequences were determined in both directions, using the Sequenase version 2.0 sequencing kit (UBS).

Species Names and Accession Numbers

Full species names and EMBL accession numbers for the αA-crystallin sequences aligned in Figure 1 and used to construct the phylogenetic trees in Figure 2 are (in the order shown in Figure 1): cow (Bos taurus; M26142), flying fox (Pteropus hypomelanus; U24063), round-eared bat (Tonatia silvicola; U24065), Syrian golden hamster (Mesocricetus auratus; X02951), Norway rat (Rattus norvegicus; U47922), house mouse (Mus musculus; J00376), mole rat (Spalax ehrenbergi; M17249), tree shrew (Tupaia glis; U24064), galago (Galago crassicaudatus; U24068), sifaka (Propithecus sp.; U24062), tarsier (Tarsius syrichta; U24066), tamarin (Leontopithecus rosalia; U24067), capuchin (Cebus albifrons; U24059), colobus (Colobus guereza; U24058), rhesus monkey (Macaca mulatta; U24061), chimpanzee (Pan troglodytes; U24057), human (Homo sapiens; U05569), rabbit (Oryctolagus cuniculus; X95382), African elephant (Loxodonta africana; U24060), two-toed sloth (C. didactylus; AJ012436), three-toed sloth (B. tridactylus; AJ012443), tamandua (T. tetradactyla; AJ012442), two-toed anteater (C. didactylus; AJ012437), eleven-banded armadillo (C. unicolor; AJ012438), nine-banded armadillo (D. novemcinctus; AJ012441), red kangaroo (M. rufus; AJ012439), southern opossum (D. marsupialis; AJ012440), hoatzin (Opisthocomus hoazin; U31947), grove-billed ani (Crotophaga sulcirostris; U31940), sooty tern (Sterna fuscata; U31944), rock pigeon (Columba livia; U31945), Eurasian blackbird (Turdus merula; U31942), mallard duck (Anas platyrhynchos; U31946), silver pheasant (Lophura nycthemera; U31943), domestic fowl (Gallus gallus; M17627), elegant-crested tinamou (Eudromia elegans; L25850), red-eared slider turtle (Trachemys scripta elegans; U31938), bullfrog (Rana catesbeiana; X85205) and common frog (R. temporaria; X00716).

Calculating the Likelihood for the Observed Distribution of Deletions

To calculate the likelihood for the observed distribution of deletions, we essentially followed the method of Kishino et al. (1990), which is a modified application of Felsenstein's (1981) algorithm. Our main modifications are that we do not allow reversal of the deletion event and use a clock-like tree. The algorithm used can be explained as follows. The evolution of indels may be modeled as a stochastic process with two possible states for each indel: 0 coding for 'indel absent', and 1 for 'indel present'. For a deletion in the protein-coding region of a gene, it is hard to envisage a mutational mechanism that could precisely reinsert the deleted
nucleotides after the deletion has been fixed in a population. Therefore, although ideally the possibility of a deletion and a compensating insertion should be considered (Kishino et al., 1990), we here assume for practical reasons that reversal of a deletion, $1 \rightarrow 0$, can be excluded. In that case, there remains at each evolutionary time step only one transition possible: $0 \rightarrow 1$ with probability $P$, while the unaltered conditions $0 \rightarrow 0$ and $1 \rightarrow 1$ have probabilities $1-P$ and 1, respectively. For a given phylogenetic tree, it is then possible to calculate the likelihood of the observed distribution of deletions among the living species, i.e., the terminal taxa, by a simplification of Felsenstein’s algorithm (1981). Each branch, internal or terminal, of the tree contributes to the likelihood function in either of two possible ways. If no deletion occurred on that branch, as can be concluded when the deletion is not present in at least one of the living descendants of this branch, the contribution is:

$$v_i = (1-P)^i,$$  \hspace{1cm} [1]

where $l_i$ is the branch length. If a deletion occurred on the branch, as evident from the presence of this deletion in all living descendants of this branch, the contribution is:

$$\delta_i = \sum_{k=0}^{l_i-1} P(1-P)^k = 1 - v_i$$  \hspace{1cm} [2]

The summation over all time steps along the branch is necessary, because the deletion may have occurred at any time. Given a bifurcating tree with two sister taxa having a similar deletion, two possible events may have occurred. One option is a deletion in the ancestral branch (denoted 1) leading to the common ancestor of the two taxa with probability $\delta_1$. The other possibility is that two independent deletion events occurred after divergence on each of the two lineages (denoted 2 and 3) leading to the terminal taxa with probability $v_1\delta_2\delta_3$. So, the contribution to the likelihood is the sum of the probabilities of these events denoted with $\xi_j$:

$$\xi_j = \delta_1 + v_1\delta_2\delta_3.$$  \hspace{1cm} [3]

If the deletion is observed in only one species of a clade, then Equation 3 reduces to 2. In our study, a further summation is required because not two, but six, species have the deletion. That is, even more possible events or series of events must be considered. Besides a single deletion in the common ancestor, six independent deletion events on the peripheral branches, or any intermediate number, also may have occurred. Obviously, it is impossible to know which of these scenarios actually applies (for a similar approach, see Ferris et al., 1979). Therefore, all possible events must be taken into account, and the calculation of $\xi_j$ will be more complex than given in Equation 3 (for a recursive way to do this, see Felsenstein, 1981).

Finally, the likelihood function $L$ for a given tree is obtained by multiplying the different contributions of the branches with and without deletions. This gives:

$$L = \prod_{i=1}^{n} v_i \prod_{j=1}^{m} \xi_j$$  \hspace{1cm} [4]

where $n$ is the number of branches in which no deletion occurred, and $m$ is the number of sums involving deletion events (note that $m$ is also the number of monophyletic groups with the deletion). The maximum likelihood estimate (MLE) of $P$ is the value that maximizes $L$. This value was found by direct iterative optimisation using different values of $P$ up to convergence.
of \( L \) to a maximum. The value of \( L \) was log-transformed before analysis. Because, of course, it is possible to consider not only one but several mutually independent deletions,

\[
L = \prod_{d=1}^{D} \left( \prod_{i=1}^{n} \nu_i \prod_{j=1}^{m} \xi_j \right) \quad [5]
\]

where \( D \) is the number of distinct deletions under analysis.

The discrete time model presented above has a continuous time equivalent as the time intervals go to zero (the approach of Kishino et al., 1990). It is obtained by substituting into Equation 1 with \( \nu_i = e^{-pl_i} \), and into Equation 2 \( \delta_i = 1 - e^{-pl_i} \). We did our computations with both alternatives and did not find any major differences, so here we present only the results for the discrete time version of the model.

**Results**

*New \( \alpha \)-A-Crystallin Sequences and Phylogenetic Tree Reconstructions*

Exon 3 sequences coding for residues 116-163 of \( \alpha \)-A-crystallin were successfully obtained for the armadillos, anteaters, and sloths and - as outgroups - the marsupials, a kangaroo and an opossum (as listed above). These eight newly obtained sequences are aligned in Figure 1 with all corresponding DNA sequences available for the \( \alpha \)-A-crystallin gene in other vertebrate species.

The DNA data immediately reveal that an identically positioned and unambiguously demarcated gap of nine nucleotides must be introduced to align the six xenarthran sequences with the others. This gap can be identified with certainty as a deletion in the xenarthran sequences, because it is not present in any outgroups to the eutherians (i.e., marsupials, birds, turtle and frogs). The most parsimonious inference obviously is to consider this deletion as a shared derived mutational event, having occurred in the last common ancestor of the xenarthrans, after its divergence from the other eutherian lineages. Similarly, a three-nucleotide deletion must be positioned, as noticed earlier (Jaworski, 1995), in the two Old World monkey \( \alpha \)-A-crystallin sequences. In addition to the two deletions, inspection of the nucleotide substitutions in Figure 1 also provides some other obvious phylogenetic information. With regard to xenarthran phylogeny, the positions indicated with arrowheads should be considered, where perfect or almost perfect synapomorphies can be observed for all six xenarthrans. Similarly convincing synapomorphies for each of the three xenarthran subgroups (sloths, anteaters and armadillos) are indicated by asterisks.

The aligned nucleotide sequences were used for neighbor joining (NJ) (Saitou and Nei, 1987), unweighted parsimony (Swofford et al., 1996), and quartet puzzling (Strimmer and von Haeseler, 1996) analyses, of which the consensus tree is shown in Figure 2 (PAUP\(^*\)4.0; Swofford, 1998, was used for the first two types of analysis). The monophyly of Xenarthra is supported, with values of 76% for NJ and parsimony and 90% for ML - which is high, considering the length of these sequences. Moreover, the consensus trees group the two species of sloths as well as the two anteaters together. The NJ analysis gives considerable bootstrap support to the grouping of sloths and anteaters in the Pilosa (80%), but the other two methods give much less support (50% and 35% for parsimony and ML, respectively), making interpretation uncertain. The two armadillos fail to group together, perhaps because of the considerable number of autapomorphic substitutions in exon 3 of *Dasypus novemcinctus* (Figure 1). The other better supported clades in Figure 2 are in reasonable agreement with prevailing opinions about vertebrate relationships. Exceptions are the grouping of sifaka - a
### Figure 1.
Alignment of a 145-bp sequence from exon 3 of the αA-crystallin gene from 39 tetrapod species. The amino acid residues encoded by the bovine DNA are given above the alignment, together with their residue numbering in the αA-crystallin protein. The gray blocks highlight the deletions (-) in xenarthran and Old World monkey αA-crystallin; (.) indicates identity with the top sequence. Undetermined sequences at the 5’ and 3’ ends in certain species are left open. Arrowheads below the alignment indicate the positions of xenarthran synapomorphies; asterisks indicate the positions of synapomorphies for the three xenarthran subgroups. Sequences were aligned with the PILEUP program from the GCG package (Devereux et al., 1984) and modified with the DIF program (written by P.J.L. Wreten, 1996). Full species names and EMBL accession numbers are given in Materials and Methods. This data set is available at www.utexas.edu/ftp/depts/systbiol/.
Figure 2. Consensus tree derived from the aligned DNA sequences in Figure 1. Phylogenetic trees were constructed with neighbor joining using observed distances, and maximum likelihood (HKY model) by using quartet puzzling. For neighbor joining, all nucleotide positions were used, the deletions and terminal open sequences being filled in by N. The consensus tree shows only those clades that are supported by bootstrap values of ≥ 30% with both methods, of which at least one was ≥ 50%. Support values of neighbor joining and quartet puzzling, for 1,000 and 10,000 replicates, respectively, are shown on the branches (top number = neighbor joining; bottom number = quartet puzzling). Arrows indicate the clades that are supported by synapomorphic deletions.
lemur - with flying fox - a megabat - and the joining of mouse with hamster rather than with rat.

Although it is satisfying that some phylogenetic signal seems present in a 145-bp sequence of the \(\alpha\)-crystallin gene, such short sequences have hardly any weight in an era where 17,000-bp mitochondrial genomes, or kilobases of nuclear genes, have become ‘the new minimum for studying vertebrate relationships’ (Penny and Hasegawa, 1997). We therefore wanted to better exploit the added phylogenetic significance of the xenarthran three-residue deletion.

**Calculating the Likelihood of the Deletion Marking Xenarthran Monophyly**

To compute the MLE of the probability \(P\) for the deletion of residues 147-149 to occur per unit evolutionary time, we used the topologies and branch lengths from eight different phylogenetic trees, as presented in Figure 3. These trees are hypothesized alternative relationships for the same 39 taxa included in our sequence analysis. In these trees, branch lengths are mainly estimated from the fossil record, complemented with molecular data (Springer, 1997; Kumar and Hedges, 1998). Trees 1 and 6 have topologies that are composites of prevailing molecular and morphological hypotheses about mammalian and avian relationships. Both assume a monophyletic xenarthran clade, in tree 1 positioned as a sister group of artiodactyls (Arnason et al., 1997), and in tree 6 as a sister group of all remaining eutherian orders (morphologically still the favored option; McKenna and Bell, 1997). Trees 2-5, and trees 7 and 8 are variants of trees 1 and 6, respectively, hypothesizing alternative unresolved and paraphyletic origins of the three xenarthran subgroups. The trees derived from our DNA sequences (Figure 2) can also be used for the likelihood analysis; in that case, \(P\) is assumed to be a fixed proportion of the rate of nucleotide substitution (Kishino et al., 1990). However, the likelihoods from these trees cannot be compared with the likelihoods computed from the trees with branch lengths derived from paleontological data because the scaling is not comparable.

The trees with the largest likelihoods are the ones that are supported by the deletion data. Table 1 presents the values obtained for \(P\) and the corresponding log likelihoods. The likelihood values for trees 1 and 6, which assume xenarthran monophyly, demonstrate convincingly and quantitatively that they much better fit the observed distribution of the deletion than do the other trees. Since \(\log L\) depends on the estimated value of \(P\), which may

<table>
<thead>
<tr>
<th>Tree</th>
<th>(P(10^{-4}))</th>
<th>(\log L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.2</td>
<td>-6.8</td>
</tr>
<tr>
<td>2</td>
<td>6.7</td>
<td>-11.2</td>
</tr>
<tr>
<td>3</td>
<td>9.6</td>
<td>-12.8</td>
</tr>
<tr>
<td>4</td>
<td>9.7</td>
<td>-13.2</td>
</tr>
<tr>
<td>5</td>
<td>9.6</td>
<td>-13.3</td>
</tr>
<tr>
<td>6</td>
<td>3.1</td>
<td>-5.5</td>
</tr>
<tr>
<td>7</td>
<td>9.3</td>
<td>-11.1</td>
</tr>
<tr>
<td>8</td>
<td>9.3</td>
<td>-11.0</td>
</tr>
</tbody>
</table>

\(^a\) The relevant parts of the trees are shown in Figure 3.

\(^b\) Maximum likelihood estimate of the probability of occurrence of the three-residue deletion per unit evolutionary time.

\(^c\) Corresponding value of the maximum log likelihood obtained with Equation 4 after log transformation.
Figure 3. Alternative hypothesized tree topologies used for the likelihood computations given in Table 1. Trees 1 and 6 assume xenarthran monophyly, placing them as a sister group of Ferungulata - (represented here by cow) (Arnason et al., 1997), and Epitheria - all remaining eutherians (e.g., McKenna and Bell, 1997), respectively. Apart from the positioning of the xenarthran clade, the remaining parts (not shown) of trees 1 and 6 are identical. Divergence of the presented parts of the trees from the other eutherians and from marsupials is taken as 90 and 130 MYA in trees 1 and 6, respectively. Trees 2 to 5 are variants of tree 1, in which various unresolved and paraphyletic origins of the three xenarthran groups are assumed; trees 7 and 8 are the corresponding variants of tree 6. The complete branching patterns and branch lengths of the trees are available with the data file at www.utexas.edu/ftp/depts/systbiol/.
have been considerably over- or underestimated as compared with the 'correct' value, we also calculated for all trees the log$L$ values for a range of values of $P$. Although for values of $P$ between 0 and 0.003, the monophyletic trees 1 and 6 apparently should always have the highest likelihoods, actually, in our estimated range of values, for $P$ between 0.0003 and 0.0010, the log$L$ values for all trees remain quite constant. Only at unrealistically high values of $P → 0.03$ – do paraphyletic trees get higher likelihoods than monophyletic trees. This agrees with the intuitive expectation that a low value of $P$ (i.e., a rare deletion) implies a monophyletic tree.

The present model does not assume that the observed occurrence of the deletion 147-149 in the six xenarthrans is due to a single event. But, given a phylogenetic tree and a value of $P$, one can compute the probability that this deletion occurred only once during evolution. Note that this probability, which we denote $p_1$, is not a parameter estimated in the same way as $P$, but is a probability that is calculated a posteriori. A value of $p_1$ close to 1 for a tree constructed by assuming xenarthran monophyly would be evidence for a unique deletion event. The values of $p_1$ for these trees are 0.995 and 0.998, respectively, again quantitatively expressing that a single origin of the deletion is by far more probable than multiple origins. Also for our (otherwise not comparable) NJ tree, this value of $p_1$ was almost equally high, 0.991.

**Discussion**

Molecular phylogenetic studies involving Xenarthra are as yet rare. Albumin immunology was unable to support xenarthran monophyly and could not resolve the lineages to sloths, anteaters, and armadillos, which were suggested (Sarich, 1985) to have separated at least 75-80 million of years ago (MYA). Mitochondrial 12S rRNA, tRNA-valine and 16S rRNA sequences gave 99% bootstrap support to a clade of armadillos and sloth (Stanhope et al., 1998), but 12S and 16S rRNA could not resolve the relationships of armadillos, anteaters, and sloths (Höss et al., 1996). Finally, mitochondrial ND1 only marginally supported xenarthran monophyly and challenged - in contrast with our results - the naturalness of the Pilosa by grouping armadillos, rather than anteaters, with sloths (Cao et al., 1998). Our αA-crystallin sequence data, and especially the three-amino-acid deletion, provide further support for xenarthran monophyly.

Various approaches can be envisaged to use indels in phylogenetic analyses. Kishino et al. (1990) used an application of the maximum likelihood algorithm of Felsenstein (1981). They estimated branch lengths for different trees based on the substitution rate of proteins, inferred the probability of insertion and deletion events, and considered the likelihood of single or double origins of indels. Also other authors have extended the use of modeling deletions to alignments, although only pairwise comparisons are presently computationally feasible (Bishop and Thompson, 1986; Thorne et al., 1991, 1992). An extension of this approach to obtain an improved computation of pairwise genetic distances has been presented by Thorne and Kishino (1992). Allison and Wallace (1994) developed a method based on resampling and simulated annealing to perform alignment of more than two taxa. Our approach is basically the method of Kishino et al. (1990), except that edge lengths are estimated by using inferred divergence times from the fossil record, complemented with molecular data. It is also inspired by the models introduced by Ferris and Whitt (1978) for evaluating the loss of duplicate genes. The present probabilistic analysis of indels is not designed to directly infer a phylogeny but to measure the support given by a specific deletion to alternative phylogenies inferred from other evidence.

As evident from Table 1, the support for xenarthran monophyly from our data is very high. Yet, our approach used only part of the available evidence on the occurrence of the deletion. At the amino acid sequence level, the presence of deletion 147-149 is known in seven xenarthran species, including a third anteater (*M. tridactyla*; de Jong et al. 1984), and the absence has been ascertained in a total of 90 other tetrapods, compared with the 6 xenarthran and 33 other
tetrapod species included in the present DNA sequence study. Even without any further computation, it is obvious that by adding more species without the deletion, the number of branches without deletion will increase in the likelihood estimates. As a consequence, $P$ (of a deletion per unit time) will decrease, which strengthens the evidence for xenarthran monophyly even further. On the other hand, differences in branching patterns and branch lengths of the trees that are assumed in our computations obviously affect the obtained likelihood estimates. Notably, the lengths of the last common ancestral branch towards a monophyletic xenarthrans and of the branches after the xenarthran radiation do have a considerable impact. However, we have taken realistic time estimates for these branches: 10 and 35 MYA for the last common xenarthran ancestral branch in trees 1 and 6, respectively, and the successive divergences within Xenarthra beginning 75 MYA ago in both trees (Sarich, 1985; see also Höss et al., 1996, and Cao et al., 1998, who corroborate such a date with independent data and methods).

The model used here is a relatively simple one and does not take into account several variables that are likely to influence the outcome. A three-residue deletion is much rarer than a one-residue deletion and thus should be given a greater weight. Insertions are rarer than deletions, but potentially reversible, and thus should be modeled differently. The ratio of occurrence of indels to base substitutions, variations in evolutionary rate, the length and the degree of variability of the analysed DNA fragment in which the indel occurs, the presence of short DNA repeats, and the adjacency of splice junctions can all be envisaged to have some impact. One can readily see two potential extensions of the modeling approach followed in this paper. First, as done by Kishino et al. (1990), one could consider the branch lengths $l_i$ as parameters of the model, and their values could be searched to maximize $L$ similarly to the maximum likelihood analysis of DNA sequences (Felsenstein, 1981). Second, the use of a stochastic model for a character to assess the likelihood of alternative phylogenies may be a relevant approach to the problem of combining data in phylogenetic analysis (e.g., Kishino et al., 1990; Huelsenbeck et al., 1996). Of course, the model used here for deletions can in fact be applied to any character that can be thought of as irreversible.

Taking indels into account in phylogenetic analyses adds weight to sequence-based tree reconstructions. However, suitable indels are rare, and thus can provide only a limited phylogenetic resolution. But whenever they are traced, they tend to form excellent markers to identify monophyletic groups.

**Acknowledgements**

We greatly appreciated valuable comments by Jeff Thorne, Peter Waddell, Masami Hasegawa, and an anonymous reviewer on various versions of this manuscript. The collection of xenarthran samples in French Guiana 1994-1995 was funded by Electricité de France under contracts ‘Faune Sauvage’ (to J.-C. Vié) and GP-0850 (to F.C.). We thank Collin van Asten for useful suggestions and Paul Werten for use of the program DIF. This study was supported by grants from the Netherlands Foundation for Life Sciences (SLW 805-41.042) and from the European Commission (HCM CHRXT930254) to W. de J.
References

CHAPTER 3

Indels in tumor necrosis factor α and islet amyloid precursor protein indicate that rodents are monophyletic and lagomorphs are their sister group
Abstract

During protein evolution, insertions and deletions (indels) occur less frequent than amino acid replacements, being the result of more complex and constrained mutational events. Therefore, indels are less prone to homoplasy, and may be valuable in reconstructing phylogenetic relationships. We here explore the phylogenetic potential of indels present in mammalian tumor necrosis factor alpha (TNFα) and islet amyloid polypeptide precursor (IAPP). Available TNFα and IAPP data sets were extended with newly determined sequences to possibly answer questions pertaining to rodent and lagomorph relationships. Morphologically, the monophyly of the order Rodentia and their sister group relation with the Lagomorpha, together forming the Glires, is undisputed. In contrast, the molecular evidence concerning these groupings is controversial. The sequences in the TNFα data set, representing 24 placental and 3 marsupial species, were analyzed to obtain the best supported molecular tree, with gaps being treated as missing data. The resulting maximum likelihood tree suggests rodent monophyly, and groups lagomorphs with rodents. Using the indels observed in TNFα as an independent source of phylogenetic information, supports these same relationships. Also the limited taxon sampling of IAPP suggests the presence of an indel characteristic for rodents. Sequence alignments aid in visualizing indels that may be suitable for identifying monophyletic groupings.
Introduction

Rare genomic changes offer an independent way of evaluating alternative phylogenetic hypotheses in cases where primary sequence data generate conflicting or equivocal results (Rokas and Holland, 2000). Several of such unique molecular markers, like long and short interspersed elements (LINEs and SINEs; e.g., Shimamura et al., 1997; Serdobova and Kramerov, 1998; Hillis, 1999; Huchon and Catzeflis, 1999), sequence signatures in proteins (e.g., Luckett and Hong, 1998; Huchon et al., 2000; van Dijk et al., 2001) and insertions/deletions (indels; e.g., van Dijk et al., 1999; Groth and Barrowclough, 1999; Matthee et al., 2001), have demonstrated their usefulness in analyzing vertebrate phylogeny. Although indels, especially in protein-coding sequences, thus may constitute a potentially valuable source of phylogenetic information, the corresponding gaps in sequence alignments are generally removed before further analysis (Swofford et al., 1996). This is done because these gaps must often be introduced in regions with the greatest sequence variability or with sequence repeats. As a consequence, the positioning of indels within a multiple alignment is often ambiguous, and their value as genuine homologous characters uncertain. Also, the mutational processes and selective constraints involved in the origin and evolution of indels are more complex and less understood than is the case for simple base substitutions. Slipped-strand mispairing and intron-exon boundary sliding account for most of the shorter indels in proteins (Graur and Li, 1999), and obviously have different modes of evolution.

To explore the phylogenetic value of the various types of indels, we searched for protein data sets containing indels that might be informative in reconstructing mammalian phylogeny. We found that published sequences of tumor necrosis factor alpha (TNFα) and islet amyloid polypeptide precursor (IAPP) contain indels that might be useful for studying problems in rodent phylogeny. This concerns especially the persistent uncertainty about rodent monophyly and the sister group relationship between rodents and lagomorphs. Morphologically, there is not the slightest doubt about the monophyly of rodents, and also their grouping with lagomorphs in the cohort Glires is generally accepted (Novacek, 1992; Luckett and Hartenberger, 1993; McKenna and Bell, 1997). However, ever since rodent monophyly was questioned, on basis of protein sequence data (Graur et al., 1991), the molecular evidence remains ambiguous. The latest and most comprehensive sequence studies either maintain that rodents are para- or polyphyletic (Reyes et al., 2000), are ambiguous in this respect (Mouchaty et al., 2000; Cao et al., 2000), or support their monophyly (Penny et al., 1999; Waddell et al., 1999; Madsen et al., 2001; Murphy et al., 2001). The Glires clade is generally not supported molecularly, although larger nuclear data sets increasingly tend to position the lagomorphs close to or with rodents (Madsen et al., 2001; Murphy et al., 2001). Also some mitochondrial analyses do not exclude the association of lagomorphs with rodent lineages (Penny et al., 1999; Reyes et al., 2000).

We complemented the available data sets of TNFα and IAPP with eleven and two newly determined sequences, respectively. The nucleotide sequence alignment of the expanded TNFα data set was used to reconstruct the best supported phylogenetic tree, gaps being treated as missing data. Both the resulting maximum likelihood tree, and the TNFα and IAPP indels provided independent sources of information to get more insight in rodent and Glires phylogeny. The data may help restore the morphologists' confidence in the potential of molecular evidence (Gura, 2000).

Materials and methods

Selection of proteins with promising indels

TNFα and IAPP were selected for this study after searching the data bases for protein-coding genes containing indels that might potentially be informative to resolve eutherian
ordinal phylogeny. Such genes had to meet the following criteria. Sequences from at least four different mammalian orders had to be available, considering Hystricognatha and Myomorpha as separate orders. These sequences should preferably represent at least three eutherian orders and a marsupial outgroup, but four eutherian orders without a suitable outgroup were also considered. A list of protein coding genes from Hovergen (Duret et al., 1994) served as a guide for the retrieval of a set of homologous sequences from the EMBL, GenBank, SWISSPROT and PIR data bases united in the Sequence Retrieval System (SRS). Multiple alignments were made for the 41 selected protein data sets, using ClustalW (Thompson et al., 1994), and visually inspected for the presence of indels. To be potentially informative for phylogenetic purposes, only those protein data sets were retained that had at least one indel with precisely the same position and length in species from at least two eutherian orders. Moreover, the positioning of the indel in the protein alignment should be reasonably unambiguous, that is, not located in regions with extensive sequence repeats or extreme sequence variation, but flanked by well conserved sequences. Finally, for practical reasons, it had to be feasible to design degenerate primers in the regions surrounding the indels. Only two proteins were found to comply with these requirements, TNFα and IAPP, and both had indels of relevance for rodent and Glires phylogeny.

**Sequence determination of TNFα and IAPP**

New TNFα sequences were obtained from squirrel (Sciurus vulgaris; AJ286824), guinea pig (Cavia porcellus; the guinea pig sequence analyzed by us is identical to the meanwhile submitted sequence in the data base: P51435), pika (Ochotona princeps; AJ286825), tenrec (Tenrec ecaudatus; AJ286826), sloth (Bradypus tridactylus; AJ286827), anteater (Cyclopes didactylus; AJ286828), armadillo (Cabassous unicinctus; AJ286829), hedgehog (Erinaceus europaeus; AJ286830), mole (Talpa europaea; AJ286831) and two outgroups, opossum (Didelphis virginiana; AJ286832), and kangaroo (Macropus rufus; AJ286833). The protein-coding sequence (705 bp) of the TNFα gene is interrupted by 3 introns, and extends over a total length of about 1638 bp in the rat. The genomic sequences, coding for amino acid residues 10-228 (rat TNFα numbering), were therefore amplified with two separate primer sets, using the Expand High Fidelity PCR system (Boehringer Mannheim) on 50-250 ng genomic DNA as template. The first DNA fragment of about 1351 bp, and coding for amino acid residues 10-133 (with introns at positions 187-684, 740-887 and 946-1219), was amplified with the forward primer 5'-ATG-AGC-ACA-GAA-AGC-ATG-ATC-CGC-GA-3' and the reverse primer 5'-CCT-GGG-AGT-AGA-YRA-GGT-ACA-RNC-C-3'. Optimal results were obtained with an initial cycle of 5 min at 95°C, 3 min at 58-62°C and 3 min at 72°C; followed by 30 cycles of 1 min at 95°C, 1 min at 58-62°C and 3 min at 72°C, and a final extension step of 30 min at 72°C. When the protein-coding sequence could not completely be determined, then the internal primers 5'-TGC-TAC-AAC-ATG-GGC-TAC-AGG-3', 5'-NCC-RTT-GGC-CAG-GGC-RTT-3' or 5'-TGC-TAC-GGG-CTT-GTC-ACT-CGG-3' were used to complete the sequence. The second DNA fragment, about 309 bp, and corresponding to amino acid residues 102-228 within exon 4, was amplified with the forward primer 5'-ATG-AGC-ACA-GAA-AGC-ATG-ATC-CGC-GA-3' and the reverse primer 5'-CTG-GGG-CTT-GTC-ACT-CGG-3'. Reactions were carried out using the following program: 5 min at 95°C, 30 cycles of 1 min 95°C, 30 s at 50-55°C, 30 s at 72°C, and finally, 30 min at 72°C. For sequence determination, the PCR products were cloned into the T/A vector (Promega), and two or if necessary three independent clones were sequenced using the Thermo Sequenase kit (Amersham Pharmacia).

TNFα sequences from the following species were obtained from the data bases: brush-tailed possum (Trichosurus vulpecula; P79374), white-footed mouse (Peromyscus leucopus; P36939), mouse (Mus musculus; P06804), rat (Rattus norvegicus; D00475), hamster (Mesocricetus auratus; AF315292), woodchuck (Marmota monax; O35734), rabbit...
(Oryctolagus cuniculus; M12846), rhesus macaque (Macaca mulatta; P48094), human (Homo sapiens; P01375), cat (Felis silvestris catus; P19101), dog (Canis familiaris; X94932), horse (Equus caballus; P29553), sheep (Ovis aries; P23383), red deer (Cervus elaphus; P51743), bovine (Bos taurus; AF011926), and pig (Sus scrofa; P23563).

New cDNA sequences for IAPP, corresponding with amino acid residues 1-70 in the mouse sequence (total length 93 residues), were determined for rabbit (Oryctolagus cuniculus; AJ286814) and hedgehog (Erinaceus europaeus; AJ286815). The kangaroo (Macropus rufus; AJ286813) sequence was partially obtained, coding for amino acid residue 46-70. First, genomic sequences, coding for amino acid residues 46-70, were amplified with the forward primer 5’-AAR-TGC-AAC-ACD-GCC-ACR-TGT-GC-3’ and reverse primer 5’-CTC-YTK-CCR-TAY-GTR-TTS-GA-3’. Using 100-200 ng genomic DNA as template, reaction products were obtained with the following program: 5 min at 94 °C, 30 cycles of 1 min at 94 °C, 2 min at 50 °C, and 30 s at 72 °C, and finally an extra extension step of 30 min at 72 °C. Then, 1 µl of these PCR products from gel was reamplified to obtain specific products with the next program: 5 min at 95 °C, 30 cycles of 1 min at 95 °C, 1 min at 60 °C, and 1 min at 72 °C, and finally an extra extension step of 30 min at 72 °C. Three specific primers that comprise the first fragments, were designed for the hedgehog and rabbit, respectively, 5’-ACA-TCA-GTA-GGC-GAG-AGA-ATG-GCA-C-3’, 5’-CAG-GTT-GTT-GCT-GGA-ACG-AC-3’, 5’-ATT-TAC-CAG-GCG-TTG-AGT-TG-3’ and 5’-CGT-GTT-GGA-TCC-CAC-ACT-GG-3’, 5’-GCC-ACC-AAA-GTT-GTT-GCT-GGA-ATG-3’, 5’-AAT-CTG-CCA-GCC-AGG-CGT-GTG-GG-3’. After total RNA isolation from the pancreas with TRIzol (GibcoBRL), these primers were sequentially used to obtain the 5’end sequences with the RACE protocol according to Boehringer Mannheim. Sequence analysis was as for TNFα.

Additional IAPP sequences were extracted from the data bases: guinea pig (Cavia porcellus; M25387), degu (Octodon degus; M57669), golden hamster (Mesocricetus auratus; X56067), mouse (Mus musculus; M25389), rat (Rattus norvegicus; M25390), dog (Canis familiaris; M59998), cat (Felis catus domesticus; M25388), human (Homo sapiens; NM_000415), and the chicken (Gallus gallus; L16955) as outgroup.

Phylogenetic analysis

Alignments of the 27 TNFα DNA sequences were made by the PILEUP program from the GCG package (Devereux et al., 1984). The evolution model best fitting the TNFα data set according to the program Modeltest version 3.06 (Posada and Crandall, 1998) was the general-time-reversible model with substitution rates assumed to follow a gamma distribution with a shape parameter of 1.6368. This model was used in reconstructing the maximum likelihood tree. Maximum likelihood bootstrap support values were determined with 100 replications and heuristic searches by using tree-bisection-reconnection branch swapping. Starting trees were obtained via the neighbor joining method. Branch lengths with ML distances were estimated for the obtained consensus tree. All phylogenetic analyses were performed with PAUP 4.0b8 (Swofford, 1998). Indels were treated as missing data, or the regions comprising the indels were removed from the alignment before the analyses. In both cases the indels thus did not contribute to the obtained topology, and can independently be explored as potential markers for supporting phylogenetic inferences.

Results

Types of indels in TNFα and IAPP

The regions containing relevant indels in the amino acid sequences deduced from the newly obtained DNA sequences of TNFα and IAPP are aligned with their corresponding sequences from the data base as shown in Fig. 1. Five of the eleven indels detected in TNFα, and both
Figure 1. Alignments of TNF (A) and IAPP (C) amino acid sequences around phylogenetically informative indels. Phylogenetically informative indels are shaded, and numbered under the alignments. Asterisks mark the species for which genomic TNF informative indels are known. Positions in the protein sequences. Note that at several positions two different types of apomorphic replacements fulfil the criteria for the applied signature search procedure (e.g., 60H and 60P). The inserted codons 1, between 2nd and 3rd codon positions, and introns sequence signatures in TNF (in black). Numbers above the alignments indicate residue positions in the protein sequences. Note that at several positions two different types of apomorphic replacements fulfil the criteria for the applied signature search procedure (e.g., 60H and 60P).

A.  | 51  |
---|---|
0  |   |
1  |   |
2  |   |
3  |   |
4  |   |

B.  | 111111122 |
---|----------|
0  |   |
1  |   |
2  |   |
3  |   |
4  |   |
5  |   |

C.  | 1  |
---|---|
0  |   |
1  |   |
2  |   |
3  |   |
4  |   |
5  |   |
**Figure 2.** DNA alignments around the indel positions for TNFα of those mammalian species for which genomic sequences are available (A, B and C), and for IAPP of all species shown in Fig. 1 (D and E). Protein-coding sequences are shaded. Asterisks mark the species for which genomic TNFα and IAPP sequences were newly determined. A, exon 2 and flanking intronic regions of the TNFα gene; B, intron 2/exon 3 splice site of the TNFα gene; C, part of exon 4 of TNFα; D, 5′-UTR and initiation ATG (bold) of the IAPP cDNA; E, intron 2/exon 3 splice site of the IAPP gene (intronic sequences only known for mouse, rat and human). Indel positions and numbering (as in Fig. 1) are given below the alignment. Splice site consensus AG and GT are in bold and italics. The numbering above the alignments corresponds with that for the amino acids in Fig. 1. Arrows indicate the positions of some crucial mutations as discussed in Results. Repeats suggested to be responsible for slipped-strand mispairing are underlined.
insertions in IAPP were present in more than one species, and can therefore be phyllogenetically informative. The polarity of the indels in TNF\(\alpha\) and IAPP could be established by comparison with the marsupials and the chicken as outgroup, respectively.

Deletion 1 in TNF\(\alpha\) (Fig. 1a) is the result of intron sliding, as can be inferred from the DNA alignment for those species for which intronic sequences are known (Fig. 2a). This sliding results from a G\(\rightarrow\)A substitution in the consensus AG of the intron 1 acceptor splice site in squirrel, human, hedgehog, cat and pig, which in squirrel is additionally affected by an A\(\rightarrow\)C mutation. As a consequence, the 3-bp downstream AG is used as a new 3' splice site in these species, resulting in the deletion of a CAG codon in the mature mRNA, and thus of a glutamine residue at the protein level. Also indels 3 and 4 in TNF\(\alpha\) - both insertions - are close to an intron position (indicated by arrowheads in Fig. 1a). However, neither is due to intron sliding, as shown in Figs. 2a and 2b. The DNA alignment rather suggests that these insertions 3 and 4 have originated by slipped-strand mispairing and consequent duplication of possibly ACACTC and TCT, respectively, corresponding at the protein level with duplicated TL and S, respectively (see e.g. rat in Fig. 1a). Also indel 2 is likely to be caused by slipped-strand mispairing, resulting in deletion of one triplet of a TCATCA repeat, as still present in rat and mouse (Fig. 2a). Similarly, the NCCNCNC repeat as present around indel 5 (see e.g. human and hedgehog, Fig. 2c), may well have facilitated the formation of this one-amino acid deletion.

IAPP of rat and mouse have an additional residue at the N-terminus (Fig. 1c), due to the generation of an ATG start codon immediately upstream of the ATG as present in the other species (Fig. 2d). Interesting is the persistence of two consecutive ATG triplets in the mouse IAPP gene, a feature that is generally avoided (Saito et al., 1999). The second indel in IAPP is a three amino acid insertion (Fig. 1c), apparently caused by intron sliding (Fig. 2e). As compared to human, the genomic DNA sequences of mouse and rat – and by inference hamster, guinea pig and degu as well- have here a splice site consensus AG nine bp upstream. Intron sliding at this position was earlier suggested by Nishi et al. (1989).

Sequence-based TNF\(\alpha\) trees and additional information from indels

Fig. 3 shows the ML tree obtained from the TNF\(\alpha\) DNA sequences, gaps being treated as missing data. The tree suggests rodent monophyly as well as the Glires clade, albeit with low support. All better supported clades (>50%) appear biologically realistic, such as a monophyletic Eulipotyphla, Primates, Carnivora and Artiodactyla. Where the ML tree deviates from reasonable expectations, as by grouping tenrec with Glires, and Eulipotyphla with Artiodactyla, this is only poorly supported.

It can be seen from Fig. 1 that indels 1, 2, 3 and 5 in TNF\(\alpha\) may provide additional information concerning the positioning of rodent and lagomorph taxa in the eutherian tree; indel 4 only occurs in hedgehog, and is not relevant in this respect. The species distribution of the various indels is indicated for the maximum likelihood tree of TNF\(\alpha\) (Fig. 3). Indel 1, the only indel in TNF\(\alpha\) supposedly caused by intron sliding, is clearly homoplasious. This is not too surprising, considering that intron sliding can be caused by a single base substitution. The other indels have supposedly originated by slipped-strand mispairing, which might be a rarer event than base substitution. Indeed, indel 5 is unique for the Glires clade. Also indel 3 might be a synapomorphous insertion for Glires, but then requires a deletion at the same location in pika. Alternatively, indel 3 occurred in parallel in the rabbit lineage and in the rodent ancestor. The same uncertainty holds true for indel 2; considering that guinea pig and squirrel/woodchuck represent two morphologically and molecularly disparate clades within Rodentia (Hystricognathi and Sciuroidea, respectively; Murphy et al., 2001; Adkins et al., 2001; Debry and Sagel, 2001), it is unlikely that indel 2 is a synapomorphy for these taxa. Yet, it is remarkable that indels 2 and 3, despite the fact that they may be prone to homoplasy, occur in lineages that are in phylogenetic proximity. It suggests that sequence conditions for the
Figure 3. Maximum likelihood tree based on TNFα DNA sequences, presented with branch lengths and support values. When the ML tree was constructed after deleting the gap regions from the alignment, rather than considering gaps as missing data, an identical topology (apart from the position of Xenarthra) and very similar support values (underlined) were obtained. The presence or absence of the 5 indels in TNFα and the 2 indels in IAPP (as far as taxon representation allows) is indicated (cf. Fig. 1).

Specific underlying slipped-strand mispairings are only favorable in a restricted region of the gene tree.

As for the IAPP data, indel 1 agrees with murid monophyly, and indel 2 with rodent monophyly. Considering the very restricted taxon sampling, these findings have as yet only a limited significance.

Amino acid sequence signatures

In addition to indels, we screened the TNFα protein alignment for sequence signatures that might be unique for rodents and Glires. For a detailed description of this approach and its background, see van Dijk et al. (2001; Chapter 4). The outgroup residue(s) of the brush-tailed possum, opossum and kangaroo are considered as the plesiomorphous condition. Searching for amino acid sites among the 24 eutherian TNFα sequences that might support the monophyly of rodents (7 species in the data set) and Glires (9 species), the search window was set at 8 ± 2. Thus, positions at which 6-10 ingroup species share the same apomorphy are then candidates for any rodent or Glires sequence signature. From the alignment of 221 amino acid residues a
Indels in TNFα and IAPP

A signature of five amino acids was found for the rodents (60P, 94E, 143Y, 172P, 206V), of which two positions (143Y, 206V) also included rabbit and pika (Fig. 1b). This again suggests that the rodents are monophyletic, and indicates that lagomorpha are their sister group. IAPP was excluded from a signature search because this data set has poor species sampling.

Discussion

The indels that we visualized in TNFα and IAPP (Fig. 1a,c) provide new evidence for rodent monophyly as sister group of Lagomorpha, together forming the cohort Glires. Various studies have already shown the additional information of indels in protein-coding DNA to identify monophyletic groups. Regular phylogenetic analyses of RAG-1 gene sequences separate the chicken- and duck-like birds (Galloanserines) from a clade consisting of all other neognathous birds (Plethornithines). This divergence is additionally supported by a unique 15-bp deletion present in the plethornithine clade (Groth and Barrowclough, 1999). A unique 9-bp deletion in exon 11 of the BRCA1 gene is present in Afrotheria, a clade of placental mammals that includes elephant shrews (Macroscelidea), golden moles (Chrysochloridae), tenrecs (Tenrecidae), aardvark (Tubulidentata) and paenungulates (elephants, sea cows, and hyraxes) (Madsen et al., 2001). It thus adds weight to nuclear and mitochondrial sequence-based tree reconstructions (Springer et al., 1997; Stanhope et al., 1998), and to protein sequence signature results (van Dijk et al., 2001). Another 15-bp deletion, also in the BRCA1 gene, supports a fundamental division in the microchiropteran bats (Teeling et al., 2001). Nevertheless, despite the increasing acceptance of indels to infer relationships between living organisms, their impact remains inferior as compared to the dominance of primary sequence data, mostly due to the fact that the phylogenetic information of gaps is difficult to assess statistically and quantitatively.

The phylogenetic tree based on DNA sequences of TNFα again supports rodent monophyly (Fig. 3), with a bootstrap value of 68%. However, the evidence for Lagomorpha as closest relatives of Rodentia is less convincing, with only 33% support. The phylogenetic evidence from the indels in TNFα and IAPP is consistent with the topology of this tree, and could easily be deduced just by inspecting the sequence alignment. A further statistical analysis of the phylogenetic significance of these indels was not attempted because this would not add much stronger evidence. The inclusion of indel events in phylogenetic analysis started with the computational models for pairwise alignment by Bishop and Thompson (1986), Thorne et al. (1991, 1992) and Thorne and Churchill (1995). Their models assume that molecular sequences consist of short fragments of one or more characters joined by imaginary links that operate by a birth-death process. Although this approach can be used to model sequence evolution in a phylogenetic tree, standard likelihood calculations would be intractable. Another method, that of Kishino et al. (1990), is based on the maximum likelihood algorithm of Felsenstein (1981). Branch lengths for alternative trees are estimated from the substitution rate of proteins. An application on this method is used in our former study (van Dijk et al., 1999; Chapter 2) which estimated branch lengths from fossil divergence times complemented with molecular data. Both these methods measure the support given by a specific indel to alternative phylogenies inferred from other evidence.

In contrast to these approaches, parsimony methods that can treat a gap as a fifth character - by an extension of the substitution models - directly infer a phylogeny. A disadvantage of this method is that it treats each gap as an independent event. Recently, a probabilistic analysis of indels was designed to incorporate a gap as fifth nucleotide within a Markov model of nucleotide substitution (McGuire et al., 2001). To overcome the disproportionate influence of multi-site gaps, the likelihood of these sites is down-weighted. Despite the efforts to implement models for indels in computational analyses, a well-supported and more realistic method for the coding of gaps is still lacking. Yet, it nevertheless is better to somehow incorporate gap
information inadequately, than to ignore it altogether.

We must be aware that indels, although being rarer than base substitutions, are also subject to homoplasy. It is possible that sequences around the indel site mutate so fast over time that DNA elements, coding for inserted or deleted amino acids, are no longer recognizable. Figures 1 and 2 reveal the relative levels of homoplasy for the various indels in TNFα. Variables like the length, types and surrounding sequences of an indel influence the chances that it may occur. It has been suggested that longer gaps are more reliable phylogenetic markers because they are unlikely to occur repeatedly at exactly the same position, and with the same length and sequence (Lloyd and Calder, 1991; Simmons et al., 2001). However, longer gaps in rDNA, ITS and intron sequences were found to have similar levels of homoplasy as shorter gaps (Simmons et al., 2001). Insertions are rarer than deletions and consequently should be modeled differently. Inserted nucleotides are also potentially reversible and deletions only if located in repeat regions. Other important factors influencing the chances of homoplasy are the degree of sequence variability around the indel, the presence of short DNA repeats, and the adjacency of splice junctions. Note that shifting of a splice junction can be caused by a single base substitution, which is a simple and frequent mutational event.

Taken these considerations into account, indel 1 of TNFα - a deletion of 1 amino acid caused by intron sliding - can easily originate. This increases the chances that homoplasy occurs, and consequently decreases the phylogenetic significance. Indel 3 of TNFα, being an insertion of 2 amino acids resulting from slipped-strand mispairing, should be more informative as a phylogenetic marker. However, assuming that the Glires clade is correct, it is peculiar that all rodents as well as the rabbit possess this insertion, while lacking in pika. Possibly, this insertion appeared in the ancestor of the Glires and is lost again in the pika or this insertion occurred independently in both the rabbit and the common ancestor of the rodents.

Gaps at the beginning and end of a sequence alignment are often artifacts of the sequencing procedure. Yet, the duplication of ATG at the beginning of the mouse IAPP gene is real. It is moreover a rare occurrence because ATG triplets around start codons are selected against in order to avoid disturbance of the accurate detection of proper start codons (Saito et al., 1999). In higher eukaryotes, the average distance between a start codon and its nearest upstream ATG is generally longer than in prokaryotes.

Although gaps can be prone to homoplasy, also the use of primary sequence data for tree reconstruction is obviously not without problems. Convergent evolution of nucleotides, varying substitution rates among sites and lineages, saturation of mutations at variable sites, non-independent substitutions among sites, and functional constraints at the molecular level are just a small sample of the potential caveats that apply when using this type of data. An advantage of indels is that they may allow a better discrimination between homoplasy and homology.

Altogether, both the sequences of TNFα and the indels provide independent evidence for the monophyly of rodents and the sister grouping of Lagomorpha with Rodentia. This is in contrast to an extensive study of the mitochondrial data set (Reyes et al., 2000) that showed data in agreement with both paraphyly and polyphly of rodents. However, another study using a similar data set and the same ProtML analysis with the mtREV-F model could not exclude rodent monophyly (Cao et al., 2000). The inclusion of more rodent species is probably needed to break the long rodent branches, which are caused by the very fast evolution of their mitochondrial genomes (Philippe and Douzery, 1994). Also, Lagomorpha were clustered with primates in a study based on molecular data (Graur et al., 1996). However, reanalysis of this study showed that it was seriously hampered by under-representation of taxa, which amplified the biases introduced by missing data and long branches (Halanych, 1998). Our present results, too, especially those for the IAPP data set, could become more reliable by adding more species from different eutherian orders. However, determining sequences of the IAPP gene is very
labor intensive and time-consuming. The taxon sampling for TNFα is more adequate, although adding species like elephant shrew and tree shrew could exclude their clustering with the rabbit as suggested in earlier studies (McKenna, 1975; Graur et al., 1996).

To overcome the backlog in statistical handling of indels as compared with base substitutions, it is necessary to obtain a better knowledge about the mechanisms that generate indels. This should provide better insight in their rates of origin, character independence, mutational biases, and reversibility. Nevertheless, even in the present qualitative manner, the phylogenetic signal contained in indels can offer an independent way of evaluating alternative phylogenetic hypotheses.

Acknowledgements

We thank Ole Madsen for assistance with phylogenetic tree analyses, and Wim Tieman for initial data base searches. We thank François Catzeflis for tissue samples. This work was supported by the Research Council for Earth and Life Sciences (ALW).
References

CHAPTER 4

Protein sequence signatures support the African clade of mammals
Abstract

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

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

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

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

Materials and methods

Searching for Afrotherian signatures

Databases were searched for sets of protein sequences that included representatives of at least four Afrotherian orders, i.e., Proboscidea (elephants), Sirenia (sea cows), Hyracoidea (hyraxes), Tubulidentata, Macroscelidea, and Afrotheria (golden moles and tenrecs; ref. 7). This yielded data sets of CRYAA, AQP2, IRBP, von Willebrand factor, α-2B adrenergic receptor, γ-fibrinogen, hemoglobin-α and -β, and cytochrome b. The AQP2 data set was complemented with newly determined sequences of pig, fin whale and sperm whale (see below). From these data sets, one or, if available, two representatives of all included eutherian orders were taken. When more than two species were available for an order, only the two most divergent sequences were retained. This increases the homoplastic background, and thus the significance of retrieved signatures. Retaining all sequences would make the taxon representation unbalanced and hamper the signature searches. The selected sequences were aligned, using PILEUP, and manually edited. Where available, two divergent Marsupialia were
Protein sequence signatures

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

Sequence determination of CRYAA and AQP2

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

Phylogenetic tree reconstruction

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

We used a two-step procedure to derive the maximum likelihood (ML or maximum average likelihood sensu Steel and Penny, ref. 25) phylogeny from our sequence data. The size of our phylogenies precluded an exhaustive search of all possible topologies to find the global ML tree. We therefore first calculated the likelihood of the sequence data on starter topologies obtained from a simple neighbor-joining (minimum evolution) analysis. Likelihood calculations were done using PAUP-ML (26) assuming the HKY85 model of evolution with gamma rate heterogeneity to allow for the possibility of unequal rates of evolution across sites. Likelihood calculations were done using PAUP-ML (26) assuming the HKY85 model of evolution with gamma rate heterogeneity to allow for the possibility of unequal rates of evolution across sites. We estimated the shape parameter (α) of the gamma distribution, and the transition/transversion ratio, from the data. This yielded a candidate topology with branch lengths based upon the ML distance calculation. Subsequently, we searched for better topologies in the region of the initial neighbor-joining topology by using the tree-bisection-reconnection branch swapping algorithm in PAUP-ML. We repeated this analysis procedure
with random sequence input orders and always found the same ML tree. ‘Constrained trees’ were constructed to conform with alternative morphology-based hypotheses for eutherian relationships (see Fig. 2). For these trees we supplied the topologies and reconstructed branch lengths by ML.

**Computation of ancestral states**

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

**Results**

**Candidate sequence signatures in CRYAA, AQP2 and IRBP**

In the alignment of 28 mammalian CRYAA sequences, six positions were found to be relevant for distinguishing any possible five-species clade (Fig. 1A). The only group of five species set apart by a combination of two or more putative apomorphies, namely 70Q, 74L and 142C, is formed by elephant, manatee, hyrax, aardvark and elephant shrew. The combination QLC at positions 70, 74 and 142 thus is a unique feature for the Afrotheria in this CRYAA data set. All Afrotheria, apart from African elephant, share in addition the apomorphy 72L. We therefore investigated the phylogenetic value of 70Q, 72L, 74L, 142C as a putative Afrotherian signature in CRYAA. Amongst 20 aligned AQP2 sequences, we traced four positions at which putative apomorphies might be diagnostic for a six-species clade (Fig. 1B).

**Figure 1.** Afrotherian signatures in CRYAA, AQP2 and IRBP. A) Positions from an alignment of 26 eutherian and 2 marsupial CRYAA sequences at which the same putatively apomorphous replacement occurs in 4-6 eutherian sequences. In black are replacements that occur in at least four of the five Afrotheria. Note that ‘.’ indicates all residues that did not pass the 4-6 search window; they may be identical to the two top out-group sequences or be apomorphies occurring in < 4 or > 6 sequences; x is unassigned residue; * denotes species that are included in the trees in Fig. 2. B) Apomorphous replacements occurring in 5-7 AQP2 sequences, considering armadillo as out-group for the other Eutheria; rodents as out-group yields the same signature (shown in Fig. 4). C) Positions in IRBP that pass the search for four- to six-species eutherian clades. Note that at certain positions different putatively apomorphous replacements fulfil the search criterium and may set apart different clades (e.g., 13E for Afrotheria, 13G for Cetartiodactyla). D) and E) Afrotherian signature positions in newly determined CRYAA and AQP2 sequences, respectively.
| A. Position nr: | 1       | 167774 | 1450 |
|               | 310242  | 0854   |
| kangaroo*     | AIKVEFS  | wombat* | opossum* |
| opossum*      | ..R.Y.   | Ind. elephant* | AAAI |
| tamandua      | ..Q.     | dugong* | 3-toed sloth* |
| Afr. elephant*| ..Q.     | aardvark* | hyrax* |
| manatee*      | ..Q.     | elephant shrew* | golden mole* |
| hyrax*        | ..Q.     | aardvark* | rat* |
| aardvark*     | ..Q.     | elephant shrew* | guinea pig* |
| elephant shrew*| ..Q.     | mouse* | mole rat* |
| beaver*       | ..Q.     | N. Am. porcupine* | rabbit* |
| pika          | ..Q.     | human* | rat* |
| rabbit*       | ..Q.     | mole* | elephant shrew* |
| galago*       | ..Q.     | aardvark* | golden mole* |
| human*        | ..Q.     | hyrax* | tree shrew* |
| tree shrew*   | ..Q.     | hedgehog* | hedgehog* |
| hedgehog*     | ..Q.     | fin whale* | fruit-eat. bat* |
| fruit-eat. bat*| ..Q.     | sperm whale | hominid |
| flying fox*   | ..Q.     | round-eared bat* | panther* |
| pangolin*     | ..Q.     | flying fox* | pig* |
| mink*         | ..Q.     | pangolin* | dog* |
| seal*         | ..Q.     | fox* | cat* |
| rhinoceros*   | ..Q.     | tapir* | tapir* |
| horse*        | ..Q.     | horse* | horse* |
| porpoise      | ..Q.     | minke whale* | minke whale* |
| minke whale*  | ..Q.     | hippo* | hippo* |
| bovine*       | ..Q.     | bovine* | bovine* |
| pig*          | ..Q.     | Q. ..Q.  | Q. ..Q. |
| D.            | Ind. elephant | dugong | Ind. elephant |
|               | tamandua | taill. tenrec | taill. tenrec |
|               | Afr. elephant* | Afr. elephant* | Afr. elephant* |
|               | manatee | golden mole* | golden mole* |
| E.            | manatee | tree hyrax | tree hyrax |
|               | sea otter | otter shrew | otter shrew |
|               | North Atlantic | North Atlantic | North Atlantic |
|               | Q. ..Q.  | Q. ..Q.  | Q. ..Q.  |

Protein sequence signatures

| B. | 111111111112344444555778481445777223455577889990011222 |
|    | 11233336889915123660493608200770256956336812892363633678 |

Protein signature sequences
The combination 10S, 55T and 104L perfectly sets apart the Afrotheria, with exception of dugong, which only shares the 10S apomorphy. The signature STL will be studied as an Afrotherian marker in AQP2. In the alignment of 28 IRBP sequences, 47 positions pass the search for a five-species grouping (Fig. 1C). There are 17 putative apomorphies in support of at least four of the five Afrotheria. The combination 18M, 19S, 76S, 147G, 226I, 272D and 328E is even perfectly unique for all five Afrotheria. At the ten other positions the signature is affected by homoplasy, within the limits allowed by our search procedure (see Materials and Methods). The ‘degenerate’ 17-residue signature is used in our further analyses. No signatures were detected in the other six proteins.

**CRYAA and AQP2 signatures in other Afrotheria**

To perform meaningful likelihood calculations for the retrieved signatures it was desirable to broaden the Afrotherian representation by sequencing CRYAA and IRBP in golden mole and tenrec, and AQP2 in tenrec. This would also be a test for the diagnostic value of these signatures; if they are genuine synapomorphies for Afrotheria, one expects to find them, completely or partially, in CRYAA, AQP2 and IRBP from other members of this clade. We sequenced exon 2 of the CRYAA gene, which encodes the signature positions 70, 72 and 74, in Indian elephant, dugong, golden mole and three Tenrecidae, including the otter shrew. All new CRYAA sequences were found to code for 70Q, 72L and 74L, including that of Indian elephant, suggesting that 72V in African elephant is a back mutation (Fig. 1D). For AQP2, additional sequences were obtained for manatee, tree hyrax and three tenrecs, again including otter shrew. All of these species have the STL signature, apart from manatee, which like dugong AQP2 misses 55T and 104L (Fig. 1E). Unfortunately, sequences for golden mole and tenrec IRBP could not be obtained.

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

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

A. ML tree

3-toed sloth
Paenungulata
Aardvark
Elephant shrew
Afrosoricida
Primates
Rodentia
Rabbit
Tree shrew
Hedgehog
Chiroptera
Pangolin
Cetartiodactyla
Perissodactyla
Carnivora

B. Afrosoricida with hedgehog

3-toed sloth
Hedgehog
Afrosoricida
Aardvark
Elephant shrew
Paenungulata
Pangolin
Cetartiodactyla
Chiroptera
Perissodactyla
Carnivora
Rodentia
Rabbit
Tree shrew
Primates

C. Elephant shrew with Glires

D. Novacek

3-toed sloth
Hedgehog
Pangolin
Cetartiodactyla
Perissodactyla
Carnivora
Chiroptera
Tree shrew
Primates
Elephant shrew
Rabbit
Rodentia
Aardvark
Afrosoricida
Paenungulata

Hedgehog
Afrosoricida
Elephant shrew
Rabbit
Rodentia
Chiroptera
Tree shrew
Primates
Aardvark
Paenungulata
Perissodactyla
Cetartiodactyla
Carnivora
3-toed sloth
Pangolin
**Likelihoods of the CRYAA, AQP2 and IRBP signatures**

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

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

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

<table>
<thead>
<tr>
<th>Protein</th>
<th>Topology</th>
<th>Log-likelihood&lt;sup&gt;*&lt;/sup&gt;</th>
<th>Log-difference from best tree&lt;sup&gt;*&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRYAA (n=27 species)</td>
<td>A</td>
<td>-90.07(-1540.09)</td>
<td>0.00(0.00)</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>-100.12(-1548.75)</td>
<td>10.05(8.66)</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>-103.65(-1563.69)</td>
<td>13.58(23.60)</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>-119.17(-1667.25)&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>29.1(127.16)&lt;sup&gt;‡&lt;/sup&gt;</td>
</tr>
<tr>
<td>AQP2 (n=20 species)</td>
<td>A</td>
<td>-51.95(-876.97)</td>
<td>0.00(0.00)</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>-61.83(-891.97)</td>
<td>9.88(15.00)</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>-61.52(-896.81)</td>
<td>9.57(19.84)</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>-98.78(-1004.13)&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>46.83(124.16)&lt;sup&gt;‡&lt;/sup&gt;</td>
</tr>
<tr>
<td>IRBP (n=28 species)</td>
<td>A</td>
<td>-541.55(-6118.73)</td>
<td>0.00(0.00)</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>-590.41(-6217.33)</td>
<td>48.86(98.27)</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>-735.74(-6419.23)&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>194.19(300.46)&lt;sup&gt;‡&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>*</sup>n; Number of sequences used for tree constructions.

<sup>‡</sup>Topologies as explained in legends of Fig. 2; tree B is lacking for IRBP, Afrosciricida not being available.

<sup>*</sup>Likelihoods in parentheses are those calculated for the entire protein sequence and agree in every case with those calculated for the signature sequence alone.

<sup>‡</sup>Likelihood of tree D adjusted to have same number of branches as other trees.

**Likelihoods of ancestral state reconstructions**

If the signatures in CRYAA, AQP2, and IRBP are synapomorphies of Afrotheria they should have evolved in the branch leading to the basal node of the Afrotherian clade. The estimated posterior probabilities of observing the signature QLCC at nodes 1 and 2 in Fig. 2A are 3.0x10<sup>9</sup> and 0.984, respectively. For the AQP2 and IRBP ML trees the corresponding probabilities are 2.0x10<sup>6</sup> and 0.987, and 7.8x10<sup>32</sup> and 0.391, respectively. The sequence signatures of all three proteins thus have a high probability of evolving in the branch leading to the basal node of the Afrotheria. Probabilities this high for the CRYAA and AQP2 signatures imply that each amino acid replacement in the signatures has a near 1.0 probability of having evolved in that branch. Even for the IRBP signature, which requires seventeen separate events in a specified branch, the combined probability is 0.391. Removing just two of the more
variable sites (e.g., 59S and 326E, each of which has an approximately 0.65 probability of having evolved in the branch), the combined probability rises to 0.94.

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

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

Phylogenetic value of the Afrotherian signatures

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

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

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

(1)
gives the probability of an \( r \)-event. Summing this product over \( r \), allowing \( r \) to range from \( r \) to \( V \), gives the probability of a signature of length \( r \) or greater. Call this probability \( p_s \), where ‘s’ denotes ‘signature’.

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

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

(2)
gives the probability of observing a signature of length \( r \) or greater in \( b \) branches of the tree.

Summing this product over \( b \) ranging from 1 to \( N \) gives the probability of observing on the tree at least one signature of length \( r \) or greater. Call this quantity \( p_t \), where the ‘t’ denotes the tree.
Table 2. Probabilities of signatures occurring in the ML trees

<table>
<thead>
<tr>
<th>Protein</th>
<th>Variable sites</th>
<th>Total changes on tree</th>
<th>$p^\Psi$</th>
<th>$p_s^*$</th>
<th>$p_t^*$</th>
<th>Probability $p$ for same signature of $r$ sites to occur</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Twice</td>
</tr>
<tr>
<td>CRYAA</td>
<td>57</td>
<td>123</td>
<td>0.014</td>
<td>0.008</td>
<td>0.328</td>
<td>2.8 x 10^{-7}</td>
</tr>
<tr>
<td>(173)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AQP2</td>
<td>31</td>
<td>78</td>
<td>0.018</td>
<td>0.019</td>
<td>0.523</td>
<td>6.76 x 10^{-5}</td>
</tr>
<tr>
<td>(111)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRBP</td>
<td>227</td>
<td>912</td>
<td>0.051</td>
<td>0.069</td>
<td>0.979</td>
<td>9.48 x 10^{-26}</td>
</tr>
<tr>
<td>(334)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Calculated by reconstructing the most parsimonious set of amino acid replacements on the ML tree.

$\Psi$ Probability of a substitution per site per branch. Calculated as total changes/length of protein/no. of branches in tree. Our conclusions are unaltered if we calculate $p$ using the number of variable sites rather than the length of the protein.

# Probability of a signature of length $r$; for CRYAA, AQP2, and IRBP, $r$ is 4, 3, and 17, respectively.

$\phi$ Probability of any $p_s$ event at least once on a tree (further defined in the text).

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

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

$$p_t \binom{N_{b-1}}{I} p_b^I (1 - p_b)^{(N_{b-1} - I)}$$

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

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

\[
P(\text{Afrotheria}) = \frac{w(\text{Afrotheria})P(\text{signature} \mid \text{Afrotheria})}{P(\text{signatures})}
\]

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

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

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

<table>
<thead>
<tr>
<th>Protein</th>
<th>Prior weight for Afrotherian clade</th>
<th>Posterior probability of Afrotheria</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRYAA</td>
<td>0.001</td>
<td>0.999</td>
</tr>
<tr>
<td>AQP2</td>
<td>0.001</td>
<td>0.921</td>
</tr>
<tr>
<td>IRBP</td>
<td>0.001</td>
<td>~1.00</td>
</tr>
<tr>
<td>Combined</td>
<td>0.0001</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Discussion

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

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

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

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

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

Acknowledgements

We thank Collin van Asten for writing SIGNWIN, Marcel Sweers for technical help, and Peter Vogel (Lausanne) for otter shrew tissue. Supported by grants from the Netherlands Foundation for Life Sciences (ALW) to W. de J., the European Commission (TMR) to W. de J., F.C. and M.S., and the Leverhulme Trust and the Natural Environment Research Council (UK) to M.P.
References

CHAPTER 5

αA-crystallin in avian phylogeny: the hoatzin groups with Cuculiformes
Abstract

The hoatzin (Opisthocomus hoazin), a South American bird, has been an issue in avian phylogeny ever since it was described in 1776 (1). Various classifications have grouped the hoatzin either with the fowl-like birds (Galliformes), with the cuckoos (Cuculiformes) or as a monotypic order, the Opisthocomiformes. In this paper we present comparative sequence data from the nuclear gene that encodes the eye lens protein αA-crystallin, and refer to the two mitochondrial genes for 12S and 16S rRNA, in relation with the phylogenetic position of the hoatzin. Both αA-crystallin and these two mitochondrial rRNA genes indicate that the hoatzin shares a more recent common ancestry with Cuculiformes than with Galliformes.
Chapter 5

Introduction

Most species of birds can be grouped into well-characterized monophyletic orders. However, the phylogenetic relationships of the hoatzin (Opisthocomus hoazin) have been uncertain. The hoatzin lives in the humid lowlands of northern and central South America, often in tropical riparian habitats (2). It feeds primarily on the tender young leaves, twigs and shoots of trees and marsh plants that are ingested into a huge, muscular crop. This crop is used for microbial foregut fermentation to convert cellulose into simple sugars, which is unique for the hoatzin and some groups of mammals (3). Normally the hoatzins form large non-breeding flocks, but with the first rains they break up into smaller groups and defend small territories. They are communal breeders and the clutch is usually two eggs.

Both morphological evidence and molecular data have been inconclusive as to their relationships. Morphologically, the most frequently held opinion is that the hoatzin groups closest to the Galliformes (4-7). However, DNA-DNA hybridization comparisons were interpreted to indicate a closer relationship with the cuckoos (8, 9), while the amino acid sequence of the bacteriolytic lysozyme of the hoatzin was more similar to that of rock pigeon (Columba livia) than to that of domestic fowl (Gallus gallus) (10). Sequence data of more than 900 bases of the mitochondrial gene for cytochrome b from 18 species of birds did not provide a clear answer either (11). Another hypothesis suggested the alliance of the hoatzin with turacos based on morphological (12, 13) and behavioral similarities (14). Turacos are a family of 20 fruit-eating species that live in the forests of sub-Saharan Africa. Traditionally they are grouped among the Cuculiformes. Yet, based on DNA-DNA hybridization analyses the turacos were not grouped with cuckoos (9). Because of these conflicting hypotheses on the phylogenetic position of the hoatzin, it appeared desirable to obtain additional sequences, including those from nuclear genes, for solving this question. In this paper we present new sequence data from the nuclear gene for αA-crystallin, and refer to the two mitochondrial genes for 12S and 16S rRNA (15).

αA-crystallin is a suitable protein for comparative sequence analyses because it is a single copy gene, which avoids the risk of comparing paralogous gene products of multi-copy genes in different species and the complicating effects of gene conversion. Furthermore, αA-crystallin has a relatively slow rate of evolution and is therefore informative in the study of vertebrate phylogenetic relationships at the level of higher taxonomic categories (16, 17). In birds, the 173-residue αA-crystallin protein sequences earlier supported that Ratites (ostrich and related flightless birds) and tinamous (pigeon-like South-American birds), form a monophyletic clade, the Palaeognathae, as the first offshoot of the avian stem (18, 19). All other birds, the Neognathae, were set apart from the palaeognaths on basis of two amino acid replacements in αA-crystallin: 122 S→A and 147 Q→P. Amongst the neognathous birds, the Galliformes and Anseriformes (duck-like birds) are the first to branch off, as an unresolved trichotomy. Three synapomorphic amino acid replacements, 127 S→A, 135 S→N and 152 P→A, distinguish the other investigated neognathous orders from the Galliformes and Anseriformes.

These five phylogenetically diagnostic replacements are all encoded in the third exon of the αA-crystallin gene. Therefore, although the third exon of αA-crystallin is only 207 bp in length, it should be informative to distinguish whether the hoatzin is closer to Galliformes or to Cuculiformes. We amplified and determined the relevant sequence of the αA-crystallin gene in hoatzin and in representatives of Cuculiformes and Galliformes, as well as from a sampling of additional neognathous orders (Anseriformes, Charadriiformes, Passeriformes and Columbiformes). Unfortunately, no turaco sequences could be obtained. Some further amniote αA-crystallin DNA sequences were obtained from the data bases. Phylogenetic analyses were performed on this data set, separately and in combination with sequence data from the two mitochondrial genes for 12S and 16S rRNA (15), to help resolve the position of the hoatzin within the Neognathae.
Materials and methods

Full species names and EMBL accession numbers for the αA-crystallin sequences aligned in Fig. 1 are: hoatzin (U31947), groove-billed ani (Crotophaga sulcirostris; U31940), silver pheasant (Lophura nycthemera; U31943), sooty tern (Sterna fuscata; U31944), and a thrush, the Eurasian blackbird (Turdus merula; U31942). Corresponding fragments of αA-crystallin cDNA sequences were obtained from rock pigeon (U31945) total lens RNA according to the method described by Caspers et al. (19), and from a mallard duck (Anas platyrhynchos; U31946) lens cDNA library in phage λgt11 (20). Previous studies also provided corresponding sequences from elegant-crested tinamou (Eudromia elegans; L25850) and red-eared slider turtle (Trachemys scripta elegans; U31938) (19, 21). DNA sequences for αA-crystallin from domestic fowl (M17627), house mouse (Mus musculus; J00376), and human (Homo sapiens; U05569) were obtained from the database for comparison.

Most of the third exon of the αA-crystallin gene was amplified by the polymerase chain reaction (PCR) using a Biometra TRIO-thermoblock. Degenerated primers were designed, using the program OLIGO 4.0, to amplify a 146-bp sequence coding for amino acids 112-159 of the αA-crystallin chain. The sequences of these primers, synthesized by Eurogentec S.A., were 5'-GAY GAC CAY GGC TAC ATN TC-3' and 5'-TTY TCC TCC YGN GAC ACN G-3'. A total reaction volume of 50 µl was used. The dNTPs were from Boehringer; the Taq polymerase was a gift from Dr. Wiljan Hendriks and the buffer contained 2.0 mM MgCl2, 20 mM Tris-HCl, 50 mM KCl, 0.27% Tween-20, 0.27% P-40 and 0.2 mg/ml bovine serum albumin. As template we used total DNA isolated from the liver of a Eurasian blackbird, and nuclear DNA of hoatzin, groove-billed ani, silver pheasant and sooty tern, provided by Drs Charles G. Sibley and John C. Avise. To each reaction 200 ng template was added. Optimalisation of the PCR conditions led to the following program: initial denaturation for 10 min at 95°C; 30 cycli of 2 min at 95°C for denaturation, 1 min at 54°C for annealing, and 30 sec at 72°C for extension; and finally an extra extension step for 7 min at 72°C. To obtain a specific reaction, 5 µl PCR product of the hoatzin, the groove-billed ani and the silver pheasant were reamplified using the same program. The PCR products were directly ligated into a T-vector using a TA cloning kit (Invitrogen or Promega). All sequences were amplified in triplicate and sequenced in both directions using the Sequenase version 2.0 DNA sequencing kit (United States Biochemical).

Sequences were aligned with the program Pileup from the GCG package (22). Phylogenetic analyses were performed with neighbor-joining (23), MEGA (24), maximum parsimony (DNAPARS in PHYLIP; 25), and maximum likelihood (PHYLIP; 25). The neighbor-joining method employed Kimura distances (24, 26) using transversions only. Statistical confidence in the neighbor-joining analyses was assessed by the interior branch test (27) in PHYLTEST (28).

Results and discussion

A closer relationship of the hoatzin to either Galliformes or Cuculiformes was examined on basis of a 146-bp sequence from exon 3 of the αA-crystallin gene. This small DNA fragment codes for residues 112-159 of the protein and contains the information for five amino acid replacements that are pertinent for resolving some major avian groupings, as discussed above (18, 19). Sequencing the complete 3500-bp αA-crystallin gene was considered as too laborious relative to the expected additional phylogenetic information; the three protein-coding exonic regions are interrupted by two large and phylogenetically uninformative introns. Also, it was not possible to obtain αA-crystallin cDNA sequences from lens mRNA, because fresh lenses of hoatzin and most other birds are obviously not available. Therefore, the 146-bp sequence of exon 3, coding for residues 112-159 of αA-crystallin, was determined from nuclear DNA of the hoatzin.
Figure 1. Comparison of hoatzin αA-crystallin sequences with those of other amniotes. a) Alignment of the variable positions in a 146-bp sequence from exon 3 of the αA-crystallin gene. (.) indicates identity with the top sequence. Full species names and EMBL accession numbers are given in Materials and methods. b) Alignment of the amino acid sequences deduced from the 146-bp DNA sequences in Fig. 1a. Positions are numbered as in the complete αA-crystallin protein.

and some relevant other birds. The phylogenetically informative positions in the obtained sequences are aligned in Fig. 1a, together with the corresponding nucleotides in the only other available avian αA-crystallin DNA sequences, those from domestic fowl and tinamou, and with the most relevant amniote outgroups (i.e., turtle, human and mouse). The corresponding deduced amino acid sequences are aligned in Fig. 1b. The latter alignment confirms that the replacements 122 S→A and 147 Q→P apparently have originated in the lineage to the neognathous birds, after the divergence of the palaeognaths, as represented by tinamou. Moreover, after the divergence of the Galliformes and Anseriformes, the remaining neognaths, including the hoatzin, share the amino acid replacements 127 S→A and 152 P→A (with the exception of 152-T in ani). This already indicates that hoatzin groups closer to the other neognaths than to the galliform-anseriform clade. The fact that the hoatzin shares 135-S with galliforms and paleognaths would suggest that the hoatzin is the first to split off from the remaining neognaths. However, the hoatzin and the groove-billed ani share the unique replacement 155 S→V, which requires two base substitutions. The weight of evidence of these various amino acid replacements thus favors a closer affinity of the hoatzin to the Cuculiformes than to the Galliformes.
Figure 2. Phylogenetic relationships of the hoatzin as inferred from the αA-crystallin DNA sequences as presented in Fig. 1a. a) Neighbor-joining tree constructed from transversion distances of the complete data set. b) Similarly constructed neighbor-joining tree for the four taxa hoatzin, ani, Galliformes and turtle as outgroup. Confidence probability values ($P_c$) are indicated at the nodes. Distances are proportional to the minimum number of mutations per residue.

Because of the scarcity of amino acid replacements, and the consequent paucity of phylogenetic information, the various tree reconstructions were performed on the nucleotide sequences only. The neighbor-joining tree, using the Kimura method with transversions only, places the hoatzin as the closest relative of the Cuculiformes (Fig. 2a; $P_c = 79\%$). When only hoatzin, ani and galliforms are compared, with turtle as outgroup, the support for the hoatzin-cuckoo grouping increases to 91% (Fig. 2b). Branch lengths are highly unequal in both trees; this is to be expected because of the small numbers of sites involved in the analyses, which is even further reduced by considering only transversions. However, the more frequently occurring transitions are excluded in order to diminish the effect of superimposed substitutions. The parsimony consensus tree (not shown) yielded exactly the same topology as in Fig. 2a. Also in the maximum likelihood tree the hoatzin and ani were grouped together (not shown). All the different analyses showed the hoatzin-ani clade as well as a sister group relation between Anseriformes and Galliformes.
Figure 3. Phylogenetic tree of hoatzin, ani, domestic fowl and a chimeric outgroup inferred from concatenated DNA sequences of the $\alpha$A-crystallin gene third exon and the mitochondrial 12S-16S rRNA and cytochrome b genes (2832 bp in total). The tree was reconstructed by the neighbor-joining method with transversion distances. Distances are proportional to the minimum number of mutations per residue.

The relationship of the hoatzin was further analyzed on basis of the combined data set of the nuclear $\alpha$A-crystallin gene and the mitochondrial 12S-16S rRNA and the earlier studied cytochrome b sequences. In this case the phylogenetic problem was treated as a four-taxon statement, using only the sequences of hoatzin, ani, domestic fowl and an outgroup. The concatenated outgroup sequence consisted of the combination of turtle $\alpha$A-crystallin, alligator 12S-16S rRNA and human cytochrome b. This resulted in an even higher support for the hoatzin-cuckoo relationship, with confidence probability ($P_c$) and bootstrap values of 100% (Fig 3 and Table 1). Also separate four-taxon analyses of the $\alpha$A-crystallin, 12S and 16S rRNA, and cytochrome b data sets provided statistically significant support for the hoatzin-cuckoo relationship, giving bootstrap values from 96 to 99%, and $P_c$ values from 91 to 100% (Table 1).

Within birds, the evolutionary affinities of the hoatzin form one of the most difficult problems due to its morphological divergence (29). The closer affinity of the hoatzin to the cuckoos than to the galliforms, as concluded from the present analyses, is in contrast with most morphological studies. This relationship is, however, in agreement with the earlier DNA-DNA hybridization evidence (9).

### Table 1. Closest relative of the hoatzin in four-taxon analyses, as inferred from DNA sequences for 12S and 16S rRNA, cytochrome b and $\alpha$A-crystallin.

<table>
<thead>
<tr>
<th></th>
<th>no. of sites</th>
<th>closest relative (support in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>total</td>
<td>Variable</td>
</tr>
<tr>
<td>12S and 16S rRNA</td>
<td>1725</td>
<td>651</td>
</tr>
<tr>
<td>cytochrome b</td>
<td>961</td>
<td>435</td>
</tr>
<tr>
<td>$\alpha$A-crystallin</td>
<td>146</td>
<td>42</td>
</tr>
<tr>
<td>Combined</td>
<td>2832</td>
<td>1128</td>
</tr>
</tbody>
</table>

Bootstrap $P$ values (2000 replications) and $P_c$ values (in parentheses) in support of a hoatzin-galliform or a hoatzin-cuculiform relationship are presented. These values are obtained by neighbor joining analyses using transversion distances and four taxa: hoatzin, ani, domestic fowl and an outgroup (alligator for 12S-16S rRNA; human for cytochrome b and turtle for $\alpha$A-crystallin).
Recently, after completion of our sequence studies, six mitochondrial genes and one nuclear gene, totaling over 5.4 kb of aligned sequences, were analyzed to resolve the phylogenetic position of the hoatzin (30). This resulted in support for a close relationship between hoatzin and turacos, as a sister clade to cuckoos. Since turacos were not included in our data set, it might be of interest to determine the sequence of their $\alpha$A-crystallin third exon. The presence of nine nucleotide differences in this region between hoatzin and ani, might well allow to confirm this more precise positioning of the hoatzin in this clade. Considering our present results, we may conclude that even 146 bp of the $\alpha$A-crystallin gene are able to provide meaningful phylogenetic information.

Acknowledgements
This work was carried out under the auspices of the Netherlands Foundation for the Life Sciences (SLW) with financial aid from the Netherlands Organization for Scientific Research (NWO). Dennis Uit de Weerd and Freek Bouwman determined the pigeon $\alpha$A-crystallin sequence. Nuclear DNAs were provided by Drs Charles G. Sibley and John C. Avise, and Taq polymerase by Dr. Wiljan Hendriks.
References

CHAPTER 6

The evolution of an alternatively spliced exon in the αA-crystallin gene
Abstract

The evolutionary aspects of alternative splicing, as a mechanism to increase the diversity of gene products, are poorly understood. Here we analyse the evolution of a 69-bp exon that is alternatively spliced in the primary transcript of the gene for the mammalian eye lens protein αA-crystallin. In rodents, the skipping of this exon 2 is attributed to the presence of a non-consensus 5’ splice site GC, and results in the expression of 10-20% of αA\textsuperscript{ins}-crystallin, with an insert of 23 residues as compared with normal αA-crystallin. αA\textsuperscript{ins}-crystallin is also expressed in some non-rodent mammals, including kangaroo, while lacking in others. We now demonstrate that the alternatively spliced exon 2 is present in mammals from different orders that do not express αA\textsuperscript{ins}-crystallin. The expression of this exon has thus been silenced independently in various lineages. Sequence comparison in 16 species reveals that - whether or not αA\textsuperscript{ins}-crystallin is expressed - exon 2 is always flanked by the non-consensus donor splice site GC, while a consensus branch point sequence and 3’ pyrimidine-rich region are hardly detectable in the downstream intron. Increased numbers of amino acid replacements in the peptide encoded by exon 2 indicate that it is subject to much lower selective constraints than the exons that code for normal αA-crystallin. The absence of any apparent advantage at the protein level may suggest that exon 2 DNA sequences are conserved as \textit{cis}-acting factors for proper splicing of the αA-crystallin transcript.
More than one third of human genes may be alternatively spliced (Hanke et al. 1999; Croft et al. 2000). It thus is a major mechanism for enhancing genetic complexity and protein diversity. Alternative splicing is regulated by highly complex arrays of cis- and trans-acting factors that are often gene-, species-, cell- and development-specific (Kramer 1996; Lopez 1998). The evolutionary aspects and functional advantages of the various forms of alternative splicing - exon skipping, intron inclusion and the use of cryptic splice sites - and the development of species-specific splicing patterns are poorly understood (e.g., Conboy et al. 1992; Ito et al. 1994; Mollat et al. 1994; Reddy and Swarup 1995; Pret and Fiszman 1996; Laverdière et al. 2000; D'Souza and Schellenberg 2000). Here we try to reconstruct and understand the evolutionary fate and consequences of one of the first recognized cases of exon skipping, αA<sup>ins</sup>-crystallin (King and Piatigorsky 1983).

αA<sup>ins</sup>-crystallin is the alternatively spliced form of the eye lens protein αA-crystallin, a member of the small heat shock protein family which displays chaperone-like activity by suppressing the aggregation of unfolding proteins (Horwitz 1992). This property may contribute to maintaining the transparency of the lens. As compared to the 173-residue αA-crystallin, αA<sup>ins</sup>-crystallin contains an insertion of 23 amino acid residues between positions 63 and 64. This insert is encoded by exon 2, which as a ‘cassette exon’ (Smith et al. 1989) is normally spliced out, but spliced into 10-20% of the mature mRNA in rodent lenses (King and Piatigorsky 1983). The skipping of exon 2 is due primarily to mutation of its 5’ donor splice site, changing the consensus GT into GC (King and Piatigorsky 1983; Smulders et al. 1998).

The expression of low levels of αA<sup>ins</sup>-crystallin has also been observed in eye lenses of several non-rodent mammals, and even in kangaroo, but not in other placental mammals (Wistow 1995; see Fig. 1A). Since αA<sup>ins</sup>-crystallin is not observed in non-mammalian vertebrates (Hendriks et al. 1988), the alternatively spliced exon 2 probably originated in the mammalian lineage before the divergence of marsupials and placental mammals, and has vanished independently in different eutherian lineages. The physiological relevance, if any, of the expression of αA<sup>ins</sup>-crystallin is unclear. It is readily incorporated in the large and soluble α-crystallin complexes in the lens, but has considerably reduced chaperone-like capacity (Smulders et al. 1995).

The presence of exon 2 in eutherian αA-crystallin genes

To reconstruct the evolutionary rise and fall of αA<sup>ins</sup>-crystallin, we first established to what extent the known expression of αA<sup>ins</sup>-crystallin in lenses of different mammals correlates with the presence of genomic exon 2 sequences. Southern blotting was performed with an exon 2-specific probe on an appropriate PCR-amplified fragment (see legends of Fig. 1) from the αA-crystallin gene of 22 eutherian species, selected to represent the 18 generally accepted eutherian orders. Kangaroo was included to represent the nearest outgroup to the placental mammals. Fig. 1A indicates in which of these species the expression of αA<sup>ins</sup>-crystallin had earlier been established.

The Southern blot in Fig. 1B shows that exon 2-specific sequences are recognized in all included species, apart from golden mole and hedgehog. The absence of a positive signal in hedgehog DNA must be due to inadequate annealing of the probe since sequence determination (see below) confirmed that exon 2 is present in hedgehog, but indeed not in golden mole. Exon 2 is also detected in species like elephant, human, seal, bovine and whale, which have been found not to express αA<sup>ins</sup>-crystallin.

Fig. 1C presents the lengths of the amplified fragments, which span the αA-crystallin gene sequence from the 5’ end of exon 1 to the 3’ end of exon 3. The small fragment length of golden mole (874 bp) suggests that the absence of exon 2 is the result of deletion. The lengths of the other fragments ranges between 1400 bp in elephant shrew and 2250 bp in guinea pig. Since intron 1 is always small, between 63 bp in hedgehog and 322 bp in aardvark (as
determined by direct sequencing; see below), most of the length variation is accounted for by intron 2, of which the length is calculated to range between 1180 bp in tenrec and 1995 bp in guinea pig.

Figure 1. Presence of the alternatively spliced exon 2 in kangaroo and representatives of all eutherian orders. Asterisks indicate species for which exon 2 sequences were determined (cf. Fig. 2). (A) Known expression of αAins-crystallin, as determined by immunoblotting of eye lens extracts (Hendriks et al. 1988; Jaworski et al. 1990; Wistow 1995; aardvark and tenrec, M.S., unpublished data): +, present; -, absent; nd, not determined; *, determined in other species than used in present study, but from same order. (B) Southern blot of PCR fragments, spanning the region from exon 1 to exon 3, amplified from genomic DNA and hybridized with an end-labeled probe for exon 2. Positions of 1444-, 849- and 476-bp markers (top to bottom) are indicated by arrows. The considerable variation in intensity of exon 2 bands is due to differences in probe binding, and to the varying conditions used to obtain adequate results for all species in three separate analyses. Methods: PCR primers were designed to amplify the DNA sequence that codes for amino acid residues 13-117 of αA-crystallin, i.e. from the 5‘ end of exon 1 to the 3‘ end of exon 3, and thus including the optional exon 2, if present, and its flanking introns 1 and 2. The forward primer had the sequence 5’ CCA-TTC-AGC-ACC-CTT-GGT-TYA-ARC-G 3’ and the reverse primer 5’ TCT-CGT-TGT-GCT-TGC-CRT-GDA-TYT-C 3’. The PCR protocol as described in van Dijk et al. (2001) was followed. For Southern blot analysis, 50 ng of each PCR product, purified with the Qiaquick PCR Purification Kit (Qiagen), was blotted onto nitrocellulose membrane, and hybridized with a radioactive end-labeled probe (5’ TGG-ACC-AAG-CATG-CCTG-CAA-ACC-CAA-AGA-ACA-ACC 3’, designed on basis of exon 2 sequences of human, rodents and kangaroo). (C) Estimated size, in bp, of the PCR fragments. Methods: The lengths of aardvark, golden mole and human fragments (bold) were directly derived from their completely determined sequences. Together with the three marker bands (panel B), this provided six length values that were plotted against the mobilities on the ethidium bromide stained gel, as measured in mm relative to the fastest band (476 bp). An exponential curve was drawn through these six points, and the equation for this
curve, \( y = 507.08 \exp (0.2253 \, x) \), was calculated. For the other PCR products the length \( y \) in base pairs was then calculated by taking for \( x \) the mobility of the PCR fragment in mm relative to the 476 bp marker band. Full species names are: red kangaroo (Macropus rufus, Marsupialia), three-toed sloth (Bradypus tridactylus, Xenarthra), Indian elephant (Elephas maximus, Proboscidea), dugong (Dugong dugon, Sirenia), rock hyrax (Procavia capensis, Hyracoidea), aardvark (Orycteropus afer, Tubulidentata), elephant shrew (Macroscelides proboscideus, Macroscelidea), golden mole (Amblysomus hottentotus, Chrysochloridae), tail-less tenrec (Tenrec ecaudatus, Tenrecidae), hedgehog (Erinaceus europaeus, Erinaceidae), rat (Rattus norvegicus, Rodentia), guinea pig (Cavia porcellus, Rodentia), pika (Ochotona princeps, Lagomorpha), tree shrew (Tupaia tana, Scandentia), human (Homo sapiens, Primates), flying lemur (Cynocephalus variegatus, Dermoptera), big-eared bat (Macrotus californicus, Chiroptera), Blanford’s fruit bat (Cynopterus sphinx, Chiroptera), pangolin (Manis sp., Pholidota), grey seal (Halichoerus grypus, Carnivora), horse (Equus caballus, Perissodactyla), cow (Bos taurus, Artiodactyla) and fin whale (Balaenoptera physalus, Cetacea).

Sequencing of exon 2 and adjacent introns

To find out how exon 2 has evolved, we determined from the amplified fragments the sequences of exon 2, the 3’-end of exon 1, and the 5’-end of exon 3, as well as the adjacent regions of introns 1 and 2 in twelve species. These included the nine species marked with an asterisk in Fig. 1 (seven of which known to express \( \alpha A_{\text{ins}} \)-crystallin), and in addition, the orangutan and the loris (to pursue the silencing of \( \alpha A_{\text{ins}} \)-crystallin in human; Jaworski and Piatigorsky 1989), as well as Madagascar hedgehog, another tenrec species. Relevant parts of the obtained sequences are aligned in Fig. 2 with the known sequences of the kangaroo, human, mouse, mole rat, and hamster. With the exception of the golden mole, sequences resembling exon 2 can readily be distinguished in all newly determined sequences, and all of them have the same non-consensus GC donor splice site for exon 2 (Fig. 2, left black column in lower panel). Apart from the T \( \rightarrow \) C substitution, this 5’ splice site corresponds with the consensus sequence GTRAGT, with the exception of human and orangutan, which have GCATGC. The exon 1 donor splice sequence (Fig. 2, left black column in upper panel) deviates in all species from the consensus GTRAGT by a mutation of the 3’ T to A or G, with the exception of golden mole which has a perfect consensus sequence, possibly due to duplication of a GTAA sequence.

Figure 2. Sequence alignment of the optional exon 2, the nearest ends of exons 1 and 3, and the flanking intronic regions in the \( \alpha A \)-crystallin genes. *, species with newly determined sequences. *, nucleotides conserved in all sequences; ▲, deletion in human and orangutan exon 2; - - -, gaps introduced to optimize the alignment; --- ---, intronic regions omitted from the alignment. The 5’ and 3’ splice site dinucleotides are in black. Pyrimidines in the 3’ intronic regions are shaded dark grey; adenines and cytosines in the exons are shaded light grey. Arrowheads indicate the possible branch point A. Methods: \( \alpha A \)-crystallin genomic sequences were determined from the PCR-amplified fragments of aardvark (acc. nr. AJ272231 for exons 1 and 2, AJ299582 for exon 3), Indian elephant (AJ272232, AJ299583), tail-less tenrec (AJ272233, AJ299584), small Madagascar hedgehog (Echinops telfairi, Tenrecidae) (AJ272234, AJ299585), golden mole (AJ299591), hedgehog (AJ272235, AJ299586), guinea pig (AJ272236, AJ299587), pika (AJ272237, AJ299588), slow loris (Nycticebus coucang, Primates) (AJ272238), orangutan (Pongo pygmaeus, Primates) (AJ272239), Blanford’s fruit bat (AJ272240, AJ299589) and big-eared bat (AJ272241, AJ299590), following the protocol in van Dijk et al. (2001). The \( \alpha A \)-crystallin DNA sequences for human (U05569), golden hamster (Mesocricetus auratus, X02951), mole rat (Spalax ehrenbergi, M17249), and mouse (Mus musculus, J00376) were obtained from the database, and the exon 2 sequence of the grey kangaroo (Macropus fuliginosus, Marsupialia) was kindly provided by Drs. Cynthia Jaworski and Graeme Wistow. Alignment was performed with ClustalW (Higgins et al. 1996), and can be retrieved at the EBI WWW server (URL ftp://ftp.ebi.ac.uk/pub/databases/embl/align_000089).
The evolution of an alternatively spliced exon

**Exon 1**

- **Kangaroo**
- **Aardvark**
- **Ind. Elephant**
- **Tenrec**
- **Mad. hedgehog**
- **Golden mole**
- **Hedgehog**
- **Guinea pig**
- **Mouse**
- **Mole rat**
- **Hamster**
- **Pika**
- **Koris**
- **Orang-utan**
- **Human**
- **Fruit bat**
- **Big-eared bat**

**Exon 2**

**Exon 3**

**Intron 1**

**Intron 2**

**Diagram Note:**

- The diagram shows the evolution of an alternatively spliced exon across different species.
- Each species is represented by a different color, with the exon in black and the introns in red.
- The evolutionary relationships are depicted through the alignment of exonic and intronic sequences.
As for the 3’ splice site consensus Y(n)AG, a pyrimidine-rich region can readily be detected in intron 1, but hardly so in intron 2 (dark grey shading in Fig. 2). The consensus sequence around the mammalian branch point A is TNCTRACY, preferably with a G preceding the A (Mount et al. 1992). Although conserved A residues are present at the appropriate positions in introns 1 and 2 (18 to 40 nucleotides upstream from the AG acceptor; arrowheads in Fig. 2), consensus branch point sequences can hardly be recognized, least so in intron 2. At the 3’ end of all exon 2 sequences a potential splicing enhancing AC-rich region is reasonably well conserved (light grey shading in Fig. 2). The 1-bp deletion in this region of human exon 2 (Jaworski and Piatigorsky 1989) is also present in orangutan, but not in the loris sequence (black triangle in Fig. 2).

Rate of change of the insert peptide

The normal αA-crystallin is a moderately slowly evolving protein (3.0% amino acid sequence change in 100 million years; Caspers et al. 1995). One may wonder whether or not the inserted 23 residue sequence in αA<sub>ins</sub>-crystallin is subject to similar evolutionary constraints. Fig. 3 compares the deduced amino acid sequences for all known eutherian exon 2 DNA sequences with that of the outgroup marsupial. It reveals that the residues encoded by exon 2 are much more variable than the flanking regions encoded by exons 1 and 3. This becomes clearer when the 23 positions encoded by exon 2 and the 173 positions encoded by the constitutively expressed exons 1, 3 and 4 are pairwise compared for marsupial and eutherians (Table 1).

<table>
<thead>
<tr>
<th>EXON 1</th>
<th>EXON 2</th>
<th>EXON 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>kangaroo + RMQLFQTVLESGISE LMTHVFEMHKPHAGNPKNPTK VRSDRDFVIFLDVK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>aardvark +............D.............M...I...Q.............A.............Q...L...L...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ind. elephant -............D.............M...V...Q.............S...A.............Q...L...L...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tenrec +............D.............S...M...V...Q.............A.............Q...L...L...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mad. hedgehog ............D.............S...M...A...R...Q.............A.............Q...L...L...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hedgehog +............D.............A...M...L...V...Q...Q...G......A...Q.............</td>
<td></td>
<td></td>
</tr>
<tr>
<td>guinea pig +............D.............M...V...Q.............A.............</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mouse +............D.............M...V...Q.............V.............</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rat +............D.............M...V...Q.............G.............</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mole rat +............D.............R...V...Q.............E...I.............</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hamster +............D.............M...V...Q.............I.............</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pika +............D.............R...M...S...V...R...R...G...S...G.............</td>
<td></td>
<td></td>
</tr>
<tr>
<td>loris +............D.............R...M...V...N...Q...S...G.............</td>
<td></td>
<td></td>
</tr>
<tr>
<td>orang-utan ............D.............(.....A...C...V...R...Q...T...SS...SR)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>human −............D.............(.....C...V...R...Q...T...SS...SR) .............</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fruit bat +............D.............S...M...L...V...N...Q...S.............</td>
<td></td>
<td></td>
</tr>
<tr>
<td>big-eared bat +............D.............M...V...N...Q...P...S...S.............</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

** Figure 3. **Amino acid sequences encoded by exon 2 and adjacent regions of exons 1 and 3. Only the differences with the top sequence are shown. The presence (+) or absence (-) of αA<sub>ins</sub>-crystallin expression, if known, is indicated after the species names. Strictly conserved residues (*) and chemically similar replacements (.) are indicated below the alignment. Exon 2 in human and orangutan (in brackets) can not be expressed because of a 1-bp deletion in the codon for residue 85; exon 3 sequences for orangutan have not been determined. For kangaroo, exon 2 is from Macropus fuliginosus, and exon 1 and 3 from M. rufus. Loris exon 2 is combined with exon 1 and 3 sequences of the closely related galago. Rat is included in this figure, while not in Fig. 2, since only its αA<sub>ins</sub>-crystallin cDNA sequence is known [can be retrieved at the EBI WWW server (URLftp://ftp.ebi.ac.uk/pub/databases/embl/align_000089)].
The evolution of an alternatively spliced exon

In all comparisons the rate of amino acid replacement is faster in the insert peptide than in the remainder of the αA-crystallin sequence, the ratio’s ranging from 1.5 to 5.5. Because of the small number of residues in the insert peptide, the significance of this broad range is difficult to interpret. However, it is obvious that the insert peptide is under relaxed constraints as compared to normal αA-crystallin.

Evolutionary implications

In addition to the faster rate of change of the insert peptide, the lack of selective advantage is also obvious from the fact that expression of αA\textsuperscript{ins}-crystallin has disappeared in many mammals after it came into existence in the early mammalian lineage. It is indeed difficult to envisage any physiological advantage for the expression of a small amount of αA\textsuperscript{ins}-crystallin next to the normal αA-crystallin, especially since the insert peptide considerably reduces chaperone-like activity (Smulders et al. 1995). Yet, if αA\textsuperscript{ins}-crystallin as an expressed protein were an evolutionarily neutral, let alone slightly deleterious character, one would expect it to have vanished since long in all species, and all traces of the exon 2 DNA sequence to be obliterated. The fact that intact exon 2 sequences are present in all investigated mammals, apart from golden mole, while the flanking introns are difficult to align and riddled with gaps, suggests that selective constraints may have been retained at the DNA level. This also appears from the observation that human exon 2, since its silencing some 30 to 40 million years ago (Jaworski and Piatigorsky 1989), has not changed faster than in mammals like pika and hedgehog that still express αA\textsuperscript{ins}-crystallin (Table 1).

It is tempting to relate these constraints to cis-acting sequence elements required for proper splicing of the αA-crystallin gene transcript. Inspection of the cis-acting elements adjacent to the constitutively spliced exons 1 and 3 (Fig. 2) reveals in intron 2 a weak branch point, lacking the preceding G, and the absence of a downstream pyrimidine tract.

Table 1. Accelerated evolution of the insert peptide in αA\textsuperscript{ins}-crystallin

<table>
<thead>
<tr>
<th></th>
<th>exon 2</th>
<th>αA-crystallin</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aardvark</td>
<td>19.8</td>
<td>13.3</td>
<td>1.5</td>
</tr>
<tr>
<td>Guinea pig</td>
<td>19.8</td>
<td>10.6</td>
<td>1.9</td>
</tr>
<tr>
<td>Mouse</td>
<td>19.8</td>
<td>9.9</td>
<td>2.0</td>
</tr>
<tr>
<td>Rat</td>
<td>19.8</td>
<td>9.9</td>
<td>2.0</td>
</tr>
<tr>
<td>Mole rat</td>
<td>32.1</td>
<td>16.1</td>
<td>2.0</td>
</tr>
<tr>
<td>Hamster</td>
<td>19.8</td>
<td>9.9</td>
<td>2.0</td>
</tr>
<tr>
<td>Pika</td>
<td>54.8</td>
<td>9.9</td>
<td>5.5</td>
</tr>
<tr>
<td>Human</td>
<td>54.8</td>
<td>11.2</td>
<td>4.9</td>
</tr>
<tr>
<td>Hedgehog</td>
<td>64.0</td>
<td>13.3</td>
<td>4.8</td>
</tr>
<tr>
<td>Bat</td>
<td>46.5</td>
<td>10.6</td>
<td>4.4</td>
</tr>
</tbody>
</table>

* Corrected numbers of amino acid replacements per 100 residues are shown for the insert peptide encoded by exon 2 (first column), and for the constitutively expressed αA-crystallin (second column) in pairwise comparisons of ten eutherian species with the outgroup kangaroo. The ratio (third column) indicates that the insert peptide has changed from 1.5 to 5.5 times faster than normal αA-crystallin. For all species for which the αA\textsuperscript{ins}-crystallin amino acid sequence is available, the insert and normal αA-crystallin sequences were aligned separately (PILEUP, GCG package; Devereux et al. 1984). For both alignments pairwise comparisons were made (DISTANCES, GCG package) to estimate the number of replacements per 100 amino acids, corrected for multiple substitutions with the Kimura Protein Distance method (Kimura 1980). A comparison of substitutions per synonymous and nonsynonymous site in exon 2 and in the constitutively expressed exons might have provided further information, but was not possible since complete DNA sequences are not available for most eutherian species, nor for kangaroo.
Also the 5’ splice site in intron 1 is suboptimal, deviating from the consensus GTRAGT in all species, but for golden mole. These features are expected to adversely affect the splicing of the constitutive αA-crystallin mRNA. Considering that the production of αA-crystallin must proceed very efficiently in the lens, where it constitutes up to 25% of the total protein, we may speculate that compensating cis-acting elements must be present. These elements might be located in and around exon 2. Notably the potential AC-rich splicing enhancer in exon 2 might be relevant. The reason why golden mole can afford the loss of a recognizable exon 2 might be that it has a perfect 5’ splice site in intron 1; moreover, having degenerated underskin eyes, the golden mole can probably do with a less efficient production of the lens protein αA-crystallin.

Considering the poorer chaperoning capacities of αA\textsuperscript{ins}-crystallin, it certainly would have been disadvantageous if the cassette exon 2 had evolved into a constitutively spliced exon. The alternative splicing of exon 2 might thus be a good compromise to safeguard the constitutive splicing of the normal αA-crystallin mRNA. The combination of a non-consensus 5’ splice site (Talerico and Berget 1990; Robberson et al. 1990), small exon size (Talerico and Berget 1994; Smulders et al. 1998), mutations in the AC-rich enhancer element (Valentine 1998), and the competition for branch sites (Noble et al. 1988; Mullen et al. 1991) may all contribute to maintaining the skipping of exon 2.

The evolutionary history of αA\textsuperscript{ins}-crystallin would seem to present a paradox: having no obvious functional advantage, yet persisting at the protein level in many mammals, and at the DNA level in almost all. Our speculation that its evolutionary maintenance might relate to constraints on cis-acting sequence elements involved in the normal splicing process, can readily be tested in future experiments.

Acknowledgements

We thank Drs. Cynthia J. Jaworski and Graeme Wistow for making the kangaroo exon 2 sequence available. This work was supported by the Research Council for Earth and Life Sciences (ALW) with financial aid from the Netherlands Organization for Scientific Research (NWO).
The evolution of an alternatively spliced exon

References

CHAPTER 7

General discussion
1. Introduction

In general, phylogenetic trees are resolved by analyzing ever longer and better-sampled sequences of DNA and proteins. However, it becomes increasingly clear that such a quantitative approach will not give the answer to all phylogenetic problems (Hillis, 1996). For instance, high bootstrap supports in phylogenetic analyses do not guarantee that the ‘true’ clade has been found (Naylor and Brown, 1997). Therefore, additional molecular markers are desired, like insertions or deletions (indels), sequence signatures, and alternatively spliced exons, which may be more unique than base substitutions. These qualitative markers might be able to identify monophyletic groups, but probably also make a better distinction between homoplasy (convergence, parallel and back mutations) and homology, which is necessary for finding the ‘true’ trees (Doyle, 1996). Until now, they were often mentioned as potential sources of information, but commonly left out or ignored in phylogenetic analyses. The aim of this thesis was to apply these molecular markers as a tool for phylogenetic purposes. This required a better understanding of the molecular evolutionary mechanisms responsible for the appearance of these markers. Additionally, these markers were used to search for specific clades within the vertebrates, in particular the placental mammals. In this chapter, the results presented in this thesis are summarized and discussed.

1.1 Indels

As a first step in achieving our goal, we studied a unique three amino acid deletion that had been found in the eye lens protein αA-crystallin (CRYAA) of two species of sloths and two species of anteaters (chapter 2). These two families represent the Pilosa, one of the two infraorders of Edentata, or Xenarthra (Simpson, 1945; Wilson and Reeder, 1993; McKenna and Bell, 1997). We now established that this indel was also present in two species of armadillos, representing the Edentata infraorder Cingulata, and in another species of anteater, while lacking in 55 species from 16 other eutherian orders, 2 species of marsupials and 34 non-mammalian vertebrates. The neighbor joining and quartet puzzling trees of a 145-bp sequence of the CRYAA gene of 39 tetrapod species supported the monophyly of Edentata with 76% and 90%, respectively (van Dijk et al., 1999). Moreover, the probability that this deletion is a genuine shared derived character for Edentata was calculated to be more than 0.99.

The first important step in identifying monophyletic groupings by the use of indels is obtaining the ‘correct’ multiple alignment. The 9-bp deletion in the edentate αA-crystallin gene is located in the middle of an exon, flanked by conserved regions, and not surrounded by obvious tandem repeats. Therefore, positioning of this deletion in the multiple alignment is unambiguous, which makes the possibility of incorrect aligning negligible. To compute the likelihood of this indel for different hypothetical trees, we basically used the method of Kishino et al. (1990). The branch lengths were estimated by using divergence times from the fossil record, complemented with molecular data. This relatively simple model considers deletions as irreversible. This is acceptable, because it appears very unlikely that a mutational mechanism can precisely restore the deletion of nine nucleotides in a protein-coding gene, after such a deletion has become fixed in a population. The method also considers the 9-bp deletion as a single event, and therefore unrealistically assumes that the chances of indel occurrences with various lengths are equal. On the other hand, the implementation in parsimony methods of substitution models that treat a gap as a fifth character has been dismissed, because this approach treats each position of a multi-site indel as an independent event. An important future task would be the design of a more realistic model which circumvents the shortcomings of these two methods.
Morphological evidence for the monophyly of Edentata is irrefutable (Engelmann, 1985; Rose and Emry, 1993). Interestingly, therefore, the scarce molecular data could until recently not, or only marginally support this monophyly, and the phylogenetic relationships within this order remained unresolved (Sarich, 1985; Stanhope et al., 1998; Hös et al., 1996; Cao et al., 1998). In our study, both the 9-bp deletion and the phylogenetic trees of the CRYAA gene independently support the monophyly of Edentata, increasing the reliability of this finding. To unravel the relationships of sloths, anteaters, and armadillos within the Edentata, additional molecular data would be very helpful.

For this CRYAA study, only one indel was examined as a potential phylogenetic source. To explore the phylogenetic potential of additional indels of various types, a pool of indels was collected from several protein-coding genes. Tumor necrosis factor alpha (TNFα) and islet amyloid polypeptide precursor (IAPP) were selected from the data bases of protein-coding genes because of the presence of indels that might be informative for resolving the phylogenetic position of both the guinea pig and the rabbit (chapter 3). The DNA sequences of TNFα were used for phylogenetic analyses by maximum likelihood (ML) treating indels as missing data, so that they do not contribute to the obtained topology and can independently be explored as potential markers for supporting phylogenetic inferences. This TNFα phylogenetic tree supported the monophyly of rodents, with as closest relative the Lagomorpha. The additional phylogenetic information of the indels under study was consistent with the topology of this tree, and could easily be identified just by visual inspection of the sequence alignments.

Consistent with the indels of TNFα and IAPP, phylogenetic information from another molecular marker, the retroposon element BC1, also suggested rodent monophyly (Martignetti and Brosius, 1993). The disadvantages of using SINEs, like BC1, are 1) the possible integration of several SINE insertions at the same time, although at different sites (Miyamoto, 1999); 2) incomplete lineage sorting (Miyamoto, 1999; Hillis, 1999) and 3) also the practical limits of detection beyond 30% difference in sequences flanking orthologous SINE elements (Shedlock and Okada, 2000). Other molecular markers used to infer phylogenetic relationships of vertebrates are the gene order in mitochondrial genomes, indels in intronic regions, gene duplications, signature sequences and genetic code variants (Rokas and Holland, 2000). The use of these rare genomic changes has gained interest because the analyses of primary sequence data are not without problems. The variation in rates of nonsynonymous and synonymous substitutions among sites, genes and at the level of species, but also functional selection at the molecular level and homoplasy of nucleotides are only some of the difficulties that have to be considered when using primary sequence data.

However, indels are also not free from homoplasy. Therefore, various variables should be taken into account when analyzing their phylogenetic potential. It is important to include in the analysis the length of indels. A three-residue deletion is much rarer than a one-residue deletion, and thus less prone to homoplasy. Secondly, the frequency ratio of deletions to insertions varies from 1.3 to 4.0 within different sequence data (de Jong and Ryden, 1981; Graur et al., 1989; Golenberg et al., 1993; Saitou and Ueda, 1994; Gu and Li, 1995; Ophir and Graur, 1997), and therefore should be modeled differently. Also important is the position of the indel in the protein-coding gene; it is most preferably associated with the less-constrained surface loops (Pascarella and Argos, 1992; Benner et al., 1993). We finally propose to distinguish indels on basis of their underlying mutational mechanism. It is known that long indels are mainly created by unequal crossing over and DNA transposition, while slipped-strand mispairing and intron-exon boundary sliding account for indels up to 20-30 nucleotides in length. Still, it would be quite difficult to design a phylogenetic method for the coding of gaps,
because several variables of this complex event are not yet explored and probably are not independent. Thus, it is necessary to get a better understanding of the underlying mechanisms by which indels originate.

1.2 Sequence signatures

The above studies show the convincing added value of indels as a phylogenetic marker. However, they occur infrequently and thus other quantitative markers such as sequence signatures are desired. Protein sequence signatures can be defined as characteristic combinations of amino acid replacements that represent synapomorphous character states for a group of species. We developed an objective approach to search for protein sequence signatures in support of the Afrotheria, that includes the elephant shrews (Macroscelidea), golden moles (Chrysochloridae), tenrecs (Tenrecidae), aardvark (Tubulidentata), and the paenungulates (elephants, sea cows, and hyraxes). The Afrotheria are very well supported by nuclear and mitochondrial gene sequences (Springer et al., 1997; Stanhope et al., 1998; Gatesy et al., 1999; Springer et al., 1999; Madsen et al., 2001). However, morphologically this assemblage is disputed, and challenges the monophyly of Lipotyphla (Insectivora) which includes golden moles and tenrecs (MacPhee and Novacek, 1993; Butler et al., 1988). After scrutinizing all nine protein sequence data sets that have been reported to support Afrotheria, we found signatures for this clade in CRYAA, aquaporin-2 (AQP2), and interphotoreceptor retinol-binding protein (IRBP) which were statistically highly significant (chapter 4). The posterior probability of the mammalian tree containing the afrotherian clade is effectively 1.0 using likelihood, combinatorial, and Bayesian methods. Furthermore, these signatures appear highly diagnostic for identifying newly investigated species in the Afrotheria, as is shown for the otter shrew (Micropotamogale lamottei).

To search for signatures characterizing the afrotherian clade, we allowed 20% back or otherwise superimposed replacements, and the same absolute number of parallel replacements in the other ingroup sequences. Therefore, combinations of apomorphous amino acid replacements were found that did not exclusively occur in a subset of the ingroup sequences. However, great variation in rates of nonsynonymous substitutions occurs within and among genes, and consequently their level of homoplasy is not equal. Relative levels of homoplasy between base characters within the data set could have been measured by using the consistency (CI) or retention index (RI). However, accepting more homoplasy at sites that have a greater variation in rate of substitution is not very helpful in finding synapomorphic characters. Actually, we prefer to find a specific combination of apomorphous positions that is exclusively present in a set of species. Thus, by allowing only 20% of homoplasy we selected for the most conserved sites which are of main interest.

We realize that the sites of signatures already contribute to the topology and support values obtained by regular tree reconstruction procedures, especially those based on parsimony and likelihood. Therefore, to calculate the likelihood of signatures we needed the topology and branch lengths from data sets independent of the protein data set from which the signatures were extracted. This resulted in highly significant likelihoods for the signatures in these three proteins. We assumed that the sites in a signature evolve independently, but it is possible that they co-evolve. For example, in primates the middle-wavelength-sensitive and long-wavelength-sensitive visual pigments can be distinguished by the amino acids AFA in the green pigment and SYT at the corresponding positions in the red pigment. Amino acid substitutions at these sites cause a shift in the maximum wavelength of absorption ($\lambda_{\text{max}}$) (Yokoyama and Radlwimmer, 1998). However, it seems very unlikely that such parallel evolutionary mechanisms, also at the genomic level, occur in three functionally unrelated proteins.
In the data sets under study the afrotherian signatures were the most easily seen. However, searching with a window suitable to detect Afrotheria also revealed the Cetartiodactyla as another superordinal clade that is supported by noticeable sequence signatures in the protein IRBP. After adapting the search window for paenungulates as a superordinal clade, sequence signatures were found most notably in $\alpha$-2B adrenergic receptor (A2AB) and IRBP. This would suggest that relatively large numbers of nonsynonymous substitutions were fixed in the last common ancestral lineages to Afrotheria, Cetartiodactyla and Paenungulata. Since these lineages can not have been particularly long in terms of evolutionary time, accelerated rates of evolutionary change are likely to be the reason. Whether drift, founder effects or other population genetical bottlenecks have been involved remains an interesting issue for future research.

Sequencing the complete gene can in some cases be too laborious relative to the expected additional phylogenetic information. Therefore, it could be meaningful to restrict the sequencing efforts to phylogenetically informative regions. In birds, five phylogenetically diagnostic replacements are encoded in the third exon of the CRYAA gene that might distinguish whether the hoatzin (Opisthocomus hoazin) is more closely related to Galliformes (the fowl-like birds) or to Cuculiformes (the cuckoos) (chapter 5). Two of these replacements, 122 $S \rightarrow A$ and 147 $Q \rightarrow P$, apparently originated in the lineage to the neognathous birds, after the divergence of the palaeognaths (rattles and tinamous), and favor the grouping of the hoatzin within the neognaths. Within this group, after the divergence of the Galliformes and Anseriformes (duck-like birds), the remaining neognaths, including the hoatzin, share the amino acid replacements 127 $S \rightarrow A$ and 152 $P \rightarrow A$. (with the exception of 152-T in ani (Crotophaga sulcirostris; Cuculiformes)). This already indicates that hoatzin groups closer to the other neognaths than to the galliform-anseriform clade.

The fact that the hoatzin shares one of the five replacements, 135-S, with galliforms and paleognaths would suggest that the hoatzin is the first to split off from the remaining neognaths. Additionally, a shared derived character was found for the hoatzin and the groove-billed ani, the replacement 155 $S \rightarrow V$, which requires two base substitutions. In conclusion, the weight of evidence of these various amino acid replacements thus favors a closer affinity of the hoatzin to the Cuculiformes than to the Galliformes. Obviously, these few phylogenetically informative sites can not compete with very long sequences, but they have definitely a great diagnostic value. Additionally, the phylogenetic trees of the CRYAA DNA sequences performed with the neighbor joining, maximum parsimony, and maximum likelihood analyses showed the hoatzin-ani clade. Further analysis based on the combined data set of the nuclear CRYAA gene, the mitochondrial 12S-16S rRNA and the earlier studied cytochrome b sequences showed an even higher support for the hoatzin-cuckoo relationship (Hedges et al., 1995).

The closer affinity of the hoatzin to the cuckoos than to the galliforms, as concluded from our analyses, is in contrast with most morphological studies (Bock, 1992). This relationship is, however, in agreement with earlier DNA-DNA hybridization evidence (Sibley and Ahlquist, 1990). Recently, after completion of our sequence studies, six mitochondrial genes and one nuclear gene, totaling over 5.4 kb of aligned sequences, were analyzed to resolve the phylogenetic position of the hoatzin (Hughes and Baker, 1999). This resulted in support for a close relationship between hoatzin and turacos, as a sister clade to cuckoos. In our data set turacos were not included and, therefore, it might be of interest to determine the sequence of their CRYAA third exon to examine the more precise positioning of the hoatzin within this clade.
General discussion

In conclusion, once phylogenetically informative sites are identified, they appear to be useful synapomorphic characters to identify the inclusion of newly studied species into monophyletic groupings.

1.3 Alternative splicing

The number of partial and complete genome sequences, also from mammalian species, rapidly increases, and comparing chromosomes will therefore be potentially useful for exploring phylogenetic hypotheses in the future (O’Brien et al., 1999). One of the evolutionary events occurring at higher genomic level is alternative splicing, a mechanism to increase the diversity of gene products. We examined the molecular evolution of the alternatively spliced 69-bp exon 2 of the CRYAA gene, resulting in the formation of $\alpha^\text{ins}$-crystallin (chapter 6). It is expressed in some unrelated mammalian species like pika and bat, and in all studied rodents while lacking in other mammals (Hendriks et al., 1988). Southern blot analysis showed the presence of exon 2 in all investigated species, representing the 18 eutherian orders, except in golden mole and hedgehog, despite the fact that mammals like elephant, human, seal, bovine and whale do not express $\alpha^\text{ins}$-crystallin. Thus, the expression of exon 2 has been silenced independently in various lineages.

Comparison of DNA sequences revealed the non-consensus GC donor splice site for exon 2, instead of the consensus GT (King and Piatigorsky, 1983; van den Heuvel et al., 1985; Hendriks et al., 1987; Smulders et al., 1998), in 16 investigated species. In rodents, the non-consensus splice site attributes to alternative splicing of exon 2 resulting in the expression of 10-20% of $\alpha^\text{ins}$-crystallin (King and Piatigorsky, 1983). Other factors probably associated with exon skipping are the hardly detectable branch point sequence and 3’ pyrimidine-rich region in the adjacent downstream intron. In mammalian introns, the branch point sequence is usually located 18 to 40 nucleotides upstream from the AG acceptor, and the most common nucleotide preceding the branched nucleotide is the G (Mount et al., 1992). The strength of the branch point is mainly determined by the match of the branch point sequence to U2 snRNP but also by the presence of the downstream pyrimidine-rich region and of the conserved AG dinucleotide (Libri et al., 1992). Thus, probably the weak branch point, lacking the preceding G, and the absence of a downstream pyrimidine tract in the intron following exon 2 also account for exon skipping in CRYAA. However, further research is necessary to unravel the underlying evolutionary changes responsible for this exon skipping event. Therefore, it would be very useful to compare in more detail the sequences of rabbit, not expressing $\alpha^\text{ins}$-crystallin, and the related pika, which does express the alternatively spliced product. This will reveal potential nucleotide substitutions in the insert peptide and its adjacent introns associated with exon skipping. Consequently, these mutations would be useful in designing rabbit CRYAA gene constructs for expression studies. Splicing could then be monitored by determining the expression of $\alpha^\text{A}$- and $\alpha^\text{ins}$-crystallin after transfecting these genes in Chinese Hamster Ovary cells.

The higher level of nonsynonymous substitutions in exon 2 than in the remainder of the CRYAA sequence implies a lower functional constraint of the insert peptide. The functional constraints are expected to be even lower in the human and orang utang, which do not express $\alpha^\text{ins}$-crystallin because they share a 1-bp deletion in exon 2 that causes a frame shift. Although the human clearly shows an accelerated rate of change in its pseudo-exon, the sequences are too short to draw conclusions on possible higher rates compared to the expressed alternatively spliced exon 2. The functional relevance of the $\alpha^\text{Ins}$-subunit was questioned by earlier studies, especially because the insertion considerably reduces chaperone-like activity (Smulders et al., 1995). The $\alpha^\text{Ins}$-crystallin sequence thus probably is functionally not
important, and evolves under selectively neutral constraints, possibly being maintained by containing as cis-acting elements for proper splicing of the CRYAA transcript.

2. Concluding remarks and perspectives

During the past few years, the number of sequences in the data bases has greatly increased, and it is therefore worthwhile to screen them for rare indels and signatures which are phylogenetically promising. Despite their limited phylogenetic resolving power, screening alignments for synapomorphic characters like sequence signatures and indels may help to visualize monophyletic groups. These phylogenetic markers can also function as a diagnostic character when sequence data of a newly investigated species is scrutinized. They further can be useful as independent sources of information to decide between alternative phylogenetic hypotheses.

Although indels and signatures provide phylogenetic information and are to a certain extent immune for some of the problems that affect primary sequence data, they are also hampered by convergence, parallelism and reversion. Moreover, the lack of adequate statistical analyses for these markers prevents their regular use in molecular phylogeny. Therefore, it is worthwhile to explore the mechanisms that are involved in the evolutionary processes giving rise to these molecular markers, and implement these in the alignment programs. To achieve this goal, it might be helpful to compare sequences of closely related viruses, or to follow the evolution of one type of virus in time during exposure to various selective factors. The advantage of using viruses is their relatively high rate of evolution. However, extrapolation of such a system to vertebrate evolution can be problematic. Yet, the results of such studies may contribute to obtaining the ‘correct’ alignment, which is equally important for searching phylogenetic markers as for the reconstruction of trees from primary sequence data. Reconstructing trees from misaligned sequences results in extra-long branch lengths, and will likely have incorrect branch points with a tendency to migrate to the root of the tree (Swofford, 1996).

I think that, in the future, when the evolutionary processes of the various indels and sequence signatures have been further explored, we will be able to make a better discrimination between homoplasy and homology. This is essential for obtaining the ‘right’ phylogenetic tree. In addition to the molecular markers that we used, comparison of chromosomes will reveal inversions, translocations and duplications that also seem promising sources in unraveling phylogenetic problems.
References


Summary

Phylogenetic trees are nowadays generally resolved by analyzing ever longer and better-sampled sequences of DNA and proteins. Unfortunately, DNA sequences are seriously affected by convergent, parallel and back mutations, jointly referred to as homoplasy. Therefore, molecular characters are desired that are more unique than simple base substitutions, such as insertions or deletions (indels), protein sequence signatures or alternatively spliced exons. This thesis focuses on the potential of these qualitative markers as tools for phylogenetic purposes. Additionally, they are used to identify monophyletic groups within the vertebrates, in particular the placental mammals.

Although indels in protein-coding DNA are relatively rare, and thus a potentially valuable source of phylogenetic information, they are generally removed from multiple alignments before further analysis. This is done because they mostly occur in regions with the greatest sequence variability, corresponding with surface loops in the encoded proteins. Consequently, their positioning within a multiple alignment is often ambiguous. In Chapter 2, we analyse a deletion of three amino acids that occurs uniquely in the eye lens protein αA-crystallin of the Edentata. Using phylogenetic trees with alternative hypothesized relationships and branch lengths, the probability was calculated that this deletion originated repeatedly or only once during mammalian evolution. A single occurrence had a probability of more than 0.99, and the deletion thus strongly supports the monophyly of Edentata. In this study, only a single indel was examined as potentially phylogenetically informative. However, the origin and evolution of different indels varies according to the involved mutational processes and constraints. Therefore, a pool of indels was traced from several protein-coding genes to explore their value as independent phylogenetic markers. Two genes, coding for tumor necrosis factor alpha (TNFα) and islet amyloid polypeptide precursor (IAPP), were found to contain putatively informative indels to help resolve the phylogenetic position of both guinea pig and rabbit (Chapter 3). Phylogenetic analysis by maximum likelihood (ML), using the DNA sequences of TNFα and treating indels as missing data, showed the monophyly of rodents, with Lagomorpha (rabbits and pikas) as their sister group. The additional phylogenetic information of the indels in TNFα and IAPP could easily be identified by visual inspection of the sequence alignments, and was in agreement with the topology of this tree.

Both these studies showed that indels can provide convincing additional information as phylogenetic markers. However, suitable indels are not frequently observed, and other quantitative markers such as sequence signatures were thus desirable. Protein sequence signatures can be defined as characteristic combinations of amino acid replacements that represent synapomorphous character states of a group of species. They provide ‘protein-morphological’ characters that may aid in visualizing candidate novel monophyletic groupings. In Chapter 4, we identified signatures in three protein sequences, CRYAA, AQP2, and IRBP, which support a clade of mammals, now called Afrotheria, that includes elephants, sea cows, hyraxes, aardvark, elephant shrews, golden moles and tenrecs. Morphologically, this appears an unlikely assemblage, which challenges - by including golden moles and tenrecs - the monophyly of the order Insectivora (or Lipotyphla). The signatures also functioned as a diagnostic tool to identify newly investigated species, namely the morphologically divergent African insectivore the otter shrew (Micropotamogale lamottei) was assigned to the Afrotheria. The statistical support for the sequence signatures as unambiguous synapomorphic evidence for the naturalness of the afrotherian clade, was calculated to be highly significant.

Sequencing just a selected small part of a gene can in specific cases be adequate to retrieve phylogenetically informative sites. This is illustrated in Chapter 5, where a 146-bp sequence in the third exon of the αA-crystallin gene is used to solve a problem in avian phylogeny. Five phylogenetically informative sites encoded in this exon group the hoatzin (Opisthocomus hoazin), a South American bird of disputed affinities, with the cuckoos (Cuculiformes) instead
of the fowl-like birds (Galliformes). Phylogenetic trees constructed from the two mitochondrial genes for 12S and 16S rRNA, together with αA-crystallin sequences, also indicated that the hoatzin shares a more recent common ancestry with Cuculiformes than with Galliformes.

In Chapter 6 we studied the molecular evolution of the alternatively spliced exon 2 of the αA-crystallin gene, which results in the formation of αA<sup>ins</sup>-crystallin. αA<sup>ins</sup>-Crystallin is expressed in all studied rodents and in some unrelated mammalian species like hedgehog, pika and bat, while it is lacking in other mammals. The presence of exon 2 was now assayed by Southern blotting of genomic DNA of 23 species, representing all 18 orders of placental mammals. Surprisingly, hybridization signals for the insert peptide were detected in all species, apart from the golden mole. It thus appears that exon 2 was present in the common mammalian ancestor, and that its expression was independently silenced in various mammalian lineages.

The evolutionary processes that lead to alternative splicing, as a means to increase the diversity of gene products, are poorly understood. The mutational changes underlying the skipping of exon 2 were therefore revealed by sequence determination of this exon and its adjacent introns in species known to express αA<sup>ins</sup>-crystallin. In all species exon 2 was flanked by the non-consensus donor splice site GC, involved in the alternative splicing of rodent αA-crystallin genes. Of further relevance, the adjacent downstream intron had a poorly recognized branch point, lacking the preceding nucleotide G, and no detectable downstream pyrimidine-rich region, both factors that are important in splice site competition. Moreover, the insert peptide is apparently subject to lower functional constraints as compared to the remainder of the αA-crystallin protein. This indicates that the expression of αA<sup>ins</sup>-crystallin is a selectively neutral character.

This thesis presents informative data on the application of indels, protein sequence signatures and an alternatively spliced exon as markers in molecular phylogeny. The obtained insights lead to new scientific questions, and additional studies are therefore required to further explore their potential usefulness. It will be a challenging task to implement the obvious phylogenetic signal of these molecular markers as algorithms in the currently used programs for phylogenetic analysis.
Samenvatting


Om de wereld om ons heen te ordenen en beter te leren begrijpen is de evolutiebioloog geïnteresseerd in de ontrafeling van ‘de boom van het leven’ die de evolutie van soorten uitbeeldt. Tot voor kort werd de evolutiebioloog gezien als een natuurliefhebber die vlinders, bloemen en andere levende organismen in het veld verzamelde, maar ook fossielen van o.a. dinosauriërs, om ze daarna te kunnen vergelijken. Met de komst van nieuwe technieken en methodologische toepassingen, en door de verbreding van onze kennis over de evolutie, heeft de studie van de evolutiebiologie het afgelopen decennium een revolutie doorgemaakt. Mede door de ontwikkeling van het PCR-apparaat en de computertechnologie heeft de moleculaire evolutie zijn intree gemaakt. Met de *polymerase chain reaction* (PCR) kun je een gewenst stukje DNA uit een grote mix van dit erfelijk materiaal amplificeren. DNA is voor te stellen als een ketting waarin vier verschillende kleuren kralen op oneindig veel manieren aan elkaar gerijgd kunnen zijn. Elk stuk DNA heeft een specifieke volgorde van deze kralen, waarin de erfelijke informatie vastligt. Hoewel elke soort (muis, olifant of mens) dezelfde informatie nodig heeft om te leven zijn er verschillen in de volgorde van de kralen bij verschillende soorten. Soorten die nauw verwant zijn, zoals muis en rat, zullen DNA hebben dat meer op elkaar lijkt dan bijvoorbeeld dat van muis en mens. Voor de reconstructie van de evolutie kunnen we bijvoorbeeld een gen - een stukje DNA dat de erfelijke informatie bevat voor een bepaalde eigenschap - van verschillende soorten dieren vergelijken. Hierdoor worden de genetische overeenkomsten en verschillen zichtbaar. Met computeranalyses kunnen van deze gegevens dan evolutiebomen gemaakt worden om de verwantschappen tussen levende organismen te bestuderen.

Kennis uit evolutiestudies kan ook gebruikt worden in andere wetenschapsgebieden, zoals de bioinformatica, om bijvoorbeeld meer te weten te komen over de functie van een gen. Misschien minder duidelijk zichtbaar is het belang van evolutieonderzoek voor sociaal-economische en juridische vraagstukken. Zo is er in Amerika een rechtzaak geweest waarbij gebruik is gemaakt van moleculair evolutionaire analyses om de bron van een HIV-overdracht te kunnen vaststellen. Dit is mogelijk door het pad waarlangs deze infectieuze ziekte heeft gelopen te reconstrueren. Maar ook bij het oplossen van een probleem als resistentie tegen antibiotica en pesticiden, veroorzaakt door overmatig gebruik hiervan, kan kennis van evolutionaire processen behulpzaam zijn.

In dit proefschrift worden resultaten van moleculaire studies beschreven met als hoofdhema het onderzoek naar monofyletische groepen - oftewel soorten die één gemeenschappelijke voorouder hebben - binnen de vertebraten (vissen, amfibieën, reptielen, vogels en zoogdieren). Mutaties kunnen de evolutie van een soort beïnvloeden. Een mutatie houdt meestal in dat op een willekeurige plaats in de lange DNA keten één van de bouwstenen, de z.g. nucleotide-basen, wordt vervangen door een andere. Parallele, convergente en terugmutaties van deze basen, tezamen ook wel *homoplasie* genoemd, treden vaak op in de evolutie en vertroebelen de analyse van DNA-volgorden. Daarom is het gebruik van andere genetische kenmerken, zoals inserties en deleties (*indels*), *protein sequence signatures* of alternatief gespliceote exonen, van potentieel belang omdat deze unieker zijn dan baseveranderingen. De mogelijke toepassing van deze kwalitatieve kenmerken voor fylogenetische doeleinden is daarom onderzocht. Indels worden vaak verwijderd uit een alignment van DNA-volgorden van de onderzochte soorten.
voordat de fylogenetische analyses plaatsvinden. De reden is dat indels vaak moeilijk zijn te plaatsen doordat ze meestal liggen in gebieden met de grootste variatie in sequentie. Deze gebieden komen in eiwitketens overeen met lussen die aan de oppervlakte liggen.

In hoofdstuk 2 hebben we de toepassing onderzocht van een deletie met een lengte van drie aminozuren. Deze indel komt voor in het ooglenseiwit αA-crystalline (CRYAA) en is uniek voor Edentata (de ‘tandarme’ zoogdieren van Zuid-Amerika). De kans dat deze indel slechts één keer is opgetreden, en wel in de gemeenschappelijke voorouder van de Edentata, is groter dan 0.99. Indels kunnen op verschillende manieren ontstaan, waarbij factoren als mutatieprocessen en selectiedruk van invloed zijn. Daarom hebben we ons onderzoek uitgebreid met indels in de eiwit-coderende genen voor tumor necrose factor (TNFα) en islet amyloid polypeptide precursor (IAPP). Onafhankelijk van elkaar tonen zowel de maximum likelihood boom afgeleid van TNFα DNA-sequenties als ook de visuele inspectie van TNFα en IAPP indels aan dat knaagdieren monofyletisch zijn. Verder laten dezelfde analyses zien dat de Lagomorpha (konijnen, hazen en fluithazen) de naaste verwanten van de knaagdieren zijn (hoofdstuk 3).

Hoewel indels een toegevoegde waarde hebben als fylogenetische kenmerken, komen ze niet frequent voor. Het is daarom noodzakelijk andere kwalitatieve kenmerken te gebruiken, zoals protein sequence signatures. Wij definiëren deze als specifieke combinaties van aminozuurveranderingen die een gezamenlijk kenmerk vormen voor een groep organismen met één gemeenschappelijke voorouder. In hoofdstuk 4 hebben wij protein sequence signatures geïdentificeerd in drie eiwitsequenties - CRYAA, aquaporine-2 (AQP2) en interphotoreceptor retinol-binding protein (IRBP) - die de monofylie van de Afrotheria ondersteunen. Tot deze groep behoren olifanten, zeehonden, klipdassen, aardvarken, olifantspitsmuizen, goudmollen en tenreks. Hoofdstuk 5 laat zien dat ook korte DNA-sequenties belangrijke fylogenetische informatie kunnen verschaffen, in dit geval over de verwantschap van de Hoatzin (Opisthocomus hoazin), een Zuid Amerikaanse vogel, met de koekoekachtigen in plaats van met de hoenderachtigen. Vijf fylogenetisch informatieve posities in een stukje van 146 baseparen, coderend voor het derde exon van het αA-crystalline gen, bleken hiervoor voldoende.

Hoofdstuk 6 beschrijft de moleculaire evolutie van het alternatief gespliceerde exon 2 in het αA-crystalline gen. Het product van deze alternatieve splicing, αAins-crystalline, komt tot expressie in alle onderzochte knaagdieren en in sommige niet verwante zoogdieren zoals egel, fluithaas en vleermuis. In andere zoogdieren is geen expressie. Met behulp van een Southern blot werd niettemin nu aangetoond dat exon 2 aanwezig is in het DNA van vertegenwoordigers van alle 18 ordes van placentale zoogdieren. Het lijkt er dus op dat exon 2 aanwezig was in de gemeenschappelijke voorouder van de zoogdieren en dat de expressie ervan in verschillende zoogdiergroepen onafhankelijk van elkaar is verdwenen. Over de evolutionaire mechanismen die alternatieve splicing veroorzaken, nodig om de diversiteit van genproducten te vergroten, is slechts weinig bekend. Daarom is de DNA-volgorde van exon 2 en zijn aangrenzende intronposities bepaald bij zoogdieren die αAins-crystalline tot expressie brengen. Dit laat zien dat al deze dieren een non-consensus donor splice site GC bezitten, waarvan is aangetoond dat het verantwoordelijk is voor de alternatieve splicing van het αA-crystalline gen bij knaagdieren. Andere factoren die waarschijnlijk van invloed zijn op de skipping van exon 2 zijn de slecht herkenbare branch point die ook nog de voorafgaande G nucleotide mist en het niet aanwezig zijn van het pyrimidine-rijke gebied in het aangrenzende stroomafwaartse intron. Verder onderzoek is nodig om de ingewikkelde evolutionaire mechanismen die alternatieve splicing op moleculair niveau veroorzaken te kunnen verklaren.

Hoe de evolutie van dieren werkelijk heeft plaatsgevonden blijft moeilijk te bestuderen omdat het onmogelijk is om miljoenen jaren terug te kijken in de tijd. Daarom zullen door zowel paleontologen als molecular biologen geen exacte bewijzen geleverd kunnen worden. Dit kan nogal eens leiden tot heftige discussies waarin een ieder met volle overtuiging zijn
gelijk probeert te krijgen. Deze discussies motiveren de onderzoekers om verder te speuren naar mogelijke oplossingen van hun probleemstellingen. De resultaten van deze studies zullen naar mijn mening uiteindelijk heel dicht de werkelijke boom van het leven benaderen.
List of publications


Dankwoord

Het volbrengen van een promotie kan gezien worden als het leggen van een puzzel. Sommige stukjes ontbreken, de ene is makkelijk te leggen en de ander moeilijk, weer andere worden opgelost met hulp. Meegeholpen met het oplossen van deze puzzel hebben de volgende mensen:

Natuurlijk mijn promotor, Prof. Dr. Wilfried de Jong. Als er iemand van puzzelen (met alignments) houdt dan ben jij het wel. Je hebt laten zien hoe belangrijk het is om in de wetenschap alles zo precies en volledig mogelijk te beschrijven. Bedankt voor je geduld en met name je hulp bij de afronding van het manuscript.

Dan de labgenoten van de groep de Jong:
Ole, lange tijd vormde wij samen de fylogenie “groep”. Ik heb het altijd leuk gevonden om met jou samen te mogen werken. Je hebt een boeiende kijk op de wereld en ik heb bewondering voor de rust die jij uitstraalt. Heel erg bedankt voor je hulp met de fylogenetische computer problemen en je handige tips voor het uitvoeren van de experimenten.
Guido, de hele dag krijg je een hoop vragen naar je hoofd geslingerd van mensen die iets willen weten over computers, ook ik was er één van. Je was altijd bereid om te helpen.
Bedankt, ik heb er veel van opgestoken.
Ook denk ik terug aan de hulpvaardigheid van met name Pauline en Bas, zelfs als ik er niet om vroeg. Ook buiten het lab hebben wij veel tijd met elkaar doorgebracht. Ik hoop jullie nog regelmatig te zien.
Verder hoop ik dat jullie niet al te veel last hebben gehad van mijn gezang en drukke buien die ik vooral toonde in het bijzijn van Marcel. Want ja Marcel, je was mijn enige student maar je hebt dat dan ook gecompenseerd door extra lang te blijven. Je begin-periode ben je goed doorgekomen ondanks mijn afwezigheid. Dit komt mede doordat je zelfstandig bent en initiatief neemt. Heel erg bedankt voor de fijne discussies over het onderzoek en de lol die we hebben gemaakt aan de labtafel. Het werk werd er een stuk leuker door en heeft ook zeker bijgedragen aan dit eind-resultaat.
Ook Els mag niet ontbreken. Je belangstelling voor mensen is groot. Zelfs toen ik van het lab verdwenen was vond ik altijd bij mijn werk nog wel een briefje met wat wijze woorden van jou. Bedankt voor je belangstelling en goede zorgen.
Ik heb nu nog een heleboel andere lab-genoten niet bij naam genoemd maar jullie waren minstens zo belangrijk voor de goede werk-sfeer op het lab. Door de vele etentjes, sportavonden en zelfs weekendjes weg ontstond er een hele hechte groep. Bedankt mensen van het lab, ik heb een fijne tijd bij jullie gehad.

Dan natuurlijk mijn familie. Lieve papa, mama en Peetje, ik kon altijd bij jullie terecht voor de nodige advies en goede zorgen. Bedankt dat ik zo’n fijn “thuis” heb.
En als er iemand een bijdrage aan dit proefschrift heeft geleverd dan is het Collin wel. Lieve Collin, je hebt me heel goed geholpen met onder andere het maken van een computer-programmaatje en de uitleg van ingewikkelde wiskundige formules. Maar bovenal ben je zeer geduldig met mij geweest. Je bent een grote steun.

De puzzel is klaar.

Marjon
Curriculum vitae

Marjon van Dijk werd geboren op 10 juli 1970 te Kampen

1982 - 1988 VWO aan het St. Bonifatius College te Utrecht

1988 - 1994 Medische Biologie aan de Universiteit Utrecht (UU)

1994 - 2000 Promotie aan de Katholieke Universiteit Nijmegen bij de afdeling Biochemie (Faculteit Natuurwetenschappen, Wiskunde en Informatica)

2000 - 2001 Postdoc (moleculaire biologie) aan de Erasmus Universiteit Rotterdam bij de afdeling Immunologie (Faculteit Geneeskunde en Gezondheidswetenschappen)