Some histories of molecular evolution: amniote phylogeny, vertebrate eye lens evolution, and the prion gene

Teun van Rheede

Some Histories of Molecular Evolution: Amniote Phylogeny, Vertebrate Eye Lens Evolution, and the Prion Gene

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
op gezag van de Rector Magnificus Prof. Dr. C.W.P.M. Blom,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op dinsdag 18 mei 2004
des namiddags om 4.00 uur precies
door

Teun van Rheede

geboren op 23 januari 1973
te Utrecht

Omslagillustratie: Aletta van Rheede
Lay-out: Alex Verkade en Pleuni Pennings
Druk: PrintPartners Ipskamp BV, Enschede

Promotor:

Prof. dr. W.W. de Jong

Manuscriptcommissie:

Prof. dr. J.A.M. Leunissen (WUR)
Prof. dr. J.J.H.H.M. de Pont
Prof. dr. W.J. van Venrooij

Every man takes the limits of his vision for the limits of the world.

Schopenhauer

The research described in this thesis was performed at the Department of Biochemistry, Faculty of Science, Nijmegen Center for Molecular Life Sciences, University of Nijmegen, The Netherlands.
This work was supported by the Netherlands Foundation for Scientific Research (NWO-ALW).

Table of Contents

Voorwoord en verantwoording

Ieder proefschrift is het eindproduct van een intensieve wisselwerking tussen promovendus, begeleiders en overige collega's. Het is dus nooit duidelijk wat precies de bijdrage van de promovendus zelf is. Maar wél is de promovendus uiteindelijk verantwoordelijk voor de gehele inhoud van het proefschrift zoals dat ten slotte verdedigd wordt. Bij dit postume proefschrift van Teun van Rheede is dit niet het geval. Het is daarom goed om duidelijk te maken welke delen van dit proefschrift wél door Teun (mede)geschreven zijn en gezien, en welke niet meer.

In januari 2003, toen Teun's fatale ziekte werd vastgesteld, waren al drie artikelen voltooid, de hoofdstukken 3, 5 en 7 van dit proefschrift. Voor twee andere artikelen, de hoofdstukken 2 en 4, had hij de eerste versies geschreven. Daarnaast lagen er gegevens en ideeën voor nog meerdere artikelen, de meeste in samenwerking met anderen. Daaruit is hoofdstuk 6 voort gekomen en andere artikelen zullen nog volgen. Zijn OIO-aanstelling liep tot 1 januari 2004 en hij zou dus alle tijd gehad hebben om er een mooi proefschrift van te maken. Toen duidelijk werd dat hij niet meer zou genezen, was het Teun's diepe wens om toch nog zo veel mogelijk van zijn proefschrift af te ronden. Tijdens zijn ziekte bleef hij, wanneer zijn toestand dat toe liet, vooral bezig met de inleiding van zijn proefschrift, maar hij had ook voortdurend vragen en suggesties voor de resterende hoofdstukken. Het gaf hem afleiding en ook voldoening om de resultaten van zijn promotieonderzoek te zien.

Wij hadden nog enkele keren "werkoverleg", in Amersfoort en via e-mail, maar Teun vorderde uiteraard langzaam en vond zijn inleiding nog lang niet geschikt om aan mij voor te leggen. Toen het einde naderde heeft hij met veel moeite zijn proefschrift "losgelaten" en alles wat hij aan ideeën en fragmenten geschreven had van zijn laptop op flop gezet en mij toegestuurd. Op de flop vond ik de files die de basis vormen voor de introductie van dit proefschrift. Ik heb zoveel mogelijk getracht Teun's bedoelingen en stijl intact te laten. Doordat Teun nog geen referenties noemde in zijn fragmenten voor de inleiding, ontbreken die hier ook. Teun had "boxjes" willen gebruiken om zijn proefschrift '*wat op te leuken*'. Eén zo'n boxje had hij wat uitgewerkt en is in aangepaste vorm bij de inleiding gevoegd. Ook had hij een motto voor zijn proefschrift en het begin van een collectie stellingen, die hij echter '*wat zwak om ook echt te gebruiken*' vond. Ik noem er toch enkele, omdat hij er kennelijk plezier aan gehad heeft:

** DNA is toch voordeliger.*

** Een moleculair phylogeneet is wel degelijk blij te maken met een dode mus.*

** A miracle in the first place, cows will be cows. (Wislawa Szymborska)*

** Het is een misvatting dat je door jezelf hardnekkig center of excellence te noemen, je kan uitgroeien tot een center of excellence.*

** Als het aantal slachtoffers in aanmerking wordt genomen, krijgt de gekke koe (BSE) in vergelijking tot de 'heilige koe' (de auto) onevenredig veel aandacht.*

** Het prion wordt met uitsterven bedreigd.*

Op de flop stond verder Teun's dankwoord, zoals op bladzijde 120 ongewijzigd opgenomen, en het volgende voorwoord:

Losse flard

Noodgedwongen is het een heel ander boekje geworden dan het had moeten zijn. Niet het ambitieuze proefschrift met veel hoofdstukken en goede publicaties. Door m'n ziekte werd het minder een stap in een wetenschappelijke loopbaan, en meer een boekje waar ik zoveel mogelijk van mezelf in wil leggen. Dat wil zeggen dat de ambitie er nog makkelijk in terug te vinden is, maar dat ik meer probeerde los te komen van de formele en vaak saaie manier van rapporteren die gebruikelijk is in de wetenschap.

Ik hou van lesgeven, dus ik wil dingen uitleggen op een manier die mensen enthousiast maakt. Ik hou van schrijven, dus ik wil taal gebruiken om extra lagen toe te voegen aan de 'proefschrift-eenheidsworst'. Ik hou van diepgang, dus ik veronderstel de basis bekend en gooi er wat mijmeringen tegenaan over leven en hoe dat nou eigenlijk moet. Enerzijds wil ik een academisch acceptabel boekje fabriceren, anderzijds wil ik proberen te ontsnappen aan de banaliteit die een proefschrift kan krijgen vanaf het moment dat je beseft dat het een van de laatste dingen is die je zal schrijven.

(maar heb het dus niet helemaal geschreven.)

In zijn laatste e-mail, van 4 mei, stuurde Teun me tenslotte nog de titel voor z'n proefschrift. Niet door Teun (mede) geschreven en gezien zijn de hoofdstukken 6 en 8, de samenvatting en zijn CV.

Nu het proefschrift klaar is, moeten aan Teun's dankwoord zeker worden toegevoegd Pleuni Pennings, Aletta van Rheede en Alex Verkade die de vormgeving verzorgd hebben, en Mirko Opdam en Gé Seiger die - als 'paranimfen' nog door Teun gevraagd - veel hebben bijgedragen aan de voorbereiding van de promotie.

Wilfried de Jong

Chapter 1

General Introduction

Molecular Evolution

In the early 50's, Watson and Crick formulated a working hypothesis, which later became known as 'The Central Dogma':

DNA → RNA → Protein

DNA stores heritable, genetic information in an extremely stable form. RNA carries the genetic information to the cytoplasm, for it to be decoded to protein on the ribosome, and increasingly turns out to fulfil several 'non-dogma' functions. Protein performs the catalytic and structural functions the cell requires, including replication of the DNA, transcription into RNA and translation into protein.

It are these molecules, their interactions and the resulting molecular processes that are studied in biochemistry and molecular biology. Molecular evolution adds the evolutionary perspective. It encompasses the characterisation of changes in the genetic material (DNA/RNA) and its products (RNA/protein) during evolutionary time, and in addition attempts to unravel the mechanisms responsible for those changes. Change in the genetic material is caused by mutations and subsequent evolutionary processes like natural selection and random genetic drift. As a practical objective the study of molecular evolution is simple: determine the sequence of corresponding bits of genome from the set of organisms of interest and compare the resulting data. The comparison of DNA and proteins between organisms gives insight in their evolutionary history, and allows the reconstruction of ancestral states and the estimation of molecular clock rates. It shows how evolution explores the macromolecular sequence space, resulting in ever expanding protein functions (see Box 1).

Natural selection acts on functional, living units, often called organisms. Nowadays, the enormous amounts of genetic information are largely stowed away in stable DNA molecules. Something as outrageous as the genetic code, encrypted in the organisms' translational machinery, is used to decode this information into functional proteins. Of course, the information required for life is contained not only in DNA, but also in the spatial and temporal organisation of molecules, and in all sorts of add-on molecular machinery that makes for good night-time reading. But, in contrast with determining spatio-temporal patterns, DNA sequencing has become easy, everybody is doing it, and large amounts of data are available for free on the internet. Furthermore, we have some knowledge of the mechanisms of the evolution of DNA that is not available for, let's say, bones. For example, we know that transitions (R→R and Y→Y) in the DNA are much more frequent than transversions (R↔Y), again beautifully reflected in the build-up of the genetic code. Its 4-state simplicity allows building of sophisticated mathematical models of sequence evolution. But despite this simplicity of DNA, the models in turn have soon become utterly incomprehensible again, adding greatly to the scientific air and respectability of the field.

Why is DNA such a favorite source of evolutionary information? First, it provides lots of data – stored in endless nucleotide sequences in the genomes - and is easy to obtain. Second, there is some understanding of mutational processes during evolution, which enables the generation of good evolutionary models. And third - if it concerns coding DNA - the sequence information can readily be translated into hypothetical protein sequences, opening the way to explore their structural and functional evolution.

Molecular Phylogenetics

Molecular phylogenetics uses molecular data (as opposed to morphological characters) to study species relationships. While the conditions and potential for the use of DNA sequence data in phylogenetic reconstruction are excellent, things still can go wrong. Just as paleontologists have often been misled by their interpretation of bones, molecular phylogeneticists are being misled by the complexity of molecular evolutionary processes. This has led to widely publicized but erroneous claims, such as the proposal that "the guinea pig is not a rodent". For those outside the field of molecular phylogenetics this lends an air of unreliability to the conclusions presented. When the underlying phylogenetic signal is relatively weak or biassed, it is indeed a fact that the results of a phylogenetic study may change when you add or delete some taxa, when you analyze a subset of the data, or when a 'better' method is used. Credibility does not necessarily increase when the experts tell you "it's just a long branch attraction artefact" and you really should trust that his favorite tree is the true one, and we could all use a good night of sleep.

For phylogenetic analyses DNA has the enormous advantage that it presents itself in a mathematically convenient 4-state, which can be converted in a biologically meaningful way to a two-state suitable for easy computer handling. A major problem is that mathematical models make assumptions. Even while these assumptions are violated, the results may still be OK. But it can go wrong. The key assumption is that ancestral characteristics are inherited, and change over time. Thus, a species' phylogeny can be read from its genome. An essential assumption is further that the studied sequences are orthologous (that is, related by last common descent, not by duplication events). Violations of underlying assumptions are generally indicated in broad terms or mathematical abstractions like long branch attraction, GC bias, unequal rates, deviation from stationarity, or covarion structure. Coping with these common violations by improving the mathematical models is one of the greatest challenges of molecular phylogenetic inference.

Essential, too, for improving phylogenetic analyses is a better biochemical understanding of the processes of molecular evolution and implementing this in models of sequence evolution. In the advance of biochemical understanding, biochemistry/molecular biology and molecular evolution/phylogeny become mutually enlightening. Understanding the processes of molecular evolution ranges from knowing that transitions are more common than transversions, to realizing that correlated mutations occur and that evolution of coding DNA is heavily constrained by structure and function of the encoded protein. Implementing this knowledge in models of sequence evolution thus requires that transitions be downweighted in phylogenetic analysis, and the possibility of correlated mutations and constraint at the protein level taken into account.

One also should be aware that alternative splicing is a frequent process, which may complicate the interpretation of protein and cDNA sequences, and that gene duplications are common and make the distinction between orthologs and paralogs necessary. Gene families can expand and contract, and be subject to concerted evolution by gene conversions. Each gene has its own mode of evolution, and mitochondrial genes behave differently from nuclear ones.

To study evolutionary relationships by molecular approaches, any orthologous character that can be compared between a set of taxonomic groups can serve as a phylogenetic marker. This includes especially nucleotide or amino acid sequence data. It becomes, however, increasingly clear that molecular evolutionary events rarer than base substitutions, such as insertions or deletions (indels), absence or presence of introns in genes, or the occurrence of retroposable elements, can additionally serve as qualitative markers which can be highly diagnostic for a particular clade.

Circular Reasoning or Mutual Enlightening?

In molecular evolution, phylogenetic information is used to study macromolecular evolution. In molecular phylogenetics, macromolecules are used to infer phylogenies. In the relation between molecular evolution and phylogenetics one thus may wonder whether in addition to mutual enlightening there may perhaps also be some circular reasoning.

Box 1 Close Encounters in Sequence Space

Sometimes, life can be a bit boring. People have proposed to add time as a fourth dimension. I prefer to stick to three. That is not so many, but adding time seems a bit desperate. Mathematically, you can have as many dimensions as you like. As a molecular evolutionary biologist one can take advantage of this, to liven up the day. Consider protein sequence space. A single amino acid can be in 20 different states. Twenty dots on a line, to form a first dimension, if you wish. Two amino acids, a dipeptide, yields 20 x 20 possibilities, or a two-dimensional plane with 400 states, if you wish. Three amino acids yield the three-dimensinal, 20 x 20 x 20 state space. Only three!!! A peptide almost too short to perform any physiological function, exhausts three dimensions, the number that we have got used to so much. Proteins are much longer, and multidimensional space expands further and further, beyond imagination.

The vast emptiness of sequence space begins to give a notion that comes somewhat closer to perceiving life and its complexity. A gazing-at-the-endless-sky-filled-with-stars feeling. A deep feeling of wonder, a deep knowing of your limited understanding. It is well possible to lose sight of those universal, once deeply felt questions when working on fragmented questions, slipping into routine 'nine to five' working hours and worrying about the latest impact factor of your favourite scientific journal.

Of course, nobody would want to get stuck contemplating such stuff all day. Pipetors are made for pipeting, in the end.

In a far corner of sequence space, lactate dehydrogenase was evolving. Tentatively it explored neighbouring space, wobbling and throbbing, sticking out and retracking small tentacles, smoothly changing shape like a blossoming flower. An aspartate and threonine had tried changing to glycine several times now, unknowing that this disturbed tertiary structure and was selected against.

But then glutamine-122 replaced its neutral side chain by the positive charge of arginine. At the protein level, some 3.7 ångstrom further on, the carbon backbone flexed outwards somewhat, modifying the lactate binding site so that now a malate could be accomodated instead. A minute change in sequence space had resulted in a substantial change in function of a lactate dehydrogenase to a malate dehydrogenase (Wilks et al., 1988, Science 242:1541). No one ever noticed.

Sequence space is vast and empty.

Outline of the Thesis

Part I: Amniote Molecular Phylogenetics.

Birds, reptiles and mammals are Amniota, organisms that have an amnion during their embryonal development. Even though these organisms have been studied for centuries, their interrelationships remain debated in some cases. In Chapter 2, the molecular phylogenetic position of the egg-laying mammals (Monotremata) is analysed, and in Chapter 3 an example is presented of rare genomic changes - in this case deletions in protein-coding DNA - that are very useful to distinguish relationships between the orders of placental mammals. Some preliminary results concerning the molecular phylogenetic position of the tuatara, an isolated and enigmatic lineage of reptiles only surviving in New Zealand, are given in the General Discussion and Summary (Chapter 8).

Part II: Molecular Evolution of the Vertebrate Eye Lens

The molecules that constitute the eye lens are stable, transparant proteins termed crystallins. In this part we report some typical examples of the origin and molecular evolution of the vertebrate eye lens crystallins. We present the α-crystallin genes from the platypus (Chapter 4) and the discovery in this species of a novel lens protein, upsilon-crystallin, which turns out to be overexpressed lactate dehydrogenase A (Chapter 5). We further present a paper on the evolution of regulatory sequences of the αB-crystallin gene, which are located in the bidirectional promoter between the head-to-head arranged αB-crystallin and HspB2 genes (Chapter 6). The α-crystallins originated evolutionarily from the family of small heat shock proteins. Some data about the evolution of this protein family in vertebrates and lower chordates are included in General Discussion and Summary (Chapter 8).

Part III: Molecular Evolution of the Prion Protein

Chapter 7 addresses the evolutionary aspects of the mammalian prion protein and touches on the evolution of the prion protein gene family. Specifically, the evolution of a repeat region in the vertebrate prion is discussed. The remarkable finding of a deviating prion gene in the squirrel is presented in General Discussion and Summary (Chapter 8).

Chapter 2

The Phylogenetic Position of the Monotremes: Support for the Theria Hypothesis of Mammalian Phylogeny from Nuclear Genes and Indels

Teun van Rheede, Ole Madsen*, Trijntje Bastiaans*,*
David N. Boone§, S. Blair Hedges§, Wilfried W. de Jong‡*

* Department of Biochemistry, NCMLS, University of Nijmegen, The Netherlands;
§ Department of Biology, The Pennsylvania State University, USA;
‡ Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, The Netherlands.

Abstract

Mammals can be divided into three subclasses: Monotremata (the egg-laying platypus and echidna), Marsupialia (the pouched marsupials) and Eutheria (the placental mammals). Morphology considers monotremes as an early offshoot of the major mammalian lineage, the Theria, which subsequently split into marsupials and eutherians. In contrast, analyses of 12 mitochondrial protein-coding genes (together about 9500 bp) strongly support the grouping of monotremes with marsupials: the Marsupionta hypothesis. Limited nuclear gene analyses were as yet inconclusive with regard to the phylogenetic position of monotremes. We therefore determined sequences from seven nuclear genes and obtained additional sequences from the databases to create two large and completely independent nuclear datasets. Dataset I comprised 5 genes, with a concatenated length of 2793 bp, from 21 species (2 monotremes, 6 marsupials, 9 placentals and 4 outgroup species). Dataset II comprised 9 genes and 4 proteins, with a concatenated length of 11544 bp or 4050 amino acids, from 5 taxa (a monotreme, a marsupial, a rodent, human and chicken). Both datasets were analyzed by parsimony, minimum evolution, maximum likelihood and Bayesian methods. Dataset I gave bootstrap support values for Theria between 49.2% and 99.5%, depending on method and model of evolution. Support for Marsupionta was negligible, at most 10.8%, while the third alternative - grouping Monotremata with Eutheria, a relationship which has never been proposed - received support up to 49.2%. Correcting for base compositional bias (monotremes being very GC-rich and marsupials GC-poor) increased the support for Theria, and decreased support for Marsupionta. Dataset II exclusively supported Theria, with the highest possible values, and significantly rejected Marsupionta. On basis of our datasets, the time of divergence between Monotremata and Theria was estimated at 223-198 million years ago (Mya), and between Marsupialia and Eutheria at 173-151 Mya. As an independent source of phylogenetic information, we additionally searched the sequence alignments from both datasets for insertions or deletions that could be diagnostic for resolving the trichotomy of the mammalian subclasses. Supporting evidence, consisting of two single amino acid deletions and one insertion, was only found for Theria. The morphological evidence for Theria is thus fully supported by our molecular data from nuclear genes.

(in preparation for publication)

Introduction

Ever since its discovery, the platypus, *Ornithorhynchus anatinus*, has attracted attention from both scientists and the general public. Its astounding characters like the duck-like bill, a furry skin, the poisonous spur on the male's hindlegs, and its ability to swim and dive, raised the question of its phylogenetic position. The discovery that the platypus and echidna are egg-laying and suckle their young largely solved this question: monotremes are egg-laying mammals, basal to the other major mammalian clades, marsupials and placental mammals.

Molecular data revived the Marsupionta hypothesis: complete mitochondrial genome sequences provided strong support for the grouping of monotremes with marsupials (Janke et al. 1996; Janke et al. 1997). Ever since, the analysis of the mitochondrial protein coding genes has consistently supported the Marsupionta hypothesis. Also after the addition of several in- and outgroup species and the application of different methods of analysis, the relationship between marsupials and monotremes remains well supported (Janke et al. 1997; Cao et al. 1998; Zardoya & Meyer 1998; Kumazawa & Nishida 1999; Janke et al 2001, 2002). Some other molecular studies support the Marsupionta hypothesis as well: a DNA-DNA hybridization study (Kirsch & Mayer 1998) and a study of the gene for the dental protein amelogenin (Toyosawa et al. 1998). It seems an example 'par excellence' of the morphology versus molecules debate: morphology supports the grouping of marsupials and placentals, the Theria hypothesis, while molecules place marsupials and monotremes together, the Marsupionta hypothesis (Fig. 1).

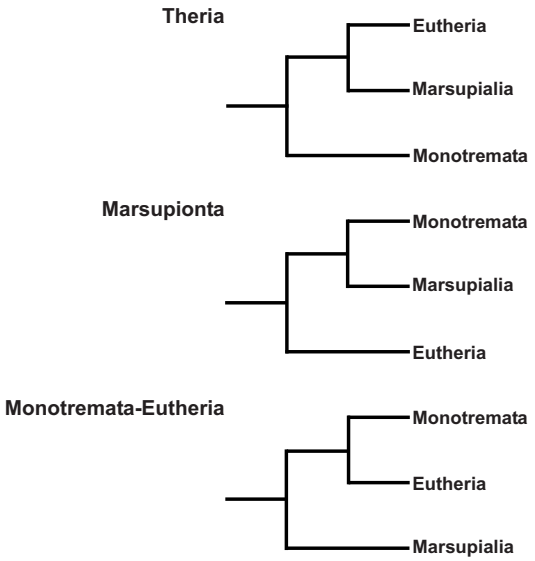


Figure 1 Alternative possibilities for the relationships between the three mammalian subclasses. Theria is supported by morphological evidence, and Marsupionta by mitochondrial protein-coding genes. A monotreme-eutherian clade has never been proposed.

Do molecules produce artefactual trees, or do derived characters of the monotremes frustrate meaningful morphological analysis? In the pursuit to shed light on this controversy, additional molecular data have been presented. Janke et al. (2002) extended the sampling of mitochondrial genomes by adding an echidna and an additional marsupial, the wombat, and also reported data from the nuclear 18S gene and an indel therein, again increasing support for the Marsupionta hypothesis. In contrast, the sequence of a single nuclear encoded gene, the large M6P/IGF2 receptor, provided support for the Theria hypothesis (Killian et al. 2001), but it was argued that this study needed further substantiation because IGF2R may have a different function in the distinct mammalian groups and also because the gene is imprinted in marsupials and eutherians, and not in monotremes and birds (Janke et al. 2002). Other molecular studies address the molecular evolution of nuclear genes along with the phylogenetic position of the monotremes (Retief et al. 1993; Kullander et al. 1997; Messer et al. 1998; Belov et al. 2002a,b,c; Miska et al. 2002; T. van Rheede, this thesis Chapter 4). These studies provide some support for the Theria hypothesis,

Table 1 Species, genes, numbers of nucleotides and accession numbers of sequences used in dataset I.

	chrm4 474 bp	*drd1a* 536 bp	*c-mos* 545 bp	*sox9* 437 bp	*adra2b* 797 bp
Mammalia					
Monotremata					
Platypus, *Ornithorhynchus anatinus*	new	new	new	new	new
Echidna, *Tachyglossus aculeatus*	new	new	new	new	new
Marsupialia					
Opossum, *Didelphis virginiana*		S67258			
Opossum, *Didelphis marsupialis*			new	new	Y15943
Kangaroo, *Macropus rufus*	new	new	new	new	AJ251183
Monito del monto, *Dromicops australis*	new	new	new	new	new
Shrew opossum, *Caenolestes fuliginosus*	new	new	new	new	new
Phascogale, *Phascogale tapoatafa*	new	new	new	new	new
Wombat, *Vombatus ursinus*	new	new	new	new	new
Eutheria					
Human, *Homo sapiens*	M16405	NM_000794	NM_005372	AI359981	M34041
Rat, *Rattus norvegicus*	M16409	NM_012546	X52952.1		M32061
Mouse, *Mus musculus*				AF421878	
Pig, *Sus scrofa*	new	U25681	X78318	AF006571	AJ251177
Seal, *Halichoerus grypus*	new		new		
Seal, *Phoca vitulina vitulina*		new			AJ251176
Bat, *Cynopterus sphinx*		new	new		AJ251180
Bat, *Macrotus californicus*	new				
Bat, *Emballonura atrata*				new	
Mole, *Talpa europaea*				new	
Shrew, *Sorex araneus*	new				
Shrew, *Sorex cinereus*		new	new		AJ315936
Manatee, *Trichechus manatus*	new	new	new	new	AJ251109
Aardvark, *Orycteropus afer*	new	new	new	new	Y12522
Anteater, *Cyclopus didactylus*	new				
Sloth, *Bradypus tridactylus*		new	new		AJ251179
Reptilia					
Squamata					
Gecko, *Eublepharis macularius*				AF217252	
Gecko, *Coleonys variegatus*			AF315386		
Gecko, *Lygodactylus picturatus*	new	new			new
Testudines					
Turtle, *Chelydra serpentina*	new	new		new	new
Turtle, *Podocnemis expansa*			AF109209		
Crocodylidae					
Caiman, *Caiman crocodilus*	new	new			new
Crocodile, *Crocodylus porosus*			AF039484		
Alligator, *Alligator mississippiensis*				AF106572	
Aves					
Neognatha					
Chicken, *Gallus gallus*	J05218	L36877	M19412.1	U12533	new

Table 2 Species, genes and accession numbers used in dataset II

	cryab	*ldha*	*eno1*	*hprt*	*rag1*	*bdnf*	*nt-3*	*ngfb*	*m6p/igf2r*	*mb*	*lalba*	*lyc*	*ins*
Eutheria													
Human, *Homo sapiens*	M28638	X02152	M14328	NM_000194	NM_000448	NM_001709	M37763	NM_002506	NM_000876	NP_005359	XP_012205	NP_000230	AAA59179
Rat, *Rattus norvegicus*	D29960	X01964											
Mouse, *Mus musculus*			BC003891	BC004686	NM_009019	NM_007540	X53257	BC011123	U04710	MYMS	AAA37208	NP_038618	NP032413
Marsupialia													
Kangaroo, *Macropus rufus*	new												
Opossum, *Didelphis virginia*													
Possum, *Trichosurus vulpecula*			new										
Opossum, *Monodelphis domestica*		AF070996			AF303968	U93379	U93372	U93373	AF342813	P02193	AAB97108	P51782	JQ0362
Wallaby, *Macropus eugenii*				AF072839									
Monotremata													
Platypus, *Ornithorhynchus anatinus*		AF545182		AF072842	AF303974	U93376	U93366	U93365	AF342814	P02196	P30805	P37156	Q9TQY7
Echidna, *Tachyglossus aculeatus*	new		new										
Aves													
Chicken, *Gallus gallus*	U26661	X53828	D37900	AJ132697	M58530	M83377	Z30092	X04003	GGU35037	P02197	CAA23711	CAA23711	P01332
Teleostei													
Flounder, *Paralichthys olivaceus*						AY074888							
Zebrafish, *Danio rerio*				BC046003				NM_131064	AL929530				
Gobiid, *Coryphopterus nicholsi*													
Carp, *Cyprinus carpio*		AF079534											
Paddlefish, *Polyodon spathula*							AF008559		U15613				

but generally suffer from the use of short sequences, limited taxon sampling, rooting with distant outgroups or paralogues, and it is not clear in all cases whether the genes are orthologous or have the same function in the different taxonomic groups.

Data from complete mitochondrial genomes are thus strongly at conflict with the generally held morphological view, but also with emerging nuclear data. To further investigate the phylogenetic position of the monotremes we obtained nucleotide sequence data from seven nuclear encoded genes. For five of these genes, the taxon sampling was increased to 17 mammalian ingroup and four outgroup taxa. The other two genes were combined with sequences from genes and proteins available in the databases, making a dataset of orthologous sequences for 9 genes and 12 proteins from four mammalian and one outgroup taxa. Thus, it is the first study applying data from nuclear encoded genes with different functions and a broad taxon sampling in both in- and outgroups.

Materials and Methods

Genes and Taxa

We obtained new sequence data from seven nuclear encoded genes. Two datasets were created, one with an extensive taxon sampling, but a shorter total length (dataset I: 21 taxa, 2793 bp; Table 1), and the other with a maximum length, but a more limited taxon sampling (dataset II: 5 taxa, 11544 bp, 4050 amino acids; Table 2). For dataset I, we sequenced segments of the genes coding for the acetylcholinergic receptor M4 (*chrm4*), proto-oncogene C-MOS (*c-mos*), dopamine receptor type 1A (*drd1a*), sex-determining transcription factor SOX9 (*sox9*), and α-2B adrenergic receptor (*adra2b*). The taxon sampling includes two monotremes, six marsupials from the orders Didelphimorpha, Paucituberculata, Microbiotheria, Dasyuromorphia and Diprotodontia (Graves and Westerman 2002), and nine Eutheria from the orders Primates, Rodentia, Cetartiodactyla, Carnivora, Chiroptera, Eulipotyphla, Sirenia, Tubulidentata and Xenarthra, representing the major superordinal eutherian clades (Murphy et al. 2001). As outgroup taxa we obtained sequences from a lizard, a turtle, a crocodile and a bird (see Table 1 for names). For dataset II, sequences are mainly from the databases, with the following taxon sampling: Primates (*Homo sapiens*), Rodentia (*Mus musculus* or *Rattus norvegicus*), a representative of marsupials, a monotreme, and as outgroup the chicken (*Gallus gallus*). We could retrieve suitable nucleotide sequences for the following genes: *ldha*, *hprt*, *bdnf*, *nt-3*, *ngfb*, *m6p/igf2r* receptor and *rag1*, and amino acid sequences for insulin (*ins*), myoglobin (*mb*), alpha lactalbumin (*lalba*) and lysozyme (*lyc*). Other monotreme sequences were found in the databases, but excluded from our analysis because of a known or suspected evolutionary history of gene conversion, duplication or concerted evolution, i.e., α- and β-hemoglobin (Lee et al. 1999), amelogenin (Toyosawa et al. 1998), olfactory receptor (Glusman et al. 2000), immunoglobulins (Belov et al. 2002a) and MHC (Miska et al. 2002), so that orthology of these genes can not be established. In addition, dataset II includes newly determined sequences for αB-crystallin (*cryab*) and α-enolase (*eno1*) genes for a marsupial and a monotreme. Exact species names and accession numbers of all sequences used in this study are presented in Tables 1 and 2.

PCR Amplication and Sequencing

Amplification of segments of 500-1200 bp was performed on the genes coding for *chrm4*, *c-mos*, *drd1a*, *sox9*, *adra2b*, *eno1* and *cryab*. Primers (Table 3) were based on alignments of known human, rat/mouse and chicken sequences, and additional tetrapod sequences when available. All PCR reactions were performed with a polymerase mix (Expand HF system, Roche) and contained approximately 50-100 ng genomic DNA or ~10 ng cDNA (reverse trancribed with Superscript II RT, Invitrogen). PCR reactions typically were in 50 µl, with 20-100 pmol of primers. Gel-extracted PCR fragments (Amersham Pharmacia GFX PCR gel extraction kit) were sequenced directly using Big Dye fluorescent technology on an ABI 3700 96-capillary sequencer.

Table 3 Primers used in amplification of segments of the *chrm4, c-mos, drd1a, adra2b, cryab* and *sox9* genes

Primer	Sequence (5' – 3')
chrm4for	CCT GTA GTC ATC ATG ACG GTN YTN TAY AT
chrm4rev	GTC ACC TTC TTC TCC CGN GCN GCC ATY T
chrm4seqfor	CCT GTA GTC ATC ATG ACG G
chrm4seqrev	GTC ACC TTC TTC TCC CG
cmosfor	CGG CTG GCC TGG TGC TYA ATV GAY TGG
cmosrev1	CTG GGA GCA RCC GAA GTC NCT DAT YT
cmosrev2	GAG GGT GAT GGC AAA GGA GTA GAT RTC NGC YT
drd1afor	TTC CAG TAT GAG AGA AAR ATG ACN CCN AA
drd1arev	GCA TGA GGG ATC AGR TAA ACN ARR TTR CA
sox9for2	CCC CAA CGC CAC CCA CMA CNC CNA ARA
sox9rev2	CTG TGT GTA GAC MGG CTG TTC CCA RTG YTG
A2ABFOR	ASC CCT ACT CNG TGC AGG CNA CNG C
A2AB52	GCA RGT AVA CNA GRA TCA TG
A2AB32	ATC ATG ATY CTN GTB TAC YTG C
A2ABREV	CTG TTG CAG TAG CCD ATC CAR AAR AAR AAY TG
αB1F	GGA YAT CRC CAT HCA YMA YCC
αB2R	GCG CTC YTC RTG YTT NCC RTG
αB2F	TCT GTM AAY CTK GAY GTR AAR CAY TT
αB3R2	CAG CAG GCT TCT CTT CAC GNG TDA TNG G

Additional gene- and species-specific primers were used in some instances (cryab, eno1, ldha). Details can be obtained from the authors.

Phylogenetic Analysis

Sequence data were assembled using the STADEN package programs PreGAP4 and GAP4 (http://www.mrc-lmb.cam.ac.uk/pubseq/). Nucleotide and amino acid alignments were produced using ClustalW, and adjusted manually. All alignments were inspected by eye for indels that could provide support for either Theria, Marsupionta or a Monotremata-Eutheria clade. For phylogenetic analysis, ambiguous positions in the alignment were excluded. Analysis of nucleotide data was performed using PAUP 4.0b10 (Swofford 2002) on the following datasets: all codon positions unweighted; first and second codon positions unweighted; first and second codon positions unweighted and third codon positions transversions only; and all codon positions transversions only. Phylogenetic criteria were: maximum parsimony, minimum evolution with LOGDET distances, and maximum likelihood (ML) with a model of sequence evolution selected using the LTR criterion in Modeltest 3.06 (Posada and Crandall 1998), except for the partitions with transversions only, where the CF+G_4+I model was selected manually as the best fitting model. In parsimony analyses stepwise addition with 100 random input orders of sequences were used and in ML analyses a neighbor joining (NJ) tree was used as starting tree. In all PAUP analyses, the tree bisection-reconnection branch swapping option was used to swap branches. ML and NJ distance analyses of amino acid sequences was performed using PROML, PROTDIST, NEIGHBOR, SEQBOOT and CONSENSE programs of the PHYLIP3.6a3 package (Felsenstein 2002) using the JTT model of sequence evolution (Jones et al. 1992). Bootstrap analyses included 1000 replicates for parsimony and minimum evolution analyses and 500 replicates for ML and NJ amino acid analyses. Bayesian analysis of both nucleotides and amino acids was performed with MRBAYES 2.1 (Huelsenbeck and Ronquist 2001) on the same data partitions and with the same models of sequence evolution as used in ML analyses, except for the amino acid data where a GTR+G_4+I model of sequence evolution was used. A Metropolis-coupled Markov chain Monte Carlo sampling approach was used to calculate posterior probabilities with initial equal probabilities for all trees, and starting trees were

random. Four Markov chains were run simultaneously two times for 500.000 generations, to check if stationary posterior probabilities had been reached. Tree sampling was done each 20 generations, and burn-in values were determined from the likelihood values.

Statistical Tests

Kishino and Hasegawa (1989; KH) and Shimodaira and Hasegawa (1999; SH) statistical tests were performed in PAUP4.0 with RELL optimization and 1000 bootstrap replicates to evaluate the a priori hypotheses about monotreme relationships: the Theria, Marsupionta and Monotremata-Eutheria hypotheses. For each hypothesis the best maximum likelihood tree and likelihood score were calculated and used for the statistical tests.

Time Estimation

Times of divergence were estimated for *Homo* versus *Mus*, marsupials versus eutherians, and monotremes versus therians. Two local clock methods were used: a Bayesian method (Divtime5b) (Kishino et al. 2001) and a maximum likelihood smoothing method (Penalized likelihood) (Sanderson 2003). The sequence alignments of the eight genes (*rag1, ngfb, eno1, ldha, nt3, m6p/igf2r, bdnf, hprt*), for which an actinopterygian fish sequence was available as outgroup (see Table 2), were concatenated for analysis. The amino acid concatenation (N=1807 sites) and nucleotide concatenation (N=5476 sites) were analyzed separately. The well-supported fossil divergence of the lineages leading to birds and mammals, 310 million years ago (Mya) (Hedges et al. 1996; Benton 2000) was used as the calibration point; no other reliable fossil calibration points were available for the datasets. For the Bayesian analysis, maximum likelihood branch lengths were calculated under F84 (Felsenstein 1984) and HKY (Hasegawa et al. 1985) models (nucleotides) and JTT (Jones et al. 1992) and gamma models (amino acids). The means of the prior distributions ("priors") for the rate parameter and the root time (rt and t, respectively) were calculated for each dataset. Divergence times "posteriors" and their 95% credibility intervals were recorded for each dataset. The penalized likelihood method was performed in r8s version 1.6 (Sanderson 2003) with maximum likelihood branch lengths calculated under a PC+gamma model (Yang 1997). A cross-validation procedure (Sanderson 2002) was used to obtain the optimal smoothing parameter for each dataset. Divergence times were recorded for each dataset.

Results

Phylogenetic Trees and Support for Theria and Marsupionta

To address the long-standing problem of the phylogenetic relationship between the three mammalian subclasses - monotremes, marsupials and placentals - we determined sequences from seven nuclear genes coding for proteins with widely different functions, like G-protein coupled receptors (*adra2b, chrm4, drd1a*), transcription factors (*sox9, cmos*), a structural protein/molecular chaperone (*crab*) and a metabolic enzyme (*eno1*). We thereby expanded the sequence data available for monotremes by 4021 nucleotides. For five of the genes (*sox9, cmos, adra2b, chrm4, drd1a*) sequences were obtained for 21 species (Table 1; dataset I: 2793 nucleotides, 931 aminoacids). A first inspection of dataset I revealed a general skew in mean base composition amongst the different mammalian subclasses and outgroup sequences, with monotremes being relatively GC rich (mean 64.4%) as compared to marsupials/outgroups (54.8% and 54.1 %, respectively) and placentals in between (59.8%) (Table 4; all positions). This GC bias is most pronounced at third codon positions, with GC contents ranging from 90.0% in monotremes to 68.5% in marsupials. Base composition can have major effects in phylogenetic analyses. We therefore made different data partitions to compensate for this base skew, such as deleting third codon positions, or coding third or all codon positions as purines and pyrimidines (RY coding) since skewness between purines and pyrimidines is minimal (Table 4).

Table 4 Mean base frequencies of dataset I

Codon position	All				1st				2nd				3rd			
	A/T	C/G	R	Y	A/T	C/G	R	Y	A/T	C/G	R	Y	A/T	C/G	R	Y
Eutheria	0.402	0.598	0.470	0.530	0.463	0.537	0.551	0.449	0.537	0.464	0.445	0.555	0.206	0.794	0.414	0.586
Marsupialia	0.452	0.548	0.472	0.528	0.492	0.509	0.553	0.447	0.549	0.451	0.461	0.539	0.315	0.685	0.401	0.599
Monotremata	0.356	0.644	0.468	0.532	0.447	0.553	0.559	0.441	0.521	0.479	0.454	0.546	0.100	0.900	0.389	0.611
Outgroups	0.459	0.541	0.496	0.504	0.530	0.470	0.598	0.402	0.564	0.436	0.463	0.537	0.283	0.717	0.425	0.575

Phylogenetic analysis of dataset I generally provided strong support for the sistergroup relationship of marsupials and placental mammals, using a range of methods and data partitions (Fig. 2; Table 5). However, some analyses, especially those that included transversions only, gave some support for joining Monotremata and Eutheria (up to 49.2%), a hypothesis that has never been proposed. In no case was there any meaningful support for Marsupionta, the highest bootstrap value being 10.8%. With regard to the other nodes in the tree (Fig. 2) it should be noted that within Eutheria, the species are correctly grouped into their respective basal clades: Afrotheria for aardvark and manatee, Euarchontoglires for human and murids, and Laurasiatheria for seal, bat, pig and eulipotyphlan insectivores. Moreover, Eurarchontoglires and Laurasiatheria are found as sistergroups, as is now well established (Murphy et al. 2001). Also the positioning of Xenarthra as sistergroup of Afrotheria, is in agreement with one of the three possible placements of the placental root (Delsuc et al. 2002). Within marsupials the rooting inside Australidelphia deviates from other recent molecular data (Amrine-Madsen et al. 2003). These authors found very strong support for a monophyletic Australidelphia with a dataset of 6363 bp. Our deviating rooting is most likely due either to long branch attraction (see the long branch to marsupials in Fig. 2), or to faulty species sampling, or both. Moreover, many sites that are informative within a mammalian subclass had to be scored as ambiguous, and thus removed from subsequent analyses, when distant outgroup sequences were added to the alignment. This leads to poor or erroneous resolution within that subclass. Also with regard to the outgroup taxa, the obtained topology (bird(caiman(gecko, turtle))) deviates from the currently most likely pattern of relationships (gecko(turtle(caiman, bird))). But reptilian relationships remain controversial, especially the position of the turtle (for review see Zardoya and Meyer, 2001).

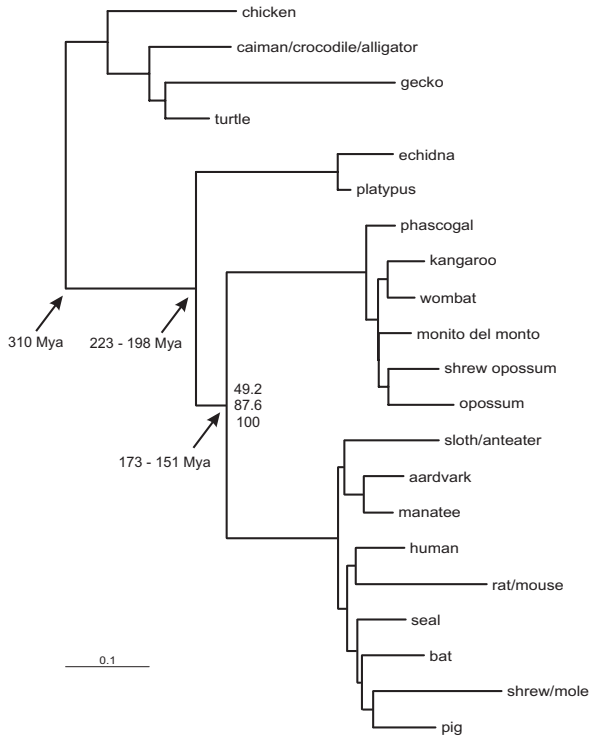


Figure 2 Maximum likelihood tree showing mammalian relationships as based on the concatenated sequences of dataset I. All codon positions were used, with a HKY+G4+I model of sequence evolution (ML partition 123 in Table 5; -ln L = 23630.56234). Chimeric sequence concatenations from different species are indicated by slashes.This tree is chosen from the various alternatives in Table 5 because it has the best agreement with established eutherian ordinal relationships (Murphy et al. 2001). Branch lengths are proportional to evolutionary distance (bar = 0.1 base substitution per site). Numbers at the eutherian node are (from top to bottom) the lowest, mean and highest support values for that clade (see also Table 5). The estimated divergence times for the three mammalian subclasses are given (Table 7) as well as the fossil-based calibration point of 310 Mya for the avian-mammalian divergence.

Table 5 Bootstrap support values and Bayesian posterior probabilities for Theria, Marsupionta and Monotremata-Eutheria. Analyses are: MP, maximum parsimony; ME/NJ, minimum evolution/ neighbor joining; ML, maximum likelihood. Models and partitions in *Materials and Methods*.

DATASET I

			Support		
Analyses	**'Model'**	**Partition**	**Theria**	**Marsupionta**	**Monotremata-Eutheria**
MP		12	76.5	5.2	18.3
		123	75	1.8	23.2
		12tv	87.6	3.3	9.1
		tv-only	50.9	7.9	41.2
		aa	89.1	5.7	5.2
ME/NJ	Logdet	12	99.5	0.2	0.3
	Logdet	123	55	0	45
	Logdet	12tv	97	0.1	2.9
	Logdet	tv-only	49.2	1.6	49.2
	JJT	aa	98.4	0.8	0.8
ML	GTR+Γ+I	12	95.4	2.6	2
	HKY+Γ+I	123	76.8	1.6	21.6
	GTR+Γ+I	12tv	94.8	2.2	3.0
	CF+Γ+I	tv-only	57.4	10.8	31.8
	JJT	aa	99.2	0.6	0.2
Bayes	GTR+Γ+I	12	1.00	0	0
	HKY+Γ+I	123	1.00	0	0
	GTR+Γ+I	12tv	1.00	0	0
	CF+Γ+I	tv-only	0.57	0.07	0.36
	GTR+Γ+I	aa	1.00	0	0
MP		12	100	0	0
		123	100	0	0
		12tv	100	0	0
		tv-only	100	0	0
		aa	100	0	0
ME/NJ	Logdet	12	100	0	0
	Logdet	123	100	0	0
	Logdet	12tv	100	0	0
	Logdet	tv-only	100	0	0
	JJT	aa	100	0	0
ML	HKY+Γ+I	12	100	0	0
	TrN+Γ+I	123	100	0	0
	HKY+Γ+I	12tv	100	0	0
	CF+Γ+I	tv-only	100	0	0
	JJT	aa	100	0	0
Bayes	HKY+Γ+I	12	1.00	0	0
	GTR+Γ+I	123	1.00	0	0
	HKY+Γ+I	12tv	1.00	0	0
	CF+Γ+I	tv-only	1.00	0	0
	GTR+Γ+I	aa	1.00	0	0

A much longer concatenation of sequences was obtained from a dataset II, comprising 9 genes and 4 proteins from five taxa: one monotreme, one marsupial, a rodent (mouse or rat), a primate (human), and chicken as outgroup (Table 2). This dataset was constructed using sequences from the database, and two newly sequenced genes (*cryab* and *eno1*), giving a concatenated length of 11544 nucleotides or 4050 amino acids. Although a deviation in base composition was also found in this dataset, phylogenetic analysis was straigthforward and unequivocal in its outcome: all types of analysis support Theria with a 100% bootstrap support (Table 5).

Statistical Tests of Prior Hypotheses

We calculated log-likelihood scores and used KH and SH tests to compare the three possible a priori hypotheses about monotreme relationships: the Theria, Marsupionta and Monotremata-Eutheria hypotheses (Table 6). With dataset II the Marsupionta and the Monotremata-Eutheria hypotheses could significantly be rejected with all data partitions (P = 0.000 to 0.029), whereas with dataset I the Marsupionta and Monotremata-Eutheria hypotheses in most cases could not be rejected. However, in all partitions Marsupionta is the most improbable outcome.

Indels

Insertions and deletions in protein-coding sequences have recently been used as markers to address several phylogenetic problems concerning for example relationships in teleost fish (Venkatesh et al. 2001), mammalian ordinal relationships (Poux et al. 2002), and the root of the placental mammalian tree (de Jong et al. 2003). The mutational events giving rise to insertions and deletions (indels) are qualitatively different from base substitutions. As such, they provide a source of phylogenetic

Table 6 Statistical results of log-likelihood scores and KH and SH tests to compare the three possible prior hypotheses about monotreme relationships (* statistically significant rejection)

DATASET I

Partition	Relationship	-ln L	Δ -ln L	P_{KH}	P_{SH}
12	Theria	9691.04365			
	Marsupionta	9699.66971	8.62606	0.109	0.065
	Monetreme-Eutheria	9699.93511	8.89146	0.101	0.052
123	Theria	23630.56234			
	Marsupionta	23640.09337	9.53103	0.067	0.034*
	Monetreme-Eutheria	23636.21457	5.65223	0.389	0.195
12tv	Theria	9691.04365			
	Marsupionta	9699.66971	8.62606	0.109	0.065
	Monetreme-Eutheria	9699.93511	8.89146	0.101	0.052
tv-only	Theria	9846.89978			
	Marsupionta	9848.83476	1.94822	0.425	0.211
	Monetreme-Eutheria	9847.75919	0.87266	0.763	0.379

DATASET II

Partition	Relationship	-ln L	Δ -ln L	P_{KH}	P_{SH}
12	Theria	24289.65533			
	Marsupionta	24325.68558	36.03024	0.001*	0.001*
	Monetreme-Eutheria	24326.68306	37.02773	0.000*	0.000*
123	Theria	47390.02351			
	Marsupionta	47432.33075	42.30724	0.000*	0.000*
	Monetreme-Eutheria	47429.94348	39.91998	0.000*	0.000*
12tv	Theria	24289.65533			
	Marsupionta	24325.68558	36.03024	0.001*	0.001*
	Monetreme-Eutheria	24326.68306	37.02773	0.000*	0.000*
tv-only	Theria	20979.47749			
	Marsupionta	20989.46735	9.98985	0.029*	0.017*
	Monetreme-Eutheria	20989.46735	9.98985	0.029*	0.017*

a)

Human	LTTLSSEPGQSQ-RTHIKTEQLS
Mouse	LTTLSSEPGQSQ-RTHIKTEQLS
Pig	LTTLSSEPGQSQ-RTHIKTEQLS
Mole	LTTLSSEPGQAQ-RTHIKTEQLS
Bat	LTTLSSEPGQSQ-RTHIKTEQLS
Aardvark	LTTLSSEPGQSQ-RTHIKTEQLS
Manatee	LTTLSSEPGQSQ-RTHIKTEQLS
Caenolestes	LTPLSSEQGQSQ-RTHIKTEQLS
Opossum	LTPLSSEQGQSQ-RTHIKTEQLS
Monito del m.	LTPLSSEQGQSQ-RTHIKTEQLS
Phascogale	LTPLSSEQGQSQ-RTHIKTEQLS
Kangaroo	LTPLSSEQGQSQ-RTHIKTEQLS
Wombat	LTPLSSEQGQSQ-RTHIKTEQLS
Echidna	LTTLSSEQGQAQQRTHIKTEQLS
Platypus	LTTLSGEQGQAQQRTHIKTEQLS
Snake	LTTLSSEPGQPQQRTHIKTEQLS
Tuatara	LTTLSSEPGQPQQRTHIKTEQLS
Turtle	MTTLSSEQGPSQQRTHIKTEQLS
Alligator	MTPLSGEQGQSQQRPHIKTEQLS
Chicken	LPALSGEQGPAQQRPHIKTEQLS
Rana	LPTLSNEQSQAQQRTHIKTEQLS
Xenopus	LSTLNSEQSQSQQRTHIKTEQLS

b)

Human	DRLVLSYVREEAG-KLDFCDGHSPAVTITFVCPS
Lemur	DRLVLSYVKEEGE-KLDFCDGHSPAVTITFVCPS
Flying lemur	DRLVLSYMKEEVG-QLDFCDGHSPAVTITFVCPS
Rabbit	DRLVLIYVKEGAE-KPDFCDGHSPAVTITFVCPS
Rat	DRLVLTYVKEEGE-KPDFCNGHSPAVTVTFVCPS
Mouse	DRLVLTYVKEEGE-KPDFCNGHSPAVTVTFVCPS
Bat	DRLVLTYGKEEAG-KPDFCEGHSPAVTITFVCPS
Hedgehog	DRLVLSYVKEDTR-KPDFCGGHSPAVTITFVCPS
Cow	DRLVLSYVKEGAG-QPDFCDGHSPAVTITFVCPS
Pig	DRLVLSYVNEEPA-PPDFCDGHSPAVTITFVCPS
Opossum	DRLVLSYVKEGED-RPAFCEGHTPAVTITFICPS
Wallaby	DRLVLSYVKQGED-QPTFCEGHTPAVTITFICPS
Echidna	DRLVLSYLKDDEESKPDFCNGHNPAVTITFICPS
Platypus	DRLVLSYLKDDEESKPDFCNGHNPAVTITFICPS
Chicken	DRLILSYERTYNDEKLNFCLGHNPAVTITFVCPS
Fish	DSLRLQYELSVESTPPERCGGHQPTVSITFICPS

c)

Human	ICQVKPNDQHFSRKVGTSDKTKYY
Lemur	ICQRKANDQHFSRKVGTSNQTRYY
Flying lemur	VCQVKSNNQQFSRKVGTSDKTKYY
Rabbit	VCQVKPNDQRFSRKVGTSERTKYY
Rat	ICQVKPNDQHFSRKVGTSDKTKYY
Mouse	ICQVKPNDQHFSRKVGTSDKTKYY
Bat	ICQVKPNDHRFSRKVGTSDKTKYY
Hedgehog	ICQVKPNDQHFSRKVGTSDMTKYY
Cow	VCQVKASDPRFGRKVGTLEKTRYY
Pig	ICQVKPNDQHFSRKVGTSDMTKYY
Opossum	ICQVKTNDQQFGRQVGSSDKTRYY
Wallaby	ICQIKTNDQQFGRQVGSSDKTRYY
Echidna	ICQVKTTDRYF-RKVGSSNTTKYY
Platypus	ICQVKTTDRYF-RKVGSSNTTKYY
Chicken	ICQVKTSERRF-RKIGWAKKAKYY

Figure 3 Support for Theria from deletions in the deduced protein sequences of sox9 (a) and m6p/igf2r (b, c). Sequences not shown in Table 1 or 2 are from the database. Grey shading shows majority consensus sequence with R=K, S=T, and D=E. Deletions are in black. Additional sox9 sequences are consistent with Theria: 5 primates (Patel et al. 2001), two bats, golden mole, hyrax, chicken and 3 reptiles (our unpublished data).

Table 7 Estimates of divergence time (Mya, ± standard error) among mammals from concatenations of nucleotide and amino acid sequences of eight nuclear protein-coding genes

Method*	Model	*Homo* vs. *Mus*	Marsupials vs. eutherians	Monotremes vs. therians
Nucleotides (N=5476)				
Divtime (Bayesian)	F84	86 ± 14	182 ± 21	235 ± 21
Divtime (Bayesian)	HKY	77 ± 14	174 ± 21	224 ± 21
Penalized likelihood	F84	85 ± 6	168 ± 11	216 ± 11
Penalized likelihood	HKY	85 ± 6	168 ± 11	216 ± 11
Summary (mean)		83 ± 10	173 ± 16	223 ± 16
Amino acids (N=1807)				
Divtime (Bayesian)	JTT	74 ± 14	155 ± 20	206 ± 20
Divtime (Bayesian)	Gamma	75 ± 15	157 ± 22	203 ± 22
Penalized likelihood	JTT	74 ± 5	147 ± 9	195 ± 9
Penalized likelihood	Gamma	72 ± 7	145 ± 12	189 ± 13
Summary (mean)		74 ± 10	151 ± 16	198 ± 16

* Priors for the Divtime analysis were: rttm = 450, lower limit = 288, upper limit = 345, rtrate (amino acid, JTT) = 0.0424, rtrate (amino acid, JTT+gamma) = 0.0521, rtrate (nucleotide, F84) = 0.000633, and rtrate (nucleotide, HKY) = 0.000875. For the Penalized Likelihood analysis, the calibration was set at 310, and the smoothing values were 0.16 (amino acid, JTT), 0.25 (amino acid, JTT+gamma), 10.0 (nucleotide, F84 and HKY); the standard deviation was derived from 100 bootstrap replicates. The gamma distribution was used with nucleotide and protein analyses.

information, independent of sequence data. We searched all available protein coding sequence data for the presence of indels that could be informative for resolving the trichotomy Monotremata-Marsupialia-Eutheria. We found three single amino acid indels that support the Theria hypothesis: one deletion in exon 3 of *sox9*, and one deletion and one insertion in the M6P/IGF2 receptor (Fig. 3). No indels supporting the two possible alternative relationships were observed.

Time Estimation

Divergence times estimated from the different methods showed some variation (Table 7), although individual estimates fell within 10% of the mean time for each divergence. However, times estimated from the nucleotide data were approximately 10% older than those estimated from the amino acid data. This would be expected if there were some sequence saturation present in the nucleotide data, because older divergences would have a greater proportion of hidden substitutions and would be compressed towards the calibration point. If true, the time estimates from the amino acid data set may be more reliable. In either case, the molecular clock analyses indicate that marsupials diverged from eutherians in the Middle or Late Jurassic (173-151 Mya) and that monotremes diverged from therians in the Late Triassic or Early Jurassic (223-198 Mya).

Discussion

Bias in base composition is known to influence phylogenetic reconstruction, particularly when the model of evolution implicit in the applied method does not fit the data. Some of our data may suffer from a bias in GC-content, monotremes being very GC-rich and marsupials being GC-poor. Our dataset I is indeed GC skewed (Table 4), as well as dataset II. However, the phylogenetic signal in dataset II is apparently so strong that it exclusively supports the Theria hypothesis, with the highest possible values. In the case of dataset I, the GC skew might contribute to the fact that certain analyses give some support to the grouping of Monotremata with Eutheria (Table 5). Since

Eutheria and Monotremata are the taxa with the highest GC-contents in dataset I, the recovery of a relationships between them might possibly be due to violation of the assumptions concerning base composition of the methods used. However, correcting for this skew using the logdet/paralinear distance, RY-coding at third codon positions, or analysis at the amino acid level did not increase support for Marsupionta. To the contrary, when correcting for GC-bias, the support for Theria generally increased, while at the same time the limited support for Monotremata-Eutheria vanished almost completely. This effect is not observed for tv-only because for nuclear genes, which evolve much slower than mitochondrial genes, this partition is too conservative and leads to loss of phylogentic information.

Our datasets I and II are large and very different in taxon and gene sampling. The congruent outcome of phylogenetic analyses of these two completely independent datasets is therefore convincing evidence for the Theria hypothesis, and against Marsupionta. In these phylogenetic analyses, those sequence regions were excluded which comprised gaps or where alignment was ambiguous. This means that the three indels that we detected in two unrelated genes (Fig. 3), grouping marsupials with placentals, provide a third source of completely independent support for the Theria hypothesis. No indels were found to support the alternative hypotheses.

While our results obtained from nuclear genes thus fully support the classical morphological Theria hypothesis (cf Figs. 1 and 2), they are in sharp contrast to the earlier mitochondrial data that strongly favor Marsupionta. How can data coming from the same organisms lead to such strongly conflicting phylogenies? Support for Marsupionta from the protein coding genes of mitochondrial genomes has been robust with different methods of analysis and increasing taxon sampling, as emphasized by Janke et al. (2002). However, it has been noted that a compositional bias is present in the mitogenomes of monotremes and some marsupials, which might contribute to the mitochondrial signal in support of Marsupionta (Phillips et al. 2001). Correcting for this bias by RY-coding along with partitioned maximum likelihood analysis indeed favored Theria over Marsupionta (Phillips & Penny 2003). This finding further extends the evidence that analyses based on mitochondrial protein-coding sequences can be misleading in deeper vertebrate phylogeny (Curole & Kocher 1999; García-Moreno et al. 2003).

As an independent source of evidence for Marsupionta, Janke et al. (2002) also presented an insertion that is present in the 3' region of the 18S rRNA gene of monotremes and marsupials. However, the indel boundaries in this gene are not located at precisely the same position in placental mammals and the outgroup species. This indel pattern can be explained by a single insertion in a common ancestor of marsupials and monotremes (thus supporting Marsupionta) plus a deletion or insertion in either outgroup or placentals (to explain the different boundaries), but equally well by two independent deletions in both outgroup species and placentals (which supports neither Theria, nor Marsupionta). Both scenarios require two steps and thus are equally parsimonious, not able to discriminate between Theria and Marsupionta. On the other hand, a highly conserved indel in the tRNA-Serine (UCN), between the acceptor and D arms, rejects Marsupionta and supports Theria (Phillips and Penny 2003). It may be noticed that indels in non-coding DNA, such as the genes for tRNA and rRNA, are generally less constrained than indels in protein coding genes, as the latter must leave the reading frame intact. However, also indels in protein sequences are certainly not free of homoplasy (de Jong et al. 2003). It is the finding of three independent deletion events in the *sox9* and *m6p/igf2* genes (Fig. 3) that makes the support for Theria compelling.

Although the branching order of the three mammalian subclasses is now well established (Fig. 2), the times of divergence of the monotremes and the subsequent split between marsupials and eutherians remain a matter of further investigation. For the marsupial-placental divergence, the paleontological estimates have ranged from ca. 125 to 97 Mya, but the most recent fossil evidence would bring this split back to at least 167 Mya (Woodburne et al. 2003). While mitochondrial

sequences date the marsupial-eutherian split at 170-140 Mya (Phillips and Penny 2003), nuclear genes have dated it at 190-182 Mya (Woodburne et al. 2003). However, based on our nuclear datasets, the marsupial-eutherian divergence occurred 167-153 Mya (Table 7), more in agreement with the mitochondrial dating. Interestingly, also our dating of the human-mouse divergence at 83-74 Mya (Table 7) is somewhat more recent than the 88-85 Mya derived from other large nuclear data sets (Springer et al. 2003).

Much less evidence is available for dating the monotreme-therian divergence. Limited paleontological data and time estimates calculated from mitochondrial DNA are in agreement with a monotreme-therian divergence of about 180-160 Mya (Phillips and Penny 2003). The dating studies based on our datasets yield considerably older divergence times for the monotreme-therian divergence, around 222-199 Mya (Table 7). Taken together, the molecular estimates for the separation of marsupials and eutherians thus range from 190-140 Mya, and for the monotreme-therian divergence from 222-160 Mya. This leaves as yet a wide time window of 20-82 million years between the two successive divergences that led to the three mammalian subclasses.

In conclusion, it can be stated that there is no longer any doubt about the position of monotremes in the mammalian tree. Morphological and paleontological data (e.g., Woodburne et al. 2003) as well as nuclear genes and indel data unambiguously support the classical Theria hypothesis. It only is the evidence from mitochondrial protein coding genes that gives robust support for Marsupionta. But this support vanishes when base compositional bias is taken into account (Phillips and Penny 2003). In the case of monotreme relationship it is no longer "morphology versus molecules", but mitochondrial proteins versus all other evidence.

Acknowledgements

This work was supported by grants from the Netherlands Organization for Scientific Research (NWO-ALW) and the European Commission.

References

Amrine-Madsen, H., Scally, M., Westerman, M., Stanhope, M.J., Krajewski, C. & Springer, M.S. (2003) Nuclear gene sequences provide evidence for the monophyly of australidelphian marsupials. *Mol. Phylogenet. Evol.* **28**: 186-196.

Belov, K., Hellman, L. & Cooper, D.W. (2002a) Characterization of immunoglobulin gamma 1 from a monotreme, Tachyglossus aculeatus. *Immunogenetics* **53**: 1065-1071.

Belov, K., Hellman, L. & Cooper, D.W. (2002b) Characterisation of echidna IgM provides insights into the time of divergence of extant mammals. *Dev. Comp. Immunol.* **26**: 831-839.

Belov, K., Zenger, K.R., Hellman, L. & Cooper, D.W. (2002c) Echidna IgA supports mammalian unity and traditional Therian relationship. *Mamm. Genome* **13**: 656-663.

Benton, M.J. (2000) *Vertebrate palaeontology*. Blackwell Science, Oxford.

Cao, Y., Waddell, P.J., Okada, N., & Hasegawa, M. (1998) The complete mitochondrial DNA sequence of the shark *Mustelus manazo*: evaluating rooting contradictions to living bony vertebrates. *Mol. Biol. Evol.* **15**: 1637-1646.

Curole, J.P. & Kocher, T.D. (1999) Mitogenomics: digging deeper with complete mitochondrial genomes. *Trends Ecol. Evol.* **14**: 394-398.

de Jong, W.W., van Dijk, M.A., Poux, C., Kappe, G., van Rheede, T. & Madsen, O. (2003) Indels in protein-coding sequences of Euarchontoglires constrain the rooting of the eutherian tree. *Mol. Phylogenet. Evol.* **28**: 328-340.

Delsuc, F., Scally, M., Madsen, O., Stanhope, M.J., de Jong, W.W., Catzeflis, F.M., Springer, M.S. & Douzery, E.J. (2002) Molecular phylogeny of living xenarthrans and the impact of character and taxon sampling on the placental tree rooting. *Mol. Biol. Evol.* **19**: 1656-1671.

Felsenstein, J. (1984) Distance methods for inferring phylogenies: A justification. *Evolution* **38**:16-24.

Felsenstein, J. (2002) PHYLIP (Phylogeny Inference Package) version 3.6a3. (Department of Genetics, University of Washington, Seattle).

Glusman, G., Bahar, A., Sharon, D., Pilpel, Y., White, J. & Lancet, D. (2000) The olfactory receptor gene superfamily: data mining, classification, and nomenclature. *Mamm. Genome.* **11**:1016-1023.

García-Moreno, J., Sorenson, M.D. & Mindell, D.P. (2003) Congruent avian phylogenies inferred from mitochondrial and nuclear DNA sequences. *J. Mol. Evol.* **57**:27-37.

Graves, J.A. & Westerman, M. (2002) Marsupial genetics and genomics. *Trends Genet.* **18**:517-521.

Hasegawa, M., Kishino, H. & Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**: 160-174.

Hedges, S.B., Parker, P.H., Sibley, C.G. & Kumar, S. (1996) Continental breakup and the ordinal diversification of birds and mammals. *Nature* **381**: 226-229.

Huelsenbeck, J.P. and Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754-755.

Janke, A., Gemmell, N.J., Feldmaier-Fuchs, G., von Haeseler, A. & Pääbo, S. (1996) The mitochondrial genome of a monotreme - the platypus (*Ornithorhynchus anatinus*). *J. Mol. Evol.* **42**:153-159.

Janke, A., Xu, X. & Arnason, U. (1997) The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupialia, and Eutheria. *Proc. Natl. Acad. Sci. USA* **94**: 1276-1281.

Janke, A., Erpenbeck, D., Nilsson, M. & Arnason, U. (2001) The mitochondrial genomes of the iguana (*Iguana iguana*) and the caiman (*Caiman crocodylus*): implications for amniote phylogeny. *Proc. R. Soc. Lond. B Biol. Sci.* **268**: 623-631.

Janke, A., Magnell, O., Wieczorek, G., Westerman, M. & Arnason, U. (2002) Phylogenetic analysis of 18S rRNA and the mitochondrial genomes of the wombat, *Vombatus ursinus*, and the spiny anteater, *Tachyglossus aculeatus*: increased support for the Marsupionta hypothesis. *J. Mol. Evol.* **54**: 71-80.

Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Cumput. Appl. Biosci.* **8**: 275-282.

Killian, J.K., Buckley, T.R., Stewart, N., Munday, B.L. & Jirtle, R.L. (2001) Marsupials and Eutherians reunited: genetic evidence for the Theria hypothesis of mammalian evolution. *Mamm. Genome.* **12**: 513-517.

Kirsch, J.A. & Mayer, G.C. (1998) The platypus is not a rodent: DNA hybridization, amniote phylogeny and the palimpsest theory. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **353**: 1221-1237.

Kishino, H. & Hasegawa, M. (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J. Mol. Evol.* **29**:170-179.

Kishino, H., Thorne, J.L. & Bruno, W.J. (2001) Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.* **18**:352-361.

Kullander, K., Carlson, B. & Hallbook, F. (1997) Molecular phylogeny and evolution of the neurotrophins from monotremes and marsupials. *J. Mol. Evol.* **45**:311-321.

Kumazawa, Y., Nishida, M. (1999) Complete mitochondrial DNA sequences of the green turtle and blue-tailed mole skink: statistical evidence for archosaurian affinity of turtles. *Mol. Biol. Evol.* **16**:784-792.

Lee, M.H., Shroff, R., Cooper, S.J. & Hope, R. (1999) Evolution and molecular characterization of a beta-globin gene from the Australian Echidna Tachyglossus aculeatus (Monotremata). *Mol. Phylogenet. Evol.* **12**:205-214.

Messer, M., Griffiths, M., Rismiller, P.D. & Shaw D.C. (1997) Lactose synthesis in a monotreme, the echidna (*Tachyglossus aculeatus*): isolation and amino acid sequence of echidna alpha-lactalbumin. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **118**:403-410.

Miska, K.B., Harrison, G.A., Hellman, L. & Miller, R.D. (2002) The major histocompatibility complex in monotremes: an analysis of the evolution of Mhc class I genes across all three mammalian subclasses. *Immunogenetics* **54**:381-393.

Murphy, W.J., Eizirik, E., O'Brien, S.J., Madsen, O., Scally, M., Douady, C.J., Teeling, E., Ryder, O.A., Stanhope, M.J., de Jong, W.W. & Springer, M.S. (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**:2348-2351.

Phillips, M.J. & Penny, D. (2003) The root of the mammalian tree inferred from whole mitochondrial genomes. *Mol. Phylogenet. Evol.* **28**:171-185.

Posada, D. and Crandall, K.A. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**:817-818

Poux, C., van Rheede, T., Madsen, O., & de Jong W.W. (2002) Sequence gaps join mice and men: phylogenetic evidence from deletions in two proteins. *Mol. Biol. Evol.* **19**:2035-2037.

Retief, J.D., Winkfein, R.J. & Dixon, G.H. (1993) Evolution of the monotremes. The sequences of the protamine P1 genes of platypus and echidna. *Eur. J. Biochem.* **218**:457-461.

Sanderson, M.J. (2002) Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.* **19**:101-109.

Sanderson, M.J. (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**:301-302.

Shimodaira, H & Hasegawa, M. (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**:1114-1116.

Springer, M.S., Murphy, W.J., Eizirik, E. & O'Brien, S.J. (2003) Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proc. Natl. Acad. Sci. USA*. **100**:1056-1061.

Swofford, D.L. (2003) PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.

Toyosawa, S., O'hUigin, C., Figueroa, F., Tichy, H. & Klein, J. (1998) Identification and characterization of amelogenin genes in monotremes, reptiles, and amphibians. *Proc. Natl. Acad. Sci. USA* **95**:13056-13061.

Venkatesh, B., Erdmann, M.V. & Brenner, S. (2001) Molecular synapomorphies resolve evolutionary relationships of extant jawed vertebrates. *Proc. Natl. Acad. Sci. USA.* **98**:11382-11387.

Woodburne, M.O., Rich, T.H. & Springer, M.S. (2003) The evolution of tribospheny and the antiquity of mammalian clades. *Mol. Phylogenet. Evol.* **28**:360-385.

Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* **13**:555-556.

Zardoya, R. & Meyer, A. (1998) Complete mitochondrial genome suggests diapsid affinities of turtles. *Proc. Natl. Acad. Sci. USA* **95**:14226-14231.

Zardoya, R. & Meyer, A. (2001) The evolutionary position of turtles revised. *Naturwissenschaften* **88**:193-200.

Chapter 3

Sequence Gaps Join Mice and Men: Phylogenetic Evidence from Deletions in Two Proteins

Celine Poux, Teun van Rheede*, Ole Madsen*, and Wilfried W. de Jong*‡*

*Department of Biochemistry, NCMLS, University of Nijmegen, The Netherlands;

‡Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, The Netherlands

Molecular Biology and Evolution Vol. 19(11), pp. 2035-2037, 2002

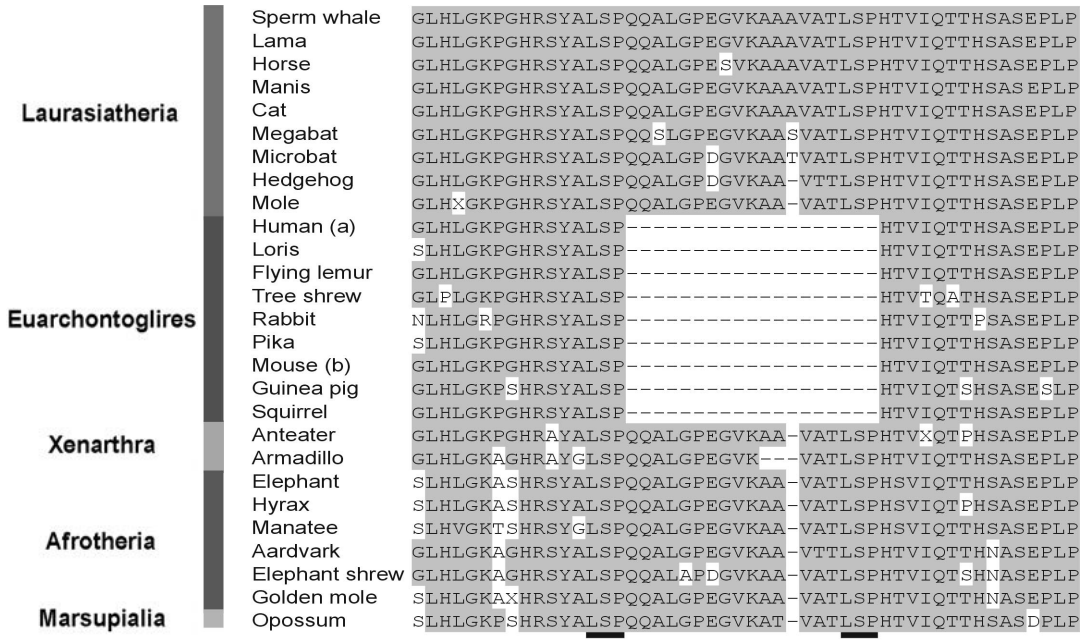


Figure 1 Deletions in the SCA1 protein (top) and the prion protein gene (bottom) support Euarchontoglires. Protein and DNA sequences, respectively, are shown as being most informative. Sequences correspond with positions 415 to 445 in the human SCA1 protein, and with nucleotides 1–44 of the coding sequence of the human *PRNP* gene. Eutherian species are grouped according to the four recently proposed basal clades of placental mammals (Murphy et al. 2000*b*). Gray shading emphasizes the overall sequence conservation; — denotes alignment gaps. The underlined Leu-Ser-Pro repeat in SCA1 is discussed in the text. Most sequences were newly determined by direct sequencing of PCR amplified genomic DNA fragments and can be found with full species names under accession numbers AJ438463–AJ438487 for *SCA1* and AJ438193–AJ438207 for *PRNP*. Human and mouse *SCA1* sequences are from the database (*a*, XM004164; *b*, NM009124), and *PRNP* sequences indicated with *c* from Wopfner et al. (1999).

Recent nuclear sequence analyses have provided evidence that primates and rodents are more closely related than previously believed (Madsen et al. 2001*a,* 2001*b*). This proposal is difficult to reconcile with morphological insights (Liu et al. 2001; Novacek 2001) and is not generally supported by current mitochondrial sequence data (Reyes, Pesole, and Saccone 2000; Nikaido et al. 2001; Arnason et al. 2002; Janke et al. 2002). Moreover, the supporting data and analyses have been criticized on methodological grounds (Rosenberg and Kumar 2001). Here we report deletions in two nuclear protein-coding genes that lend independent support to this contested grouping.

Some 18 orders of placental mammals are currently recognized, but their phylogenetic relationships remain highly controversial. Extensive sequence comparisons of mainly nuclear genes support a basal division into four major clades (Xenarthra, Afrotheria, Laurasiatheria, and Euarchontoglires), which has far-reaching implications for early mammalian biogeography and morphological diversification (Murphy et al. 2001*b*). Euarchontoglires is composed of the orders Primates, Rodentia, Lagomorpha (rabbits, hares, and pikas), Scandentia (tree shrews), and Dermoptera (flying lemurs). In contrast, morphology groups Primates, Scandentia, and Dermoptera with Chiroptera (bats) in the clade Archonta, whereas Rodentia and Lagomorpha (jointly called Glires) are in a distant clade with Macroscelidea (elephant shrews) (Liu et al. 2001; Novacek 2001). Also, sequence data from 12 proteins encoded by the mitochondrial genome generally do not support Euarchontoglires (e.g., Nikaido et al. 2001) or even maintain rodent polyphyly in many cases (Reyes, Pesole, and Saccone 2000; Arnason et al. 2002; Janke et al. 2002). Only by excluding some taxa with high or atypical substitution rates (or both) can sound mitochondrial support be obtained (Waddell, Kishino, and Ota 2001). Establishing the monophyly of the most speciose eutherian order, Rodentia, and finding its sister group has indeed been most difficult to solve on the basis of sequence evidence (e.g., Graur, Hide, and Li 1991; Adkins et al. 2001; Huchon et al. 2002). As for the molecular data sets giving support to Euarchontoglires, it has been questioned whether these are actually able to resolve the relationship of rodents and primates or whether more genes and longer sequences are needed (Rosenberg and Kumar 2001). Given, too, that Euarchontoglires is the least supported of the four major clades in some analyses (Madsen et al. 2001), additional evidence for their monophyly is certainly needed. This could be provided by "rare genomic changes," such as insertions and deletions (indels) in proteins (Rokas and Holland 2000). Indels in protein-coding DNA sequences require more complex mutational mechanisms and are generally more constrained than single base substitutions. Such indels can therefore be good indicators for monophyly, as demonstrated already for two of the other major clades, Xenarthra (van Dijk et al. 1999) and Afrotheria (Madsen et al. 2001), as well as in deeper vertebrate phylogeny (Venkatesh, Erdmann, and Brenner 2001).

While studying genes involved in various neurodegenerative disorders, we noticed two deletions that might be informative for the naturalness of Euarchontoglires. One is a large deletion in exon 8 of the gene for spinocerebellar ataxia 1 (*SCA1*), resulting in an 18-residue deletion in the encoded protein (Fig. 1, top). The other is a 6-bp deletion at the 5' end of the intronless coding region of the prion protein gene (*PRNP*; fig. 1, bottom). Both deletions perfectly distinguish Euarchontoglires from all other placentals and outgroup marsupials. Obviously, the most parsimonious interpretation is that these deletions originated once and independently in the *SCA1* and *PRNP* genes of the last common ancestor of Euarchontoglires, thus supporting their monophyly. If the morphological or mitogenomic trees are true, both deletions must have originated at least twice in exactly the same lineages.

Although reversal of the observed deletions in *SCA1* and *PRNP* is difficult to imagine, a repeated origin cannot totally be excluded. Indels are certainly not free from homoplasy, especially in regions with sequence repeats. In the *SCA1* gene, for example, a sequence repeat CTG TCN CCC, coding for Leu-Ser-Pro (underlined in Fig. 1, top), might in principle have triggered the large deletion

more than once. In the middle of this same region, a 6-bp deletion has caused the loss of two alanines in armadillo, whereas a 3-bp insertion results in an additional alanine in most Laurasiatheria (Fig. 1, top). This latter insertion might indeed agree nicely with a basal separation of Eulipotyphla (represented here by hedgehog and mole) from the other Laurasiatheria (Murphy et al. 2001*b*). However, both the deletion and the insertion are likely to be caused by the GCC (Ala) repeat in this gene region and therefore to have little phylogenetic significance. It is the congruence of independent evidence that makes the two deletions as shown in Figure 1 convincing indicators for the monophyly of Euarchontoglires. The probability of parallel origins of such deletions in two independent genes is difficult to evaluate statistically (van Dijk et al. 1999; Rokas and Holland 2000), but certainly it is extremely small. And even if these deletions were due to homoplasy, it would be a most curious coincidence that they occur in precisely the same species that are also grouped by independent sequence evidence (Madsen et al. 2001; Murphy et al. 2001*a,* 2001*b*).

Acknowledgments

This work was supported by grants from the Netherlands Organisation for Scientific Research and the European Comission.

References

Adkins, R. M., E. L. Gelke, D. Rowe, and R. L. Honeycutt. 2001. Molecular phylogeny and divergence times for major rodent groups: evidence from multiple genes. *Mol. Biol. Evol.* **18**:777–791.

Arnason, U., J. A. Adegoke, K. Bodin, E. W. Born, Y. B. Esa, A. Gullberg, M. Nilsson, R. V. Short, X. Xu, and A. Janke. 2002. Mammalian mitogenomic relationships and the root of the Eutherian tree. *Proc. Natl. Acad. Sci. USA* **99**:8151–8156.

Graur, D., W. A. Hide, and W.-H. Li. 1991. Is the guinea-pig a rodent? *Nature* **351**:649–652.

Huchon, D., O. Madsen, M. J. J. Sibbald, K. Ament, M. J. Stanhope, F. Catzeflis, W. W. de Jong, and E. J. P. Douzery. 2002. Rodent phylogeny and a timescale for the evolution of Glires: evidence from an extensive taxon sampling using three nuclear genes. *Mol. Biol. Evol.* (in press).

Janke, A., O. Magnell, G. Wieczorek, M.Westerman, and U. Arnason. 2002. Phylogenetic analysis of 18S rRNA and the mitochondrial genomes of the wombat, *Vombatus ursinus*, and the spiny anteater, *Tachyglossus aculeatus*: increased support for the Marsupionta hypothesis. *J. Mol. Evol.* **54**:71–80.

Liu, F. G., M. M. Miyamoto, N. P. Freire, P. Q. Ong, M. R. Tennant, T. S. Young, and K. F. Gugel. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. *Science* **291**:1786–1789.

Madsen, O., M. Scally, C. J. Douady, D. J. Kao, R. W. DeBry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. de Jong, and M. S. Springer. 2001. Parallel adaptive radiations in two major clades of placental mammals. *Nature* **409**:610–614.

Murphy, W. J., E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder, and S. J. O'Brien. 2001*a*. Molecular phylogenetics and the origins of placental mammals. *Nature* **409**:614–618.

Murphy, W. J., E. Eizirik, S. J. O'Brien et al. 2001*b*. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**:2348–2351.

Nikaido, M., K. Kawai, Y. Cao, M. Harada, S. Tomita, N. Okada, and M. Hasegawa. 2001. Maximum likelihood analysis of the complete mitochondrial genomes of eutherians and a reevaluation of the phylogeny of bats and insectivores. *J. Mol. Evol.* **53**:508–516.

Novacek, M. J. 2001. Mammalian phylogeny: genes and supertrees. *Curr. Biol.* **11**:R573–R575.

Reyes, A., G. Pesole, and C. Saccone. 2000. Long-branch attraction phenomenon and the impact of among-site rate variation on rodent phylogeny. Gene **259**:177–187. Rokas, A., and P. W. Holland. 2000. Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol.* **15**:454–459.

Rosenberg, M. S., and S. Kumar. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. *Proc. Natl. Acad. Sci. USA* **98**:10751–10756.

van Dijk, M. A. M., E. Paradis, F. Catzeflis, and W. W. de Jong. 1999. The virtues of gaps: xenarthran (edentate) monophyly supported by a unique deletion in alpha-A-crystallin. *Syst. Biol.* **48**:94–106.

Venkatesh, B., M. V. Erdmann, and S. Brenner. 2001. Molecular synapomorphies resolve evolutionary relationships of extant jawed vertebrates. *Proc. Natl. Acad. Sci. USA* **98**: 11382–11387.

Waddell, P. J., H. Kishino, and R. Ota. 2001. A phylogenetic foundation for comparative mammalian genomics. *Genome Informatics* **12**:141–154.

Wopfner, F., G. Weidenhofer, R. Schneider, A. von Brunn, S. Gilch, T. F. Schwarz, T. Werner, and H. M. Schätzl. 1999. Analysis of 27 mammalian and 9 avian PrPs reveals high conservation of flexible regions of the prion protein. *J. Mol. Biol.* **289**:1163–1178.

Chapter 4

The α-Crystallins of the Platypus *Ornithorhynchus anatinus*: Origin of the Alternatively Spliced Exon αA^{ins} and Implications for Mammalian Phylogeny.

Teun van Rheede and Wilfried W. de Jong

Department of Biochemistry, NCMLS, University of Nijmegen, The Netherlands

Abstract

α-Crystallins are abundant eye lens proteins that have been used extensively in early phylogenetic studies of amniotes. Since no α-crystallin sequences are available for monotremes as yet, and the relationships of the major mammalian groups remain debated, we set out to determine the αA- and αB-crystallin sequences from eye lens cDNA of a monotreme, the duck-billed platypus *Ornithorhynchus anatinus*. While both proteins were generally well conserved, the platypus αB-crystallin showed a tandem duplication of a heptapeptide FPTSFPA. This duplication was found to be further extended by an FPTFPA insert in αB-crystallin of the echidna *Tachyglossus aculeatus*, resulting in a six times repeated FP(A/T) motif. The phylogenetic signal in platypus and echidna α-crystallin sequences appeared to be insufficient to distinguish between the two competing hypotheses about monotreme relationships, which either place them at the base of the mammalian tree (the Theria hypothesis) or group them with marsupials (the Marsupionta hypothesis). Alternative splicing provides a mechanism to increase the diversity of gene products, but its role in the evolution of new lineages remains less well studied. An alternative splice variant of αA-crystallin occurs in several placental mammals. We demonstrated the mRNA expression of the alternatively spliced exon αA^{ins} in platypus, and that αA^{ins} is lacking in tegu (a lizard), caiman and duck, suggesting that αA^{ins}-crystallin originated in the last common ancestor of mammals.

(in preparation for publication)

Introduction

The vertebrate eye lens has a very high concentration of proteins, of which the majority derives from the families of α- and β/γ-crystallins. The β/γ-crystallins form a lens specific family of proteins, with 11 expressed genes in human (Lubsen et al. 1988). Two a-crystallins are present in the lens, αA- and αB-crystallin, which belong to the family of small heat shock proteins (sHsps) (de Jong et al. 1998). αB-crystallin actually occurs in heart and skeletal muscle as a molecular chaperone and in the lens as a structural protein, whereas αA-crystallin is restricted to the eye lens. Because of their chaperone-like properties, characteristic for sHsps, αA- and αB-crystallin could protect against the unfolding and aggregation of other lens proteins and thus help maintain the long term transparency of the lens (Horwitz 2003).

The αA- and αB-crystallins have in most mammals a length of 172 and 175 residues, respectively, and a sequence identity of about 55%. Their genes are located on different chromosomes (21 and 11, respectively) and have a similar organization, containing two introns. In placental mammals the first intron of the αA-crystallin gene generally contains an additional 69-bp exon, which is alternatively spliced in up to about 15% of the mature mRNA in some rodents (King and Piatigorsky 1983). As a consequence, 15% of the translation product of the αA-crystallin gene in these species has an insert of 23 amino acid residues, resulting in a 195-residue 'αAins-crystallin' chain. Even though the function of αA^{ins}-crystallin remains unknown (Hendriks et al. 1990; Smulders et al. 1995; van Dijk et al. 2001b), it does provide an example of the potential of alternative splicing to generate several transcripts and protein products from a single gene. This may well be one of the foremost evolutionary mechanisms to create new functions and complexity in specific lineages.

α-Crystallins are highly abundant proteins in the lens, making up one-third of the total lens protein in human. Because of this high abundance, αA-crystallins became one of the first polypeptides to be sequenced at the protein level, thus heralding the emerging field of molecular phylogenetics. Despite its relatively short length, αA-crystallin has been a successful protein in the study of species relationships. Several phylogenetic questions have been addressed by αA-crystallin protein sequences. Notably, the grouping of aardvark, and later on diverse other species like elephant shrews, golden moles and tenrecs with paenungulates (elephants, sea cows, hyraxes) in the superordinal clade Afrotheria was first proposed on the basis of αA-crystallin sequence data (de Jong et al. 1981; van Dijk et al. 2001a). This grouping has now rigourously been confirmed by some of the largest molecular datasets available, both from mitochondrial and nuclear encoded genes (Arnason et al. 2002; Murphy et al. 2001).

In amniotes, αA-crystallin sequences have been used extensively to study phylogenetic relationships in birds and placental mammals, and sequences are available for some marsupials and reptiles (Stapel et al. 1984; Caspers et al. 1994, 1996, 1997; Hedges et al. 1995). We set out to determine the α-crystallin sequences of a monotreme, the duck-billed platypus, *Ornithorhynchus anatinus*. Together with new data from other amniote species and the reptilian αA-crystallin intron I sequence, the platypus α-crystallin sequences turned out to contain only limited information regarding the phylogenetic position of the monotremes, but extend the insight in the origin of the alternatively spliced exon in the αA-crystallin gene.

Materials and Methods

Materials and Species Names

Eye lenses for RNA isolation were obtained from platypus (*O. anatinus*; van Rheede et al. 2003), a gecko (*Lygodactylus picturatus*; Werten et al. 2000) and Indian elephant (*Elephas maximus*), and stored in RNAlater (Ambion) (platypus) or frozen (gecko and elephant). Genomic DNA was obtained from tuatara (*Sphenodon punctatus*), tegu (*Tupinambis teguixin*), caiman (*Caiman crocodilus*), echidna (*Tachyglossus aculeatus*), opossum (*Didelphus marsupialis*), red kangaroo (*Macropus rufus*) and Indian elephant.

PCR Amplification, Cloning and Sequencing

Total RNA was isolated from eye lenses using Trizol reagent (Life technologies) and reverse transcribed to cDNA with superscript II reverse transcriptase (Invitrogen) (van Rheede et al. 2003). Most of the coding sequences of platypus, gecko and elephant α-crystallins were amplified from eye lens cDNA applying the primer combinations αA1F/αA3R2 or αB1F/αB3R2 (Table 1). Terminal sequences were obtained by 5'/3'-RACE (SMART RACE kit, Clontech), using degenerate primers along with specific primers. The tuatara αA-crystallin sequence was derived from genomic DNA using the primer pairs αA1F/αA2R and αA2F/αA3R2, spanning introns I and II, respectively. Similarly, the echidna, opossum and kangaroo αB-crystallin genes were amplified from genomic DNA using the primer pairs αB1F/αB2R and αB2F/αB3R2, spanning intron I and intron II, respectively. PCR-fragments comprising intron I of the αA-crystallin gene of tegu and caiman were obtained using the degenerate primers αA1F and αA2R. PCR products were either sequenced directly or cloned into pGEM-T easy (Promega) and subsequently sequenced using universal M13 primers. Direct sequencing was performed on an ABI 3700 96-cappilary sequencer, using Big Dye fluorescent technology. Sequences were submitted to Genbank under accession numbers: platypus αA, AJ617724; elephant αA, AJ617725; tuatara αA, AJ617726; gecko αA, AJ617727; platypus αB, AJ617728; echidna αB, AJ617729; elephant αB, AJ617732; kangaroo αB, AJ617731; opossum αB, AJ617730; tegu αA-crystallin intron I, AJ617733; caiman αA-crystallin intron I, AJ617734.

Table 1 Primers used for PCR amplification of αA- and αB-crystallin genes

Primer[1]	Sequence (5'-3')[2]
αA1F	AAC ACC CTT GGT TTA AAC GNG CNY TNG G
αA2R	GTG CTT GCC ATG GAT CTC CAC RAA RTC NT
αA2F	TGG AYG TGA AGC AYT TCT CHC CBG A
αA3R2	TTA GGA SGA GGG NGC MGA GST GGG CTT
αB1F	GGA YAT CRC CAT HCA YMA YCC
αB2R	GCG CTC YTC RTG YTT NCC RTG
αB2F	TCT GTM AAY CTK GAY GTR AAR CAY TT
αB3R2	CAG CAG GCT TCT CTT CAC GNG TDA TNG G

1 Number reflects exon of αA- or αB-crystallin gene on which primer is based; F, forward primer; R, reverse primer.
2 B: not A; H: not G; I: Inosine; K: G or T; M: A or C; N: A, G, T or C; R: A or G; S: C or G; Y: C or T.

Searching for αA^{ins} in Intron I Sequences

Initially, we attemped to detect the possible presence of the αA^{ins} optional exon in reptilian and avian αA-crystallin gene sequences by dotplot analysis (GCG package) against known mammalian intron I sequences (van Dijk et al. 2001b). Because dotplots can be relatively insensitive when short sequences are used, we also tried to detect the alternatively spliced exon by producing pairwise

alignments, using the ClustalW algorithm. A consensus sequence of the known αA^{ins} optional exons (van Dijk et al. 2001b) is expected to align with the most 'αA^{ins}-like sequence' in the intron I sequence. We produced alignments of the αA^{ins} consensus sequence with intron I sequences of tegu, caiman and duck, and as a control with several placental mammals selected to represent the major eutherian clades (Murphy et al. 2001).

Phylogenetic Analysis

DNA sequences were analyzed using the Staden package programs preGAP4 and GAP4 (http://www.mrc-lmb.cam.ac.uk/pubseq/). Nucleotide and amino acid alignments were produced using ClustalW. Phylogenetic analyses were performed with maximum likelihood (ML) using PHYML v2.0.2 (Guindon and Gascuel 2003) using a JJT+Γ_4 model of sequence evolution and with the alpha parameter of the gamma distribution estimated from the data. Bootstrap support was based on 100 replicates using the programs SEQBOOT and CONSENSE of the PHYLIP3.6a3 package (Felsenstein 2002) to generate data replicates and consensus tree, respectively.

Results and Discussion

αA- and αB-Crystallin Sequences of Monotremes

Agarose gel analysis of DNA fragments obtained from PCR on platypus lens cDNA, using the αA-crystallin primers αA1F and αA3R2, displayed two distinct bands. Both bands were sequenced and appeared to represent the messengers of αA-crystallin and its alternatively spliced form αA^{ins}-crystallin. The deduced amino acid sequence of platypus αA-crystallin is aligned in Figure 1 with those of other vertebrates, selected to represent the major classes and subclasses. The alignment includes the newly determined αA-crystallin sequences of elephant, gecko and tuatara. It is clear that the platypus and other novel αA-crystallin sequences display the high degree of sequence conservation known from other amniote α-crystallin proteins. The deduced insert sequence of platypus αA^{ins}-crystallin is aligned with all other known insert sequences in Figure 2, and is further discussed below.

The αB-crystallin coding sequence of the platypus was similarly obtained after PCR, using the primers αB1F and αB3R2 on cDNA. The deduced amino acid sequence is aligned in Figure 3 with other vertebrate αB-crystallin sequences, amongst which the newly determined ones of elephant, kangaroo and opossum. Most notable in platypus αB-crystallin is an insert of 8 residues within the exon I encoded region, resulting in a duplicated heptapeptide FPTSFPA. Considering this peculiar feature we also sequenced the largest part of the αB-crystallin gene from genomic DNA of another monotreme, the echidna *Tachyglossus aculeatus*. Most of the amino acid sequence could be deduced and showed a further insertion of a hexapeptide FPTFPA. These duplications apparently arose from replication slippages in a CT-rich repetitive sequence, resulting in a six times repeated FP(A/T) motif in echidna αB-crystallin.

The Phylogenetic Position of the Monotremes

αA-crystallin and to a lesser extent αB-crystallin have an excellent record as phylogenetic markers in amniotes. While the interrelationships of the major mammalian subclasses is still debated, no monotreme α-crystallin sequences had been determined yet. The morphology-based view places monotremes at the base of the mammalian tree, with marsupials and eutherians as each other's sistergroups (Lewis 1983). This notion, the Theria hypothesis, recently received support from nuclear sequences of the insulin like growth factor 2 receptor (IGF2R) gene (Killian et al. 2001). In contrast, a series of studies based on mitochondrial protein coding genes provides strong support for the alternative Marsupionta hypothesis, placing marsupials and monotremes as

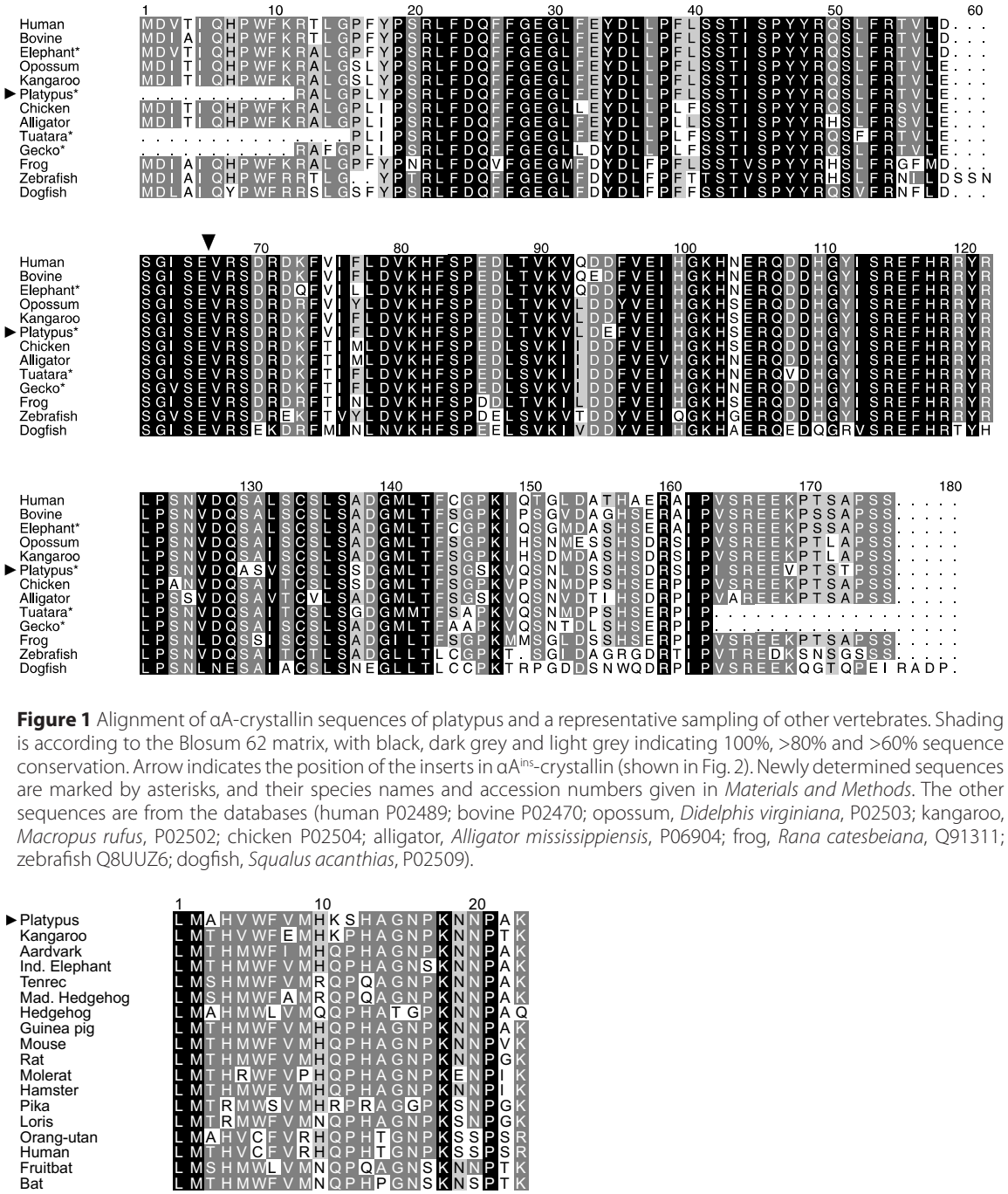


Figure 1 Alignment of αA-crystallin sequences of platypus and a representative sampling of other vertebrates. Shading is according to the Blosum 62 matrix, with black, dark grey and light grey indicating 100%, >80% and >60% sequence conservation. Arrow indicates the position of the inserts in αA^{ins}-crystallin (shown in Fig. 2). Newly determined sequences are marked by asterisks, and their species names and accession numbers given in *Materials and Methods*. The other sequences are from the databases (human P02489; bovine P02470; opossum, *Didelphis virginiana*, P02503; kangaroo, *Macropus rufus*, P02502; chicken P02504; alligator, *Alligator mississippiensis*, P06904; frog, *Rana catesbeiana*, Q91311; zebrafish Q8UUZ6; dogfish, *Squalus acanthias*, P02509).

Figure 2 The newly determined platypus αA^{ins}-crystallin insert sequence aligned with those from other mammals (as given in van Dijk et al. 2001b)

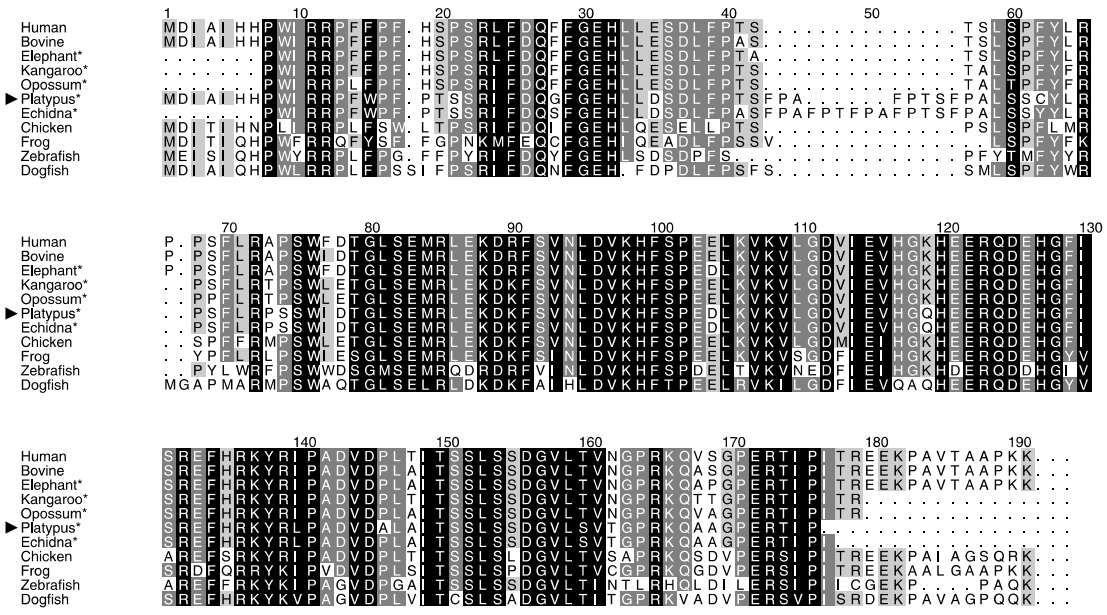


Figure 3 Alignment of αB-crystallin sequences of platypus and echidna with those from a sampling of other vertebrates. The FPx repeats in monotreme αB-crystallin cause the gap between position 42 and 57. Shading as in Figure 1. Newly determined sequences are marked by asterisks, and their species names and accession numbers given in *Materials and Methods*. The other sequences are from the databases (human P02511; bovine P02510; chicken Q05713; frog, *Rana catesbeiana*, Q91312; zebrafish NP_571232; dogfish, *Squalus acanthias,* P02512).

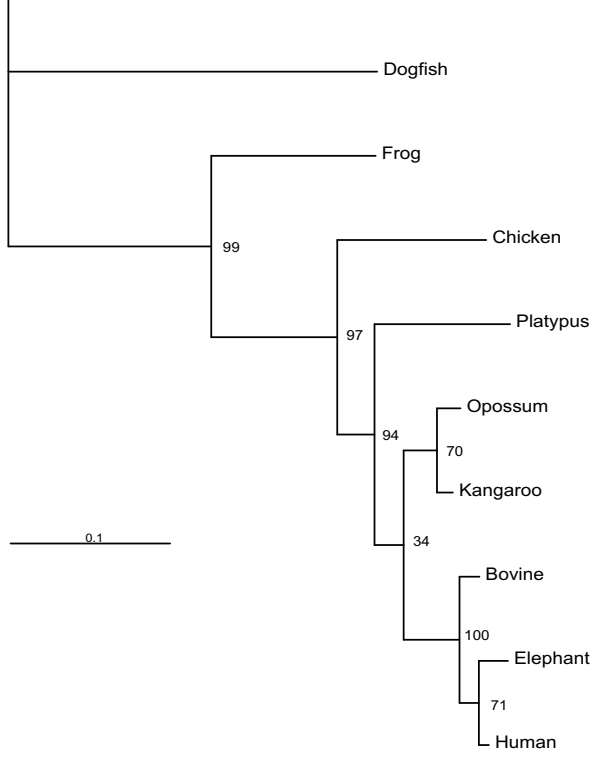


Figure 4 ML tree with bootstrap support values constructed from the concatenated amino acid sequences of αA- and αB-crystallin of platypus and 9 other vertebrates, as taken from Figures 1 and 3. Branch lengths are proportional to amino acid changes.

sistergroups to the exclusion of Eutheria (Janke et al. 1996, 1997, 2002). In view of this conflicting molecular evidence about the phylogenetic position of the monotremes, we performed phylogenetic analyses on the concatenated αA-and αB-crystallin protein sequences of platypus together with representatives of other amniote and outgroup sequences. The results were ambiguous, depending on the methods used, either placing platypus as oldest mammalian offshoot (in support of Theria) or as sister group of Eutheria, but always with weak support. Figure 4 shows a maximum likelihood tree with branch lenghts and bootstrap values based on the concatenated αA-and αB-crystallin protein alignment. As can be seen, support for Theria is insignificant. However, in no case was any support obtained for the grouping of monotremes with marsupials (Marsupionta).

Even though the separate αA- and αB-crystallin sequences are phylogenetically even less informative than the concatenated sequences, we made a tentative analysis of the position of the newly determined tuatura αA-crystallin sequence amongst the other reptiles (data not shown). It appeared that tuatara groups closest to gecko, in agreement with recent evidence from complete mitochondrial genome analyses that tuataras are the sister group of the lizards and snakes (Squamata) (Rest et al. 2003).

αA^{ins}-Crystallin in Non-Placentals

The optional αA^{ins}-exon, or remnants thereof, are present in intron I of the αA-crystallin gene in most of the studied placental mammals (van Dijk et al. 2001b). αA^{ins}-Crystallin has also been detected by western blotting in a marsupial, the tamar wallaby (*Macropus eugenii*), and by DNA hybridization in other marsupials and even in echidna (Wistow 1995). We now found that αA^{ins}-crystallin is also present, at the mRNA level, in platypus (Fig. 2). Like for αA-crystallin itself, the sequence of the insert peptide encoded by the alternatively spliced exon is well conserved in platypus, even though the rate of substitution in the optional exon in mammals appears to be higher than in αA-crystallin (van Dijk et al. 2001b).

To further assess the presence of αA^{ins} outside the mammals, we sequenced intron I of two reptiles, the tegu and caiman. Together with the available αA-crystallin gene sequence of the duck, we can now determine whether αA^{ins} occurs in the closest living relatives of mammals, that is, birds and reptiles. Dotplot analysis could not detect the alternatively spliced exon in these non-mammalian genes. We therefore adapted ClustalW to search for the αA^{ins} exon in the intron I sequences of these three genes. Pairwise alignments readily detected αA^{ins} exonic sequences in various mammalian αA-crystallin genes, as observed earlier (van Dijk et al. 2001b), while in tegu, caiman and duck no αA^{ins}-like sequences were observed. In addition, we could not detect αA^{ins}-crystallin at the cDNA level in gecko lenses (data not shown).

From this we conclude that the alternatively spliced exon coding for the additional 23 residues in αA^{ins}-crystallin is not present in birds and reptiles, or fell into disuse so long ago that it cannot be detected anymore in the species under study. In the latter case, this would be in contrast with a number of placental mammals that do not express αA^{ins}-crystallin, but still clearly show remnants of the αA^{ins} exon in their αA-crystallin genes, as for example in human (Jaworski and Piatigorsky 1989). Since αA^{ins}-crystallin could not be detected at the cDNA level in *Xenopus* either, and is absent in amphibian and teleost fish species in the database (data not shown), we conclude that αA^{ins}-crystallin originated in the common ancestor of mammals, after its divergence from the other amniotes, but before the divergence of monotremes and Theria.

An intriguing question is how the αA^{ins} exon came into existence. We addressed this question by a ψ-BLAST search of the sequence data in Genbank, using the αA^{ins} amino acid sequences in Figure 2 as query. ψ-BLAST is a sensitive, iterative search tool to detect weak similarities in molecular data (Jones and Swindells 2002). However, ψ-BLAST searches with the αA^{ins} sequence yielded no other sequences than the known αA^{ins} of mammals, even after repeated iterations. The apparent absence

of any αA^{ins}-like sequence in the Genbank database suggests that αA^{ins}-crystallin did not arise from exon shuffling or recruitment of an existing coding sequence into the αA-crystallin gene. αA^{ins}-crystallin thus may have originated from chance substitutions creating a donor and acceptor splice site in intron I of αA-crystallin, allowing an open reading frame with the exons flanking this intron. It is one of the numerous evolutionary enigmas how such a chance event could generate a viable gene product, αA^{ins}-crystallin, that has persisted next to the constitutive αA-crystallin in various mammalian lineages for more than 200 million years, without any apparent selective advantage (Smulders et al. 1995; van Dijk et al. 2001b).

Acknowledgements

We thank Niall Stewart (Hobart, Tasmania) for providing platypus eye lenses, Blair Hedges (Pennsylvania State University) for tuatara, tegu and caiman DNA, and Erik Franck and Ole Madsen for help with alignments and tree constructions. This work was supported by a grant from the Netherlands Organization for Scientific Research (NWO/ALW).

References

Arnason U., J.A. Adegoke, K. Bodin, E.W. Born, Y.B. Esa, A. Gullberg, M. Nilsson, R.V. Short, X. Xu, and A. Janke. 2002. Mammalian mitogenomic relationships and the root of the eutherian tree. *Proc. Natl Acad. Sci. USA* **99**:8151-8156.

Caspers, G.J., J. Wattel, and W.W. de Jong. 1994. αA-crystallin sequences group tinamou with ratites. *Mol. Biol. Evol.* **11**:711-713.

Caspers, G.J., G.J. Reinders, J.A. Leunissen, J. Wattel, and W.W. de Jong. 1996. Protein sequences indicate that turtles branched off from the amniote tree after mammals. *J. Mol. Evol.* **42**:580-586.

Caspers, G.J., D. Uit de Weerd, J. Wattel, and W.W. de Jong. 1997. α-Crystallin sequences support a galliform/anseriform clade. *Mol. Phylogenet. Evol.* **7**:185-188.

de Jong, W.W., A. Zweers, and M. Goodman. 1981. Relationship of aardvark to elephants, hyraxes and sea cows from α-crystallin sequences. *Nature* **292**: 538-540.

de Jong, W.W., G.J. Caspers, and J.A. Leunissen. 1998. Genealogy of the α-crystallin - small heat-shock protein superfamily. *Int. J. Biol. Macromol.* **22**:151-162.

Felsenstein J. 2002. PHYLIP (phylogeny inference package) version 3.6a3. Department of genetics, University of Washington, Seattle

Guindon, S. and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**:696-704.

Hedges, S.B., M.D. Simmons, M.A. Van Dijk, G.J. Caspers, W.W. de Jong, and C.G. Sibley. 1995. Phylogenetic relationships of the hoatzin, an enigmatic South American bird. *Proc. Natl. Acad. Sci. USA* **92**:11662-11665.

Hendriks, W., H. Weetink, C.E.M. Voorter, J. Sanders, H. Bloemendal, and W.W. de Jong. 1990. The alternative splicing product αA^{ins}-crystallin is structurally equivalent to αA and αB subunits in the rat α-crystallin aggregate. *Biochim. Biophys. Acta* **1037**:58-65.

Horwitz, J. 2003. α-Crystallin. *Exp. Eye Res.* **76**:145-53.

Janke, A., N.J. Gemmell, G. Feldmaier-Fuchs, A. von Haeseler, and S. Pääbo. 1996. The mitochondrial genome of a monotreme--the platypus (*Ornithorhynchus anatinus*). *J. Mol. Evol.* **42**:153-159.

Janke, A., X. Xu, and U. Arnason. 1997. The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupialia, and Eutheria. *Proc. Natl Acad. Sci. USA* **94**:1276-1281.

Janke, A., O. Magnell, G. Wieczorek, M. Westerman, and U. Arnason. 2002. Phylogenetic analysis of 18S rRNA and the mitochondrial genomes of the wombat, *Vombatus ursinus*, and the spiny anteater, *Tachyglossus aculeatus*: increased support for the Marsupionta hypothesis. *J. Mol. Evol.* **54**:71-80.

Jaworski, C.J. and J. Piatigorsky. 1989. A pseudo-exon in the functional human αA-crystallin gene. *Nature* **337**:752-754.

Jones, D.T. and M.B. Swindells. 2002. Getting the most from PSI-BLAST. *Trends Biochem. Sci.* **27**:161-164.

Killian, J.K., T.R. Buckley, N. Stewart, B.L. Munday, and R.L. Jirtle. 2001. Marsupials and Eutherians reunited: genetic evidence for the Theria hypothesis of mammalian evolution. Mamm. *Genome* **12**:513-517.

King, C.R. and J. Piatigorsky. 1983. Alternative RNA splicing of the murine αA-crystallin gene: protein-coding information within an intron. *Cell* **32**:707-712.

Lewis, O.J. 1983. The evolutionary emergence and refinement of the mammalian pattern of foot architecture. *J. Anat.* **137**:21-45.

Lubsen, N.H., H.J. Aarts, and J.G. Schoenmakers. 1988. The evolution of lenticular proteins: the β- and γ-crystallin supergene family. *Prog. Biophys. Mol. Biol.* **51**:47-76.

Murphy, W.J., E. Eizerik, S.J. O'Brien, O. Madsen, M. Scally, C. Douady, E. Teeling, O.A. Ryder, M. Stanhope, W.W. de Jong, and M.S. Springer. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**: 2348-2351.

Rest, J.S, J.C. Ast, C.C. Austin, P.J. Waddell, E.A. Tibbetts, J.M. Hay, and D.P. Mindell. 2003. Molecular systematics of primary reptilian lineages and the tuatara mitochondrial genome. *Mol. Phylogenet. Evol.* **29**:289-297.

Smulders, R.H., I.G. van Geel, W.L. Gerards, H. Bloemendal, and W.W. de Jong. 1995. Reduced chaperone-like activity of αA^{ins}-crystallin, an alternative splicing product containing a large insert peptide. *J. Biol. Chem.* **270**:13916-13924.

Stapel, S.O., J.A. Leunissen, M. Versteeg, J. Wattel, and W.W. de Jong. 1984. Ratites as oldest offshoot of avian stem - evidence from α-crystallin A sequences. *Nature* **311**:257-259.

van Dijk, M.A., O. Madsen, F. Catzeflis, M.J. Stanhope, W.W. de Jong, and M. Pagel. 2001a. Protein sequence signatures support the African clade of mammals. *Proc. Natl Acad. Sci. USA* **98**:188-193.

van Dijk, M.A., M.A. Sweers, and W.W. de Jong. 2001b. The evolution of an alternatively spliced exon in the αA-crystallin gene. *J. Mol. Evol.* **52**:510-515.

van Rheede, T., R. Amons, N. Stewart, and W.W. de Jong. 2003. Lactate dehydrogenase A as a highly abundant eye lens protein in platypus (*Ornithorhynchus anatinus*): upsilon-(υ)-crystallin. *Mol. Biol. Evol.* **20**: 994-998.

Werten, P.J.L., B. Röll, D.M.F. van Aalten, and W.W. de Jong. 2000. Gecko iota-crystallin: how cellular retinol-binding protein became an eye lens ultraviolet filter. *Proc. Natl Acad. Sci. USA* **97**:3282-3287.

Wistow G.J. (1995) Molecular biology and evolution of crystallins: gene recruitment and multifunctional proteins in the eye lens. Springer, New York.

Chapter 5

Lactate Dehydrogenase A as a Highly Abundant Eye Lens Protein in Platypus *(Ornithorhynchus anatinus)*: Upsilon (υ)-Crystallin

Teun van Rheede, Reinout Amons§, Niall Stewart‡ and Wilfried W. de Jong**

* Department of Biochemistry, NCMLS, University of Nijmegen, The Netherlands;

§ Leiden University Medical Center, Leiden, The Netherlands;

‡ School of Aquaculture, University of Tasmania, Hobart, Australia

Abstract

Vertebrate eye lenses mostly contain two abundant types of proteins, the α-crystallins and the β/γ-crystallins. In addition, certain housekeeping enzymes are highly expressed as crystallins in various taxa. We now observed an unusual approximately 41-kd protein that makes up 16% to 18% of the total protein in the platypus eye lens. Its cDNA sequence was determined, which identified the protein as muscle-type lactate dehydrogenase A (LDH-A). It is the first observation of LDH-A as a crystallin, and we designate it upsilon (υ)-crystallin. Interestingly, the related heart-type LDH-B occurs as an abundant lens protein, known as ε-crystallin, in many birds and crocodiles. Thus, two members of the *ldh* gene family have independently been recruited as crystallins in different higher vertebrate lineages, suggesting that they are particularly suited for this purpose in terms of gene regulatory or protein structural properties. To establish whether platypus LDH-A/ υ-crystallin has been under different selective constraints as compared with other vertebrate LDH-A sequences, we reconstructed the vertebrate *ldh-a* gene phylogeny. No conspicuous rate deviations or amino acid replacements were observed.

Molecular Biology and Evolution Vol. 20 (06), pp. 994-998, 2003

Introduction

The evolution of multicellular life forms, with specialized cell types and organs, brings the need for new functions and new building blocks. How has this problem to acquire new structures and functions been solved? The evolution of the eye provides a unique example in which housekeeping enzymes and stress proteins have been recruited for a role as structural proteins in lens and cornea, either by "gene sharing" (i.e., a gene acquiring a dual function) or after gene duplication (reviewed in Wistow 1995; Piatigorsky 1998).

The vertebrate eye lens contains large quantities of densely packed, water-soluble proteins. These proteins, aptly called crystallins, give the lens its refractive properties and long-term transparency. The α-crystallin and β/γ-crystallins are ubiquitous to the vertebrate eye lens and generally constitute the bulk of the lens protein. The α-crystallin belongs to the small heat-shock protein family, whereas the β/γ-crystallins essentially form a lens-specific family of proteins. In addition, some 10 types of taxonspecific crystallins are known to occur in various vertebrate lineages. These taxon-specific crystallins are related to or identical with common metabolic enzymes such as lactate dehydrogenase B, a-enolase, alcohol dehydrogenase, and aldose/aldehyde reductases. A single exception is iota (ι)-crystallin, which is a cellular retinolbinding protein (CRBP I), occurring in the lenses of some diurnal gecko species (Röll, Amons, and de Jong 1996; Werten et al. 2000).

The evolution of the vertebrate eye lens must have been accompanied by profound changes in gene expression and protein composition. The ubiquitous α-crystallin and β/γ-crystallins may have dominated the primordial lens, and adaptations to different visual environments were achieved by modulating their expression and recruiting additional genes in certain lineages. Such gene recruitment probably began as "gene sharing," by which a housekeeping protein, mostly enzymes, acquired an additional function as an abundant lens protein (Piatigorsky and Wistow 1991). The dual function may cause an "adaptive conflict" in which changes beneficial for the lens function may be deleterious for the housekeeping function outside the lens (Wistow 1993). This conflict can be solved when gene duplication occurs, allowing one copy to retain the housekeeping function and the other copy to further specialize for the lens function. Whereas gene sharing implies that the acquisition of a dual function may be strictly associated with changes in the regulation of such a gene (Piatigorsky 1998), gene duplication overcomes this restriction.

Although several examples of gene sharing and subsequent gene duplications have been described (Wistow 1995; Piatigorsky 1998), additional cases may contribute to better understanding this evolutionary phenomenon. We therefore present here a novel eye lens crystallin, υ-crystallin, found in platypus. Its sequence is similar to that of muscle-type lactate dehydrogenase A (LDH-A). Whereas LDH-B is expressed as a lens crystallin in many birds and crocodiles (Stapel et al. 1985; Wistow, Mulders, and de Jong 1987), this is the first example of recruitment of LDH-A as a structural protein to the eye lens. It also is the first LDH/crystallin outside the reptile/bird lineage.

Materials and Methods

Platypus (*Ornithorhynchus anatinus*) lenses were obtained from a male animal from the north of Tasmania. The platypus suffered from ulcerative mycosis, and was captured and euthanized with permission of the Tasmanian Parks and Wildlife Service. To preserve RNA, lenses were stored in RNAlater (Ambion).

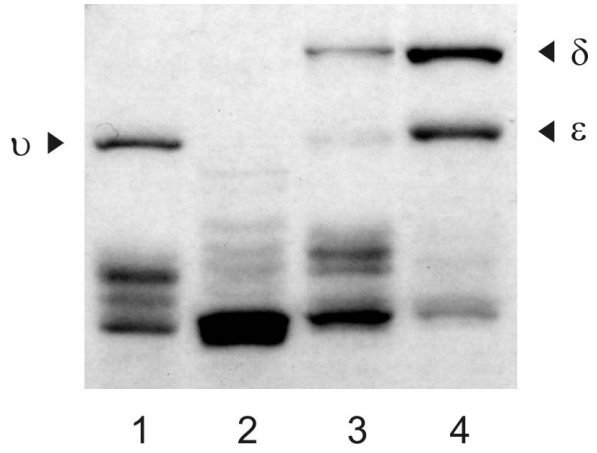


Figure 1 Analysis of lens extracts by SDS-gel electrophoresis and CBB-staining. Lane 1, platypus (*Ornithorhynchus anatinus*); lane 2, mouse (*Mus musculus*); lane 3, alligator (*Alligator mississippiensis*); lane 4, gannet (*Sula bassana*). All lanes were loaded with 6 mg of total lens protein. In addition to the novel υ-crystallin, the archosaurian δ-crystallin and ε-crystallin are indicated. The identification of avian and crocodile ε-crystallin as LDH-B, on basis of sequence analysis and enzymatic acitvity, has earlier been reported (Stapel et al. 1985; Wistow, Mulders, and de Jong 1987; Chiou et al. 1991).

SDS-PAGE and Protein Sequencing

Lens proteins were isolated with Trizol reagent (Life technologies), allowing simultaneous isolation of RNA (see below), and analyzed by SDS-PAGE. Gel pieces containing the protein of interest, with a mass of approximately 41 kd, were cut into small cubes and washed several times with water. The protein was digested in-gel with endoproteinase Lys-C (Boehringer) in 0.2 M Tris-HCl, pH 8.6, for 16 h at 37ºC. The liquid covering the gel pieces was collected, and the gel pieces were gently shaken, first with 200 ml of water for 3 h at room temperature and subsequently with 200 ml 0.1% (v/v) tri- fluoroacetic acid and 20% (v/v) acetonitrile, also for 3 h at room temperature, to elute the peptides. The extracts obtained were pooled and concentrated by lyophilization. The lyophilizate was dissolved in a small volume of 0.1% (v/v) trifluoroacetic acid and loaded onto a narrow bore (2.2×250 mm) Vydac C4 HPLC column. Chromatography was performed with a 1 h gradient from 0.1% (v/v) trifluoroacetic acid to 0.08% (v/v) trifluoroacetic acid and 80% (v/v) acetonitrile at 0.25 ml/min at room temperature. The peak fractions were collected and subjected to Edman degradation on a Hewlett Packard G1005A instrument, connected on-line with a Hewlett Packard Model 1100 HPLC. Only a few of the fractions contained peptides that could be analyzed. In one peak, the sequence SADTLWGIq was found. From another chromatographic run, two peaks were analyzed. One clearly contained a mixture of two peptides. Its analysis was: NS; AL; DH; TP; DL; L; G; T; D; A; -; -; -. The second peak gave again SADTLwGIq; -; -; -. Subtracting the latter sequence from the former, being a mixture of two sequences, gave NLHPDLGTDA.

RT-PCR and Sequencing

RNA was isolated using Trizol reagent (Life Technologies) and was reverse transcribed with Superscript II reverse transcriptase (Invitrogen). PCR on cDNA and 3' RACE (SMART RACE Kit [Clontech]) was performed with degenerate primers based on the sequenced peptides and on alignments of LDH-A and LDH-B sequences available in the data bank (see next section): 30 forward: GTTGGIGCWGTTGGNATGGCYTG; 150 forward: ATTTTGACCTATGTGGCYTGGAARAT; 280 forward: ATGGTGAAGGGCATGTATGG; 320 forward: AAGAGTGCAGAYACCYTGTGG; 3UTR reverse: AGTGCGACATACCCAATCC (numbering reflects approximate primer position in the alignment). After initial sequencing, gene-specific primers were designed to be used in 5' RACE (SMART RACE Kit [Clontech]): 5race1: CAGCTTGTGGAGTGGATGCC; 5race2: AATCCAGATTGCAGCCGCTTCC; 5race3: TCATCAGCCAAATCCTTCATC. PCR products were sequenced directly on an ABI 3700 automated sequencer. The sequence was submitted to GenBank with the accession number AF545182.

Gene Tree Construction

The newly obtained platypus LDH-A sequence was aligned with LDH sequences of mammals and reptiles from GenBank, using ClustalW. Accession numbers of sequences used in the alignment and for phylogenetic reconstruction can mostly be found in Mannen and Li (1999). The other accession numbers are AF070996 and AF070997 (Grey opossum [*Monodelphis domestica*] LDH-A and LDH-B, respectively) and AF069771 (chicken [*Gallus gallus*] LDH-B). For LDH-A sequences we performed maximum-parsimony (MP), Neighbor-Joining with Kimura two-parameter distances (NJ-K2P), and maximum-likelihood (ML) analyses on data sets of first and second codon positions with transversions only on third positions. Third codon position were used as transversions only, because some taxa had a high GC contents at third positions (e.g., platypus), thus violating assumptions of equal rates across lineages. MP analysis was performed with 100 branch-swapping replicates, random input order of sequences and 1,000 bootstrap replicates. NJ analysis was performed with K2P distance, as described in a recent study on the molecular evolution of the *ldh* gene family in vertebrates by Li and Tsoi (2002). In ML analysis, we used the HKY85 model of sequence evolution with gamma distribution and invariable sites (HKY+G+I). These analyses were performed using PAUP* 4.0 (Swofford 2002), applying the tree bisection reconnection (TBR) branch-swapping option.

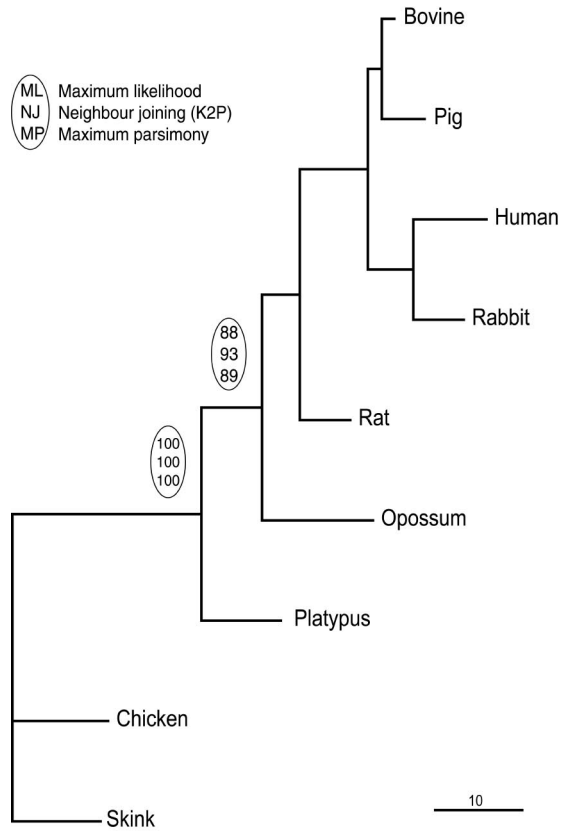
```
                    10         .        30         .        50         .        70         .        90         .       110
LDHA Platypus   MAGVKEQLIQNLLKEEY.APQNKITVVGVGAVGMACAISILMKDLADELALVDVIEDKLKGEMMDLQHGSLFLRTPKIVSGKDYSVTANSKLVIITAGARQQEGESRLNLVQRNV
     Opossum    -GT--D---L-V----Q.T-H-----------------------------------------------------K------S---A-------V-------------------
     Human      --TL-D---Y------Q.T-----------------------------------------------------------------N-------------------------
     Rat        --AL-D---V------Q.V-----------------------------------------------------K------S-------------------------------
     Pig        --TL-D---H------H.V-H---------------------E----I-----M-------------------------N-----R--V---------------------
     Skink      -.SL--K--EKVHT--HPHAH-------------------------------M--------L---------K---------A---H-----------------------
     Caiman     -.S---H--H-VH---HGHAH-------------------------------V----R---L-------------------------------------------------
     Chicken    -.SL-DH--H-VH---HAHAH---S-----------------------T----V--------L---------K----I---------H-----V---------------
LDHB Opossum    --TL--K--APVAE--AA--N---------Q----------G-S----------L----------------FQ-----AD-------G--I-VV---V---------------
     Human      --TL--K--APVAE--ATV-N---------Q----------G-S----------L----------------Q-----AD----------I-VV---V---------------
     Rat        --TL--K--APVADD-TAV-N---------Q----------G-S----------L----------------Q-----AD----------I-VV---V---------------
     Pig        --TL--K--APVAE--TTI-N---------Q----------G-S-T--------L----------------Q-----AD----------I-VV---V---------------
     Skink      --TLMDK--TSVAPPST.K-N--V------Q----------E-G-C--------V----------------N-H---AD----------V-VV---V---------D------
     Caiman     --TL--K--THIAS-ST.I-N--V------Q------V---G-S-C--------L----------------Q-H---AD---A------I-VV---V---------------
     Duck       --TL--K-MTPVAAASA.V-SS--------Q------V---G-G-C--------L----------------Q-H---AD---A------I-VV---V---------------
     Chicken    --TL--K--TPVAAGST.V-S---------Q----------G-G-C--------L----------------Q-H---AD---A------I-VV---V---------------
                  *     *     * *

                    .       130        .       150        .       170        .       190        .       210        .       230
LDHA Platypus   NIFKFIIPNVVKYSPNCKLLVVSNPVDILTYVAWKISGFPKNRVIGSGCNLDSARFRYLMGERLGIHSTSCHGWVIGEHGDSSVPVWSGVNVAGVSLKNLHPDLGTDADKEQWKD
     Opossum    ---------I-------------------------L-------------------------K-----S-----IL-----------------S---A----S-S-----
     Human      -------------------I-------------------------------------V-PL------L------------M--------T--------K------E
     Rat        -------------Q----I-------------------------------------V-PL------L----------------------S-N-Q-------------
     Pig        ---------I-------------------------------------------V-PL-----IL-----------------------E--------H--A
     Skink      -------------D-------------------------H-------------H-NPL-----IV----------------------S---EM-A-G---N---
     Caiman     ------------H--D-------------------------H-------------------PL-----IV----------------------A---E--------H--E
     Chicken    -------------D----I-------------------------H-------------H-------PL-----IV----------------------A----M-------H--E
LDHB Opossum    -V------QI-----D-TII-----------T--L--L--H-----------------S-K----PS-----IL--------A-----------QE-N-EM---N-S-N--E
     Human      -V------QI-----D-III-----------T--L--L--H-----------------A-K----PS-----IL--------A-----------QE-N-EM---N-S-N--E
     Rat        -V------QI-----D-TII-----------T--L--L--H-----------------A-K----PS-----IL--------A-----------QE-N-EM---N-S-N--E
     Pig        -V------QI-----D-III-----------T--L--L--H-----------------A-K--V-PS-----IL--------A-----------QE-N-EM---N-S-N--E
     Skink      -V------Q---H--D-II-----------T--L--L--H-------------F--A-K--V-P-----IL--------A-----------QEMN-AM-S-Q-P-S--Q
     Caiman     -V------QII----D-II-----------T--L--L--H-----------------S-K----PS-----IL--------A-----------QE-N-AM---R-S-K--E
     Duck       GV--G---QI-------TI-----------T--L--L--H----------T-------A------P-----IL--------A-----------QE-N-AM---K-S-N--E
     Chicken    -V------QI-------TI-----------T--L--L--H----------T-------A------P-----IL--------A-----------QE-N-AM---K-S-N--E
                                                                              *     *

                          .       250        .       270        .       290        .       310        .       330
LDHA Platypus   VHKQVVDSAYVIKLKGYTSWAIGLESVADLAESIVKNLRRVHPISTMIKGLYGIKDEVFLSVPCVLGQNGISDVVKITLKSEEEAHLKKSADTLWGIQKQKELQF...
     Opossum    ------E-------------------------M-----------------NED------I----------V--TT---SR--Q--------------...
     Human      ------E-------------------------M--------V----------D------I--------L--V--T-----R----------------...
     Rat        ---------------------------------M-----------------ED------I----------V--TPD---R----------------...
     Pig        ---------------------------------M-----------------ED------I----------V--TP--------------------...
     Skink      ------E-------------------------TVM--------V---V--MH----D----------YS--T----M---P---DK----------E-----...
     Caiman     ---E-------------------------TVM----------V--M-----D----------YH--T---MM-------EKIR-------------...
     Chicken    ---------------------------T-M----------AV--MH----D----------SS--T----MI--PD--EKI---------------...
LDHB Opossum    ---L-IE----------N----------I-TML---S-I--V---V--M---EN-----L--I-NAR-LTS-INQK--DD-V-Q--------D---D-KDL..
     Human      ---M--E----------N----------I--ML---S-I--V---V--M---EN-----L--I-NAR-LTS-INQK--DD-V-Q--------D---D-KDL..
     Rat        ---M-------------N----------I--ML---S-I--V---V--M---EN-----L--I-NAR-LTS-INQK--DD-V-Q-R------D---D-KDL..
     Pig        ---M--E----------N----------I--ML---S-I--V---VQ-M---EN-----L----NAR-LTS-INQK--DD-V-Q--N---------D-KDL..
     Skink      ------------------N----M------L-T-L---C----V---V--M---EN-----L-----SA-LTS-INQK--DN-V-Q-QN--T---NV--DIKDLKS
     Caiman     ------E----------N----F------L-TMM---N--Q-V---V--M---EN-----L----SAS-LTS-INQK--DD-V-Q-RS------S---D-KDL..
     Duck       ------E-----R----N---------E-C-TML---C---SV--LV--T---END----L----SAS-LTS-INQK--DD-V-Q--------S---D-KDM..
     Chicken    ------E-----R----N---------E-C-TML---Y---SV--LV--T---END----L----SAS-LTS-INQK--DD-V-Q--------S---D-KDL..
                                          *               *               *
```

Figure 2 Alignment of the platypus υ-crystallin sequence with LDH-A and LDH-B sequences from other amniotes. Platypus sequences obtained from peptide sequencing are boxed. (- - -) Indicates residues identical to the top sequence; (.) indicates an alignment gap. Asterisks indicate residues that are unique for platypus in this LDH-A data set.

Figure 3 Maximum-parsimony tree based on LDH-A amino acid sequences. Branch lengths correspond with numbers of amino acid replacements. The bar represents 10 replacements. Bootstrap support values for maximum-likelihood (HKY+G+I model), neighbor-joining (K2P distance), and maximum-parsimony analyses on the nucleotide data set are given (from top to bottom, respectively). We used skink and chicken LDH-A sequences as outgroups. For details of phylogenetic analyses, see *Materials and Methods.*



Results and Discussion

SDS gel electrophoresis of platypus lens extract revealed a conspicuous band in addition to the expected α-crystallin, β-crystallin, and γ-crystallin (Fig. 1). The protein made up 16% to 18% of the total lens protein and had an apparent molecular mass of approximately 41 kd. To identify the protein, we performed Edman degradation on peptides isolated after digestion with endoproteinase Lys- C. The sequences of two peptides could be determined (SADTLWGI and NLHPDLGTDA [boxed in Fig. 2]). A database search with these two sequences suggested that they originated from muscle-type LDH-A but could not completely exclude the possibility that it concerned hearttype LDH-B/epsilon (ε)-crystallin, as reported earlier in birds and crocodiles (Wistow, Mulders, and de Jong 1987). Degenerate PCR primers were therefore designed on basis of the sequenced peptides and of alignments of LDH-A and LDH-B. RT-PCR on platypus lens total RNA and subsequent sequencing yielded a complete coding sequence and 3' UTR. The derived amino acid sequence (Fig. 2) clearly identified it as LDH-A, rather than LDH-B. In fact, neither the Edman degradation nor the cDNA sequencing provided any evidence for the additional presence of LDH-B or any other LDH in the platypus lens. To distinguish this novel platypus lens protein from the archosaurian ε-crystallin, we designate it as upsilon (υ)-crystallin. The deduced amino acid sequence of υ-crystallin comprises 332 residues. It has a predicted molecular mass of 36.5 kd and a pI of 7.65.

Lens proteins require structural stability and the ability for close packaging. If platypus LDH-A/υ-crystallin had been under different selective constraints as compared with other vertebrate LDH-A sequences, this might be reflected by the presence of radical amino acid replacements. Nine substitutions are unique to platypus υ-crystallin, as compared with the other LDH-A sequences in Figure 2 (indicated by asterisks), but all these replacements are conservative in nature. A dual function, as an enzyme and as a lens protein, might also result in an accelerated rate of evolutionary change to adapt it to the additional function, as was found for CRBP-I/ι-crystallin (Werten et al. 2000). We therefore reconstructed the phylogeny of amniote LDH-A. From Figure 3, it appears that platypus LDH-A might have evolved somewhat slower rather than faster at the protein level than LDH-A in other mammals, suggesting that the protein is under increased selective constraints.

The related LDH-B/ε-crystallin is present at levels of up to 23% in lenses of many birds and crocodiles (Stapel et al. 1985; Wistow, Mulders, and de Jong 1987). Duck lens LDH-B/ε-crystallin is enzymatically active (Wistow, Mulders, and de Jong 1987) and is indeed the highly expressed product of the normal *ldh-b* gene (Hendriks et al. 1988). Also in the gecko genus Phelsuma, a 37-kd lens protein, present at low levels (<2%), reacts with the ε-crystallin antiserum and is associated with increased LDH activity (Röll and de Jong 1996). LDH-A, as we now found it in the platypus eye lens, is not present in echidna or other mammals (Stapel et al. 1985). This means that the recruitment of the *ldh-a* gene for expression as a lens crystallin has occurred in the platypus lineage after it diverged from the echidna lineage, estimated at around 34 MYA (Janke et al., 2002). Thus, while LDH-B is widely distributed as a lens protein in various avian and crocodilian lineages, probably having appeared and vanished repeatedly, LDH-A is as yet only found in the platypus lens.

The finding that the two paralogous *ldh-a* and *ldh-b* genes have been recruited in the eye lens in different amniote lineages raises the question of whether these genes or their products have properties that make them especially attractive for such a dual function. Recruitment of housekeeping genes in the eye lens may be a neutral process, in which the products of these genes can be used as structural proteins as long as requirements for transparency and stability are met. Some taxon-specific crystallins may also confer an evolutionary advantage to the eye lens in terms of adaptation to ecological requirements. The ability of several enzyme-crystallins to bind pyrimidine nucleotide cofactors may be adaptive because of their UV-absorbing capacity (Zigler and Rao 1991; Wistow 1995). Repeated, independent recruitment of crystallins from the same gene family, as here for the *ldh* genes, suggests that only certain gene families are suitable for this purpose. Repeated recruitement has also been observed for the zeta (ζ)-crystallins from the quinone oxidoreductase family in the eye lenses of guinea pig, camel, and Japanese tree frog (Gonzalez et al. 1995; Fujii et al. 2001) and for ϱA/ϱB-crystallins from the aldo-keto reductase superfamily, which were independently recruited in frogs (*Rana*) and geckos (*Phelsuma*) (van Boekel et al. 2001). Another example of different isoforms being recruited as lens crystallins are the δ-crystallins in birds and reptiles. Whereas the chicken lens expresses mainly δ1-crystallins and little δ2-crystallins, which are enzymatically inactive and active isoforms of the enzyme argininosuccinate lyase (ASL), respectively, the ratio δ1- rystallin to δ2-crystallin is more equal in the duck lens (Li, Wistow, and Piatigorsky 1995). δ-Crystallin/ASL is a convincing example of initial gene sharing in an ancient reptilian ancestor, with subsequent duplication of the asl gene allowing one copy (δ1) to adapt to the requirements as a lens protein, loosing its enzyme function, while the other copy (δ2) retained the enzymatic role (Wistow 1995; Piatigorsky 1998).

In the case of platypus LDH-A/υ-crystallin, we were unfortunately unable to obtain any data about enzymatic activity, expression in other tissues, or possible gene duplication because of lack of good quality material. However, considering the sequence similarities and position in the LDH-A tree (Figs. 2 and 3), it is reasonable to assume that platypus LDH-A/υ-crystallin respresents a case

of gene sharing, just as has been demonstrated for LDH-B/ε-crystallin. The recruitment of LDH-A and LDHB must have involved changes in their gene expression. Small changes in the promoter region can indeed have major effects on the expression level of genes. For example, adaptive variation in *ldh-b* gene expression between populations of the fish species Fundulus heteroclitus could be explained by a single mutation (Schulte et al. 2000). Also, the high expression of ε-crystallin in the duck lens did not require the evolution of a lens-specific promoter element (Brunekreef et al. 1996). In fact, the numerous studies on crystallin gene regulation demonstrate that their high expression in the lens is mainly the result of tissue-specific transcriptional activation, resulting from the complex interplay between lens-preferred factors, such as Pax-6, and general transcription factors (reviewed in Wistow 1995; Piatigorsky 1998).

It remains an open question whether high LDH levels in the lens are evolutionary neutral or adaptive and why the *ldh* genes are particularly suited for lens recruitment. From our data, we conclude that no major adaptive amino acid replacements are required to make LDH-A suitable for its dual function as a housekeeping enzyme and as a highly expressed lens protein. Also, duck LDH-B/ε-crystallin was actually found to be relatively unstable and to have a decreased affinity for its substrate, pyruvate (Berr et al. 2000). In any case, the *ldh* gene family has provided a versatile source of evolutionary variation. Duplication of the *ldh-a* gene led to the testis-specific LDH-C in mammals (Li and Tsoi 2002). LDH-B is expressed at levels of up to 5% in mouse oocytes (Whitt 1984) and is present as ε-crystallin in the eye lenses of many archosaurs, whereas LDH-A functions as υ-crystallin in the platypus lens.

Acknowledgments

We thank Sandor Boros for technical advice. This work was supported by a grant from the Netherlands Organization for Scientific Research (NWO-ALW).

References

Berr, K., D. Wassenberg, H. Lilie, J. Behlke, and R. Jaenicke. 2000. e-Crystallin from duck eye lens: comparison of its quaternary structure and stability with other lactate dehydrogenases and complex formation with α-crystallin. *Eur. J. Biochem.* **267**:5413–5420.

Brunekreef, G.A., H. J. Kraft, J. G. Schoenmakers, and N. H. Lubsen. 1996. The mechanism of recruitment of the lactate dehydrogenase-B/ε-crystallin gene by the duck lens. *J. Mol. Biol.* **262**:629–639.

Chiou, S. H., H. J. Lee, S. M. Huang, and G. G. Chang. 1991. Kinetic comparison of caiman epsilon-crystallin and authentic lactate dehydrogenases of vertebrates. *J. Protein Chem.* **10**:161–166.

Fujii Y., H. Kimoto, K. Ishikawa, K. Watanabe, Y. Yokota, N. Nakai, and A. Taketo. 2001. Taxon-specific zeta-crystallin in Japanese tree frog (*Hyla japonica*) lens. *J. Biol. Chem.* **276**:28134–28139.

Gonzalez, P., P. V. Rao, S. B. Nunez, and J. S. Zigler Jr. 1995. Evidence for independent recruitment of zetacrystallin/ quinone reductase (CRYZ) as a crystallin in camelids and hystricomorph rodents. *Mol. Biol. Evol.* **12**: 773–781.

Hendriks, W., J. W. Mulders, M. A. Bibby, C. Slingsby, H. Bloemendal, and W. W. de Jong. 1988. Duck lens ε-crystallin and lactate dehydrogenase B4 are identical: a single-copy gene product with two distinct functions. *Proc. Natl. Acad. Sci. USA* **85**:7114–7118.

Janke, A., O. Magnell, G. Wieczorek, M. Westerman, and U. Arnason. 2002. Phylogenetic analysis of 18S rRNA and the mitochondrial genomes of the wombat, *Vombatus ursinus*, and the spiny anteater, *Tachyglossus aculeatus*: increased support for the Marsupionta hypothesis. *J. Mol. Evol.* **54**: 71–80.

Li, Y. J., and S. C. Tsoi. 2002. Phylogenetic analysis of vertebrate lactate dehydrogenase (LDH) multigene families. *J. Mol. Evol.* **54**:614–624.

Li, X., G. J. Wistow, and J. Piatigorsky. 1995. Linkage and expression of the argininosuccinate lyase/δ-crystallin genes of the duck: insertion of a CR1 element in the intragenic spacer. *Biochim. Biophys. Acta* **1261**:25–34.

Mannen, H., and S. S. Li. 1999. Molecular evidence for a clade of turtles. *Mol. Phylogenet. Evol.* **13**:144–148.

Piatigorsky, J. 1998. Gene sharing in lens and cornea: facts and implications. P*rog. Retin. Eye Res.* **17**:145–174.

Piatigorsky, J. and G. Wistow. 1991. The recruitment of crystallins: new functions precede gene duplication. *Science* **252**: 1078–1079.

Röll, B., R. Amons, and W. W. de Jong. 1996. Vitamin A2 bound to cellular retinol-binding protein as ultraviolet filter in the eye lens of the gecko *Lygodactylus picturatus*. *J. Biol. Chem.* **271**:10437–10440.

Röll, B. and W. W. de Jong. 1996. First finding of ε-crystallin outside the archosaurian lineage. *Naturwissenschaften* **83**: 177–178.

Schulte, P. M., H. C. Glemet, A. A. Fiebig, and D. A. Powers. 2000. Adaptive variation in lactate dehydrogenase-B gene expression: role of a stress-responsive regulatory element. *Proc. Natl. Acad. Sci. USA* **97**:6597–6602.

Stapel, S. O., A. Zweers, H. J. Dodemont, J. H. Kan, and W. W. de Jong. 1985. ε-Crystallin, a novel avian and reptilian eye lens protein. *Eur. J. Biochem.* **147**:129–136.

Swofford, D.L. 2002. PAUP*: Phylogenetic analyses using parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland, Masschusetts.

van Boekel, M. A., D. M. van Aalten, G. J. Caspers, B. Röll, and W. W. de Jong. 2001. Evolution of the aldose reductase related gecko eye lens protein rhoB-crystallin: a sheep in wolf's clothing. *J. Mol. Evol.* **52**:239–248.

Werten, P. J., B. Röll, D. M. van Aalten, and W. W. de Jong. 2000. Gecko iota-crystallin: how cellular retinol-binding protein became an eye lens ultraviolet filter. *Proc. Natl. Acad. Sci. USA* **97**:3282–3287.

Whitt, G. S. 1984. Genetic, developmental and evolutionary aspects of the lactate dehydrogenase isozyme system. *Cell. Biochem. Funct.* **2**:134–139.

Wistow, G. 1993. Lens crystallins: gene recruitment and evolutionary dynamism. *Trends Biochem. Sci.* **18**:301–306.

Wistow, G. 1995. Molecular biology and evolution of crystallins: gene recruitment and multifunctional proteins in the eye lens. R. G. Landes Company, Austin, Texas.

Wistow, G. J., J. W. Mulders, and W. W. de Jong. 1987. The enzyme lactate dehydrogenase as a structural protein in avian and crocodilian lenses. *Nature* **326**:622–624.

Zigler Jr, J. S., and P. V. Rao. 1991. Enzyme/crystallins and extremely high pyridine nucleotide levels in the eye lens. *FASEB J.* **5**:223–225

Chapter 6

Sequence and Functional Conservation of the Intergenic Region between the Head-to-Head Genes Encoding the Small Heat Shock Proteins αB-Crystallin and HspB2 in the Mammalian Lineage

Linda Doerwald[§], Teun van Rheede[§*], Ron P. Dirks[‡], Ole Madsen[§], Remco Rexwinkel[§], Siebe T. van Genesen[§], Gerard J. Martens[‡], Wilfried W. de Jong[§] and Nicolette H. Lubsen[§]*

*Both authors contributed equally;

[§] Department of Biochemistry, NCMLS, University of Nijmegen, The Netherlands;

[‡] Department of Animal Physiology, University of Nijmegen, The Netherlands.

Abstract

The small heat shock protein αB-crystallin is abundantly expressed in lens and muscle and in response to a heat shock, while the related HspB2 protein is abundant only in muscle and not upregulated by a heat shock. The genes for these two proteins are located head-to-head in man and rodents as well as in duck. The intergenic distance in duck is much larger (1.6 kb) than that in man or rodents (0.9 kb). We have examined the linkage of these genes in the major mammalian lineages and in chicken. The intergenic distance in mammals ranged from 645 bp (platypus) to 1069 bp (opossum), with an average of about 900 bp; that in chicken was the same as in duck. Phylogenetic footprinting and sequence alignment showed conservation of sequence elements close to the HspB2 promoter and identified two additional conserved regions further upstream. All known regulatory elements of the mouse αB-crystallin promoter are conserved, except in platypus and birds. The lens-specific-region 1 (LSR1) as well as the heat shock elements (HSE's) are missing in the avian intergenic region; platypus has only the Pax-6 site of LSR1 and lacks the Pax-6 site in LSR2 and one of the two HSE's. Our results argue that the primordial mammalian αB-crystallin promoter had two LSR's and two HSE's and that the loss of one of the Pax-6 sites and one of the HSE's in platypus is secondary.

In transfection experiments the platypus αB-crystallin promoter retained heat shock responsiveness and lens expression. It also directed lens expression in *Xenopus laevis* transgenes as did the HspB2 promoter of rat or blind mole rat. Deletion of the middle of the intergenic region including the upstream enhancer affected the activity of both the rat αB-crystallin and HspB2 promoters, suggesting that at least some elements within the complex enhancer region work towards the HspB2 promoter.

(submitted for publication)

Introduction

Bidirectional gene pairs, located so close that promoter regions overlap, are surprisingly common in eukaryotes. For example, Adachi and Lieber (2002) recently found that almost 30% of the housekeeping genes in man were located in a bidirectional fashion less than 1 kb apart. As eukaryotic genomes are large relative to the number of genes, one would expect genes to drift apart unless there is a selective advantage to maintaining the gene pair. A recent analysis of divergent genes in the human chromosomes 20, 21 and 22 showed a biphasic distribution of the intergenic distance between such genes, with most genes being separated by an average of 25 kb and a minority by an average of 0.3 kb (Takai and Jones 2003), strongly suggesting a selective pressure in maintaining a close apposition between at least some head-to-head gene pairs. It is commonly thought that such selective pressure is imposed by the sharing of regulatory elements by the gene pair (see for example Labrador and Corces 2002; Takai and Jones 2003). Indeed, overlapping and shared promoter elements have been identified in a number of cases (see for examples Shinya and Shimada 1994; Yoshida et al. 2002; Hansen et al. 2003; Meyer et al. 2003; Otte, Schwaab and Luers 2003; Shin, Kim and Paek 2003; Zhang et. al. 2003). However, close apposition of a gene pair as seen in one genome could also be due to chance. Hence, before a functional significance can be attached to a gene pair, conservation of the close linkage needs to be shown first. Tracing these evolutionary conserved gene pairs is of considerable interest as it will help delineate the regulatory modules used in the eukaryotic genomes. In addition, comparison of the intergenic region between a conserved gene pair will provide insight into the evolution of eukaryotic promoter regions. With the elucidation of the sequence of eukaryotic genomes, it is becoming evident that changes in gene regulation rather than gene number are the driving force for phenotypic evolution (for recent review, see Wray et al. 2003).

Here we have focused on the intergenic region between two evolutionary related genes encoding the small heat shock proteins (sHsps) αB-crystallin and HspB2. These two proteins together with eight others form the sHsp family in man (Kappé et al. 2003). The αB-crystallin and HspB2 genes are located only about 0.9 kb apart in a head-to-head manner in the human, mouse and rat genomes and could thus share (at least part of) their promoter regions (Iwaki et al. 1997). The expression patterns of these two genes are, however, quite different. Products from both genes are found in heart and muscle, but only αB-crystallin and not HspB2 is expressed in lens (Iwaki et al. 1997). Also, αB-crystallin, unlike HspB2, is stress inducible (Suzuki et al. 1998). Putative elements involved

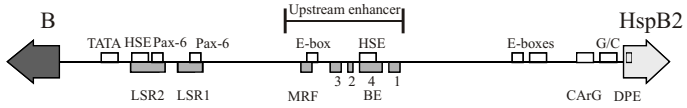


Figure 1 Schematic representation of the intergenic region between the αB-crystallin and the HspB2 genes. The known regulatory sites of the mouse αB-crystallin promoter are indicated in the figure as follows: TATA: TATA-box; HSE's: heat shock elements (Srinivasan and Bhat 1994; consensus sequence: inverted repeat of nGAAm); LSR: Lens specific regions (contain Pax-6 and RAR/RXR binding sites, Gopal-Srivastava and Piatigorsky 1994; Gopal-Srivastava, Cvekl and Piatigorsky 1996, 1998); MRF: binding site for the muscle response factor, binding to this element increases transcription in muscle (Gopal-Srivastava, and Piatigorsky 1993); αBE: αB-crystallin elements 1-4, enhance promoter activity in lens, muscle and heart (Gopal-Srivastava and Piatigorsky 1993; Gopal-Srivastava, Haynes and Piatigorsky 1995). The putative regulatory elements of the HspB2 promoter have been identified only on the basis of sequence analysis (Iwaki et al. 1997; Swamynathan and Piatigorsky 2002): E-box: binding site for MyoD family members; CArG-box: binding site for SRF; G/C: G/C-rich promoter region; DPE: Downstream promoter element.

in the regulation of expression of the HspB2 gene have been identified only at the sequence level (Fig. 1). In contrast, elements important for the regulation of expression of the mouse αB-crystallin gene have been well documented experimentally (Fig. 1). The two promoter proximal lens-specific regions (LSR's) both contain Pax-6 and RAR/RXR binding sites (Gopal-Srivastava, Cvekl and Piatigorsky 1996, 1998).The upstream enhancer encompasses four αBE elements and a muscle response factor (MRF) binding site. The αBE elements are important for expression in both lens and muscle, while the MRF is only involved in muscle expression (Dubin et al. 1991; Gopal-Srivastava and Piatigorsky 1993, 1994; Srinivasan and Bhat 1994; Gopal-Srivastava, Haynes and Piatigorsky 1995; Gopal-Srivastava, Cvekl and Piatigorsky 1996; 1998). A second, minor transcription initiation site is located just upstream from the enhancer, at -474 in the mouse (Dubin et al. 1991; Gopal-Srivastava, Haynes and Piatigorsky 1995). Two heat shock elements (HSE's) have been found in the αB-crystallin promoter (Srinivasan and Bhat 1994): one within the LSR2 and one in the upstream enhancer region (Fig.1).

As only the αB-crystallin gene is expressed in the lens or upregulated after heat shock, the elements for lens expression or heat shock responsiveness must be restricted to the αB-crystallin promoter and be isolated from the HspB2 promoter. Both αB-crystallin and HspB2 are expressed in muscle and could share the muscle specific elements in the intergenic region. However, Swamynathan and Piatigorsky (2002) concluded that the muscle enhancer acts unidirectionally towards the αB-crystallin promoter only. If the αB-crystallin and the HspB2 promoters do not share regulatory elements, there would be no obvious need to maintain the close distance between or the head-to-head orientation of these genes. We therefore checked the database whether the close head-to-head orientation was also present in species other than man and rodents. As this search showed that these genes had drifted further apart in duck, we sampled the major mammalian clades to determine whether the intergenic distance between this gene pair is also variable in mammals. We show here that the intergenic distance between the αB-crystallin/HspB2 gene pair shows little variation in mammals but is significantly larger in chicken and duck. As sequence alignment of the intergenic regions showed the absence of one of the HSE's and divergence of the LSR elements in platypus, we have measured the heat shock response and determined the lenticular activity of the platypus αB-crystallin promoter. Finally, we have tested whether shortening the intergenic region, and thus decreasing the distance between a HSE and the HspB2 promoter, could confer heat shock responsiveness to the rat HspB2 promoter.

Materials and Methods

Intergenic Region Sequences

Sequences of the intergenic regions of human (*Homo sapiens,* AP000907), rat (*Rattus norvegicus,* U04320), mouse (*Mus musculus,* NT_039473.1) and duck (*Anas platyrhynchos,* U16124) were retrieved from the GenBank database. For other species we performed PCRs on genomic DNA to obtain sequences containing the intergenic region between αB-crystallin and HspB2. One set of degenerated primers αB-1rev (TCTGAGAGYCCMGTSTCNADCCA) and Hsp2B-1rev (GGGTTGGCAAAYTCRTAYTC) was used for blind mole rat (*Nannospalax ehrenbergi*), rabbit (*Oryctolagus cuniculus*), pika (*Ochotona princeps*), cat (*Felis catus*), leaf-nosed bat (*Macrotus californicus*), shrew (*Crocidura russula*), anteater (*Cyclopes didactylus*), manatee (*Trichechus manatus*), opossum (*Didelphis marsupialis*), platypus (*Ornithorhynchus anatinus*) and chicken (*Gallus gallus*). Another set of degenerated primers was used for mole (*Talpa europaea*): αB-2rev (ATTCARCAGGTGYTCYCCRAAGA) and HspB2-2rev (GTGGCTGGGTGGGCATGYGYYA). PCRs were performed using the Expand HF kit (Roche). The DNA was first denatured at 94ºC for 3 min, and then cycles of denaturation (1 min at 94ºC), annealing of the primers (90 s at 55ºC) and elongation (2 min at 68ºC) were performed. PCR samples were taken after 40 cycles. PCR-fragments were cloned into the pGEM-T

vector and sequenced. A global pair-wise alignment of the intergenic regions was made with using the Bayes Aligner (a Bayesian block aligner; available at http://www.Bayesweb.wadsworth.org). The histogram indicates the probability that any given base j in one sequence aligns to any base k in the other sequence. Probabilities are determined from a set of alignments representative of all possible alignments of the two sequences (Zhu, Liu and Lawrence 1998; Wasserman et al. 2000). The Match program, available at http://www.gene-regulation.com, was used to find additional possible transcription factor binding sites in the conserved regions. Direct sequence alignments were produced using ClustalW and adjusted manually using Genedoc.

Constructs

The starting vector for the bidirectional reporter constructs was made by inserting the coding sequence of β-galactosidase from pCH110 (Pharmacia Biotech) *Hind*III blunt and *Bam*HI blunt into the *Sma*I site of the pGL3 basic vector (Promega).

Rat luc-αB-HspB2-β-gal: The rat intergenic region was amplified from rat genomic DNA by PCR using the Rat αB *Bgl*II primer (GAAGATCTGAGTGTAGAGTCGGTTAGC) at position +35 relative to the αB-crystallin transcription start site and the Rat HspB2 *Xho*I primer (CCGCTCGAGTGTAGCCCCAACAAGATC) at position +62 relative to the HspB2 transcription start site. When cloning this fragment *Xho*I/*Bgl*II into the bidirectional reporter vector this results in the luciferase reporter gene being driven by the αB-crystallin promoter and the β-galactosidase reporter gene being driven by the HspB2 promoter. For the Rat β-gal-αB-HspB2-luc (the reversed construct) the primers were: Rat αB *Xho*I (CCGCTCGAGTGTAGAGTCGGTTAG) and Rat HspB2 *Bgl*II (GAAGATCTAGTGTAGCCCCAACAAGA) and the PCR product was also inserted *Xho*I/*Bgl*II in the bidirectional reporter construct.

The rat deletion constructs were made by PCR using the Rat αB *Bgl*II primer and primers located at position -393 relative to +1 of the αB-crystallin gene (Δ630: AACTGCAGCCCAGGAAGATTCCAGC) for the Rat Δ630 luc-αB-HspB2-β-gal clone and at position -177 relative to +1 of the αB-crystallin gene (Δ850: AACTGCAGCCCTGCCC CGTGTTTC) for the Rat Δ850 luc-αB-HspB2-β-gal clone. These PCR fragments were recloned into the rat luc-αB-HspB2-β-gal construct using the *Pst*I at position -682 relative to +1 of the αB-crystallin gene and *Bgl*II at αB-crystallin side of the construct. The reversed constructs were made by PCR on these deletion constructs using the primers for the rat β-gal-αB-HspB2-luc construct and inserting *Xho*I/*Bgl*II into the bidirectional reporter construct.

The sequences of the platypus and blind mole rat intergenic regions obtained as described above were used to design primers to clone the respective intergenic regions into the bidirectional reporter vector. For Platypus luc-αB-HspB2-β-gal the primers used were: Platypus αB *Bgl*II (GAAGATCTGCTCTGGCTGGCTGGGCG) and Platypus B2 *Xho*I (CCGCTCGAGGGACACTGGCCGGACGC), and for Platypus β-gal-αB-HspB2-luc, the primers used were: Platypus αB *Xho*I (CCGCTCGAGGCTCTGGCTGGCTGGGCG) and Platypus B2 *Bgl*II (GAAGATCTGCGCTGCGGACACTGGCC). For the Blind mole rat luc-αB-HspB2-β-gal construct: Blind mole rat αB *Bgl*II (GAAGATCTAATGTAGGGGGTCAGCTGG) and Blind mole rat B2 *Xho*I (CCGCTCGAGGCAGCCCCAACAAGCTCAGTA) primers were used and for the Blind mole rat β-gal-αB-HspB2-luc construct: Blind mole rat αB *Xho*I (CCGCTCGAGAATGTAGGGGGTCAGCTG) and Blind mole rat B2 *Bgl*II (GAAGATCTGCAGCCCCAACAAGCTCAGTA) primers were used. PCR fragments were inserted *Xho*I/*Bgl*II into the bidirectional reporter vector.

pCSGFP2-intergenic region constructs: For *Xenopus laevis* transgenesis the intergenic regions of rat, blind mole rat and platypus were cloned blunt into pBluescript (Stratagene) and then either orientation was cloned into the pCSGFP2 construct (kindly provided by Dr. E. Amaya, Wellcome,

Cambridge, UK) using *Sal*I/*Bam*HI for the rat and blind mole rat and *Sal*I/*Eco*RI for the platypus constructs. This resulted in constructs in which either the αB-crystallin promoter or the HspB2 promoter of the intergenic region drives EGFP expression.

Cell Culture

C2 cells (mouse myoblast cells) were cultured in Dulbecco's modified Eagle's medium (Gibco) with penicillin and streptomycin (Roche) and supplemented with 20% fetal calf serum (PAA laboratories) to prevent differentiation of these cells. To obtain lens fiber cells, four lens epithelial cell explants from newborn rats were cultured in M199 medium (Sigma) supplemented with 0.1% BSA (Roche), penicillin and streptomycin, glutamax-1 (Gibco) and 50 ng/ml FGF-2 for 2 days prior to transfection (Chamberlain and McAvoy 1987; Klok et al. 1998).

Transfection

C2 cells and lens explants were transfected using lipofectAMINE plus (Invitrogen). Approximately $6.5x10^4$ C2 cells were plated in Dulbecco's modified Eagle's medium with penicillin and streptomycin and 10% fetal calf serum in 6 well plates and cultured for 24 h. Lens fiber cells (4 explants per 35 mm dish) were transfected after 48 h FGF-2 treatment. Both C2 cells and lens explants were transfected with a total of 1 µg DNA per well using 6 µl plus reagent and 4 µl of lipofectAMINE in Dulbecco's modified Eagle's medium. The total of 1 µg DNA per well was divided into 0.9 µg of the various bidirectional reporter constructs and 0.1 µg of pEGFP (Clontech) as a transfection control. After 4 hours, the medium was replaced by Dulbecco's modified Eagle's medium with 10% FCS for C2 cells or M199 with 0.1% BSA and 50 ng/ml FGF-2 for lens explants. C2 cells were harvested 48 h after transfection or heat shocked 48 h after transfection and harvested after 6 h of recovery at 37ºC and assayed for reporter gene activity. Lens explants were harvested 72 h after transfection and assayed for reporter gene activity.

Heat Shock

C2 cells were heat shocked 48 hours after transfection by submerging the 6 well plates into a 45ºC water bath for 30 min, harvested after 6 h of recovery at 37ºC and assayed for reporter activities (β-galactosidase and luciferase) as described in the next section.

Reporter Assays

Cells were harvested by vigorously shacking in 200 µl reporter lysis mix (25 mM Bicine pH 7.5, 0.05% Tween-20 and 0.05% Tween-80) per well. Lens explants were harvested and lysed by vigorously shaking in an eppendorf tube with 100 µl of reporter lysis mix. 20 µl of these lysed cells or explants was used for the reporter assays.

For the β-galactosidase assay, galacton (Tropix) was diluted 1:100 in 100 mM phosphate buffer pH 8.1, 5 mM $MgCl_2$; of this dilution 200 µl was added to 20 µl of the cell lysate. After 30 minutes incubation at room temperature, 300 µl of light emission accelerator (Tropix) was added. For the luciferase assay 100 µl of luciferase reagent (Promega) was added to 20 µl of the lysate immediately before measurement. Measurements were performed on a Lumat LB 9507 luminometer for 10 seconds. All experiments were performed at least in duplicate, all data shown are the averages of at least two independent experiments.

Xenopus Transgenesis

Xenopus laevis unfertilized eggs and sperm nuclei were obtained as described previously (Jansen et al. 2002). The intergenic region-EGFP DNA fragments obtained by digesting the pCSGFP2-intergenic region constructs with *Sal*I and *Not*I, were purified using a Qiaex II gel extraction kit (Qiagen). The fragment (250 ng/5 µl) was mixed with $2.5x10^5$/2.5 µl sperm nuclei, diluted to 500 µl and ~10 nl was injected per egg as described previously (Jansen et al. 2002). After 3 hours at 18ºC, embryos at the 4 cell stage were separated and put in 0.1x MMR, 6% Ficoll and 50 µg/ml gentamycin; after 24 hr at 18ºC gastrulas were again transferred to a new dish containing 0.1x MMR and gentamycin and kept at 22ºC. Pictures of several differentiation stages were taken using Leica MZFLIII fluorescence microscope.

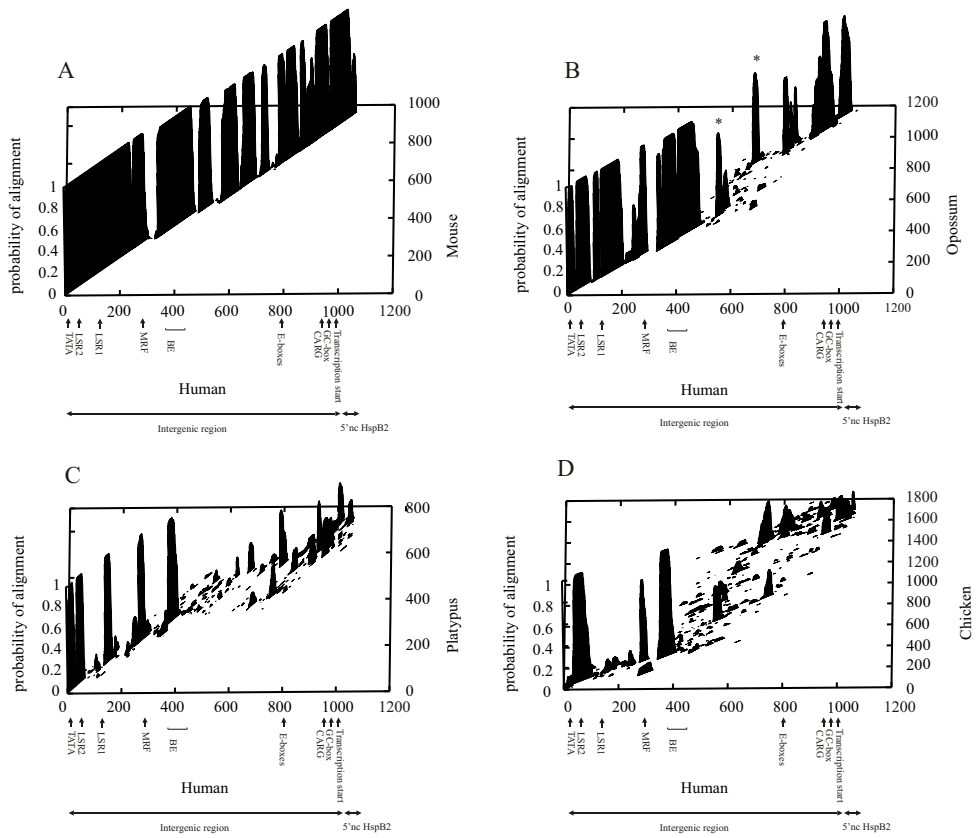


Figure 2 Phylogenetic footprinting of the intergenic region between the αB-crystallin and HspB2 genes. Two-dimensional histogram output of the Bayesian block aligner (Bayes Aligner; Zhu, Liu and Lawrence 1998; Wasserman et al. 2000) The histogram indicates the probability that any given base j in the human sequence aligns to any base k in the mouse (A), opossum (B), platypus (C) or chicken (D) sequence. Probabilities are determined from a set of alignments representative of all possible alignments of the two sequences. The known regulatory sites of the mouse αB-crystallin promoter are indicated in the figure, the sites indicated close to the HspB2 promoter are based on sequence analysis only (Iwaki et al. 1997). The two conserved elements identified by the human-opossum sequence comparison are indicated with an asterisks (*).

Results

Phylogenetic footprinting and alignment of mammalian and avian sequences of the intergenic region between the αB-crystallin and HspB2 genes shows conservation and divergence of regulatory elements

An initial database search for genomic sequences of the mammalian αB-crystallin and HspB2 genes showed that sufficient sequence information to map these two genes head-to-head with an intergenic distance of about 0.9 kb was only available from man, mouse, and rat. A head-to-head orientation was also found in the sequence from duck. Here, the 5' region of the HspB2 gene was present about 1.6 kb upstream of the αB-crystallin gene (N.B. at the time the duck αB-crystallin sequence was published [Wistow and Graham 1995], the link with the HspB2 gene could not be noted as the HspB2 gene was only identified later [Iwaki et al. 1997]). The larger distance between the αB-crystallin and the HspB2 genes in duck raised the question whether the close linkage found in man and rodents was the exception or the rule in mammals. Sampling of a broad variety of mammalian taxa, including a marsupial (opossum) and a monotreme (platypus), showed that the average length of the intergenic region in mammalian species is about 0.9 kb, with platypus having the shortest one (645 bp) and opossum the longest (1069 bp; Table 1). The length of the intergenic region in chicken was found to be almost the same as in duck (1687 bp versus 1640 bp; Table 1). Thus close linkage of the αB-crystallin/HspB2 gene pair is a conserved feature of the mammalian genome, while a larger distance is found in two avian genomes.

Table 1 Length of the intergenic regions between the αB-crystallin and HspB2 genes

Species	Length o intergenic region bp
Mouse	866
Rat	908
Blind mole rat	906
Rabbit	928
Pika	928
Human	964
Cat	938
Bat	884
Mole	973
Shrew	912
Anteater	958
Manatee	923
Opossum	1069
Platypus	645
Chicken	1687
uck	1640

* Distances between the transcription start sites of the αB-crystallin and HspB2 genes.

To identify conserved sequence elements, the intergenic regions and the 5' non-coding regions of the HspB2 gene were aligned using the Bayes Aligner (Zhu, Liu and Lawrence 1998; Wasserman et al. 2000; for examples, see Fig. 2). A remarkable conservation of sequence, particularly of the 5' flanking sequence of the αB-crystallin gene, was found (see for example the comparison between the human and opossum sequence, Fig. 2). Only when more distant species were compared, are phylogenetic footprints corresponding to most of the known regulatory sites of the mouse αB-crystallin promoter (see Fig. 1) clearly visible (Fig. 2). Sequence similarity decays faster on the HspB2 side of the intergenic region. The phylogenetic footprints mark putative E- and CArG boxes, a G/C rich promoter region and a conserved region in the 5' non-coding region of the HspB2 gene (see also Fig. 1). Two additional possible regulatory sites emerged from this analysis, one around 600 and one around 700 (see asterisks Fig. 2, numbering for the human sequence). The first contained the consensus binding site for c-Rel, a member of the REL/NF-κB/IκB superfamily of transcription factors which are involved in (anti-)apoptosis and cellular transformation (Foo and Nolan 1999). The second region matched the consensus sequence for the octamer sequence recognized by Oct transcription factors, members of the POU protein family (for review, see Phillips and Luisi 2000). This region also contained the sequence recognized by Elk-1, one of the proteins which forms part of the ternary complex with the Serum Response Factor (SRF), but lacks the SRF recognition site, which is usually adjacent (for review, see Shaw and Saxton 2003). A search for the recognition site of the CTCF transcription factor (consensus sequence CCGCNNGGNGGCAG; Ishihara and Sasaki 2002), which

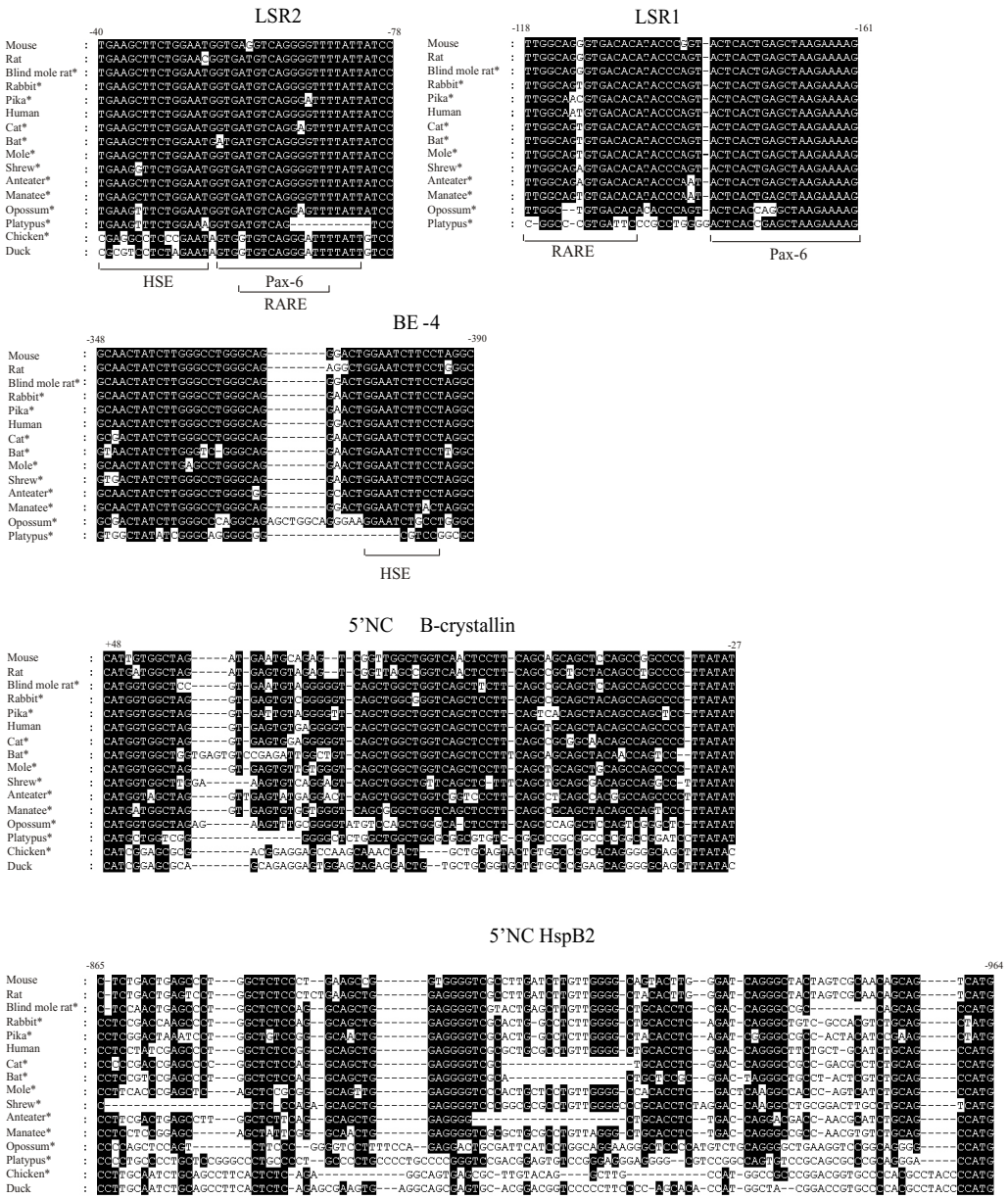


Figure 3 Alignment of some conserved elements in the intergenic regions between the αB-crystallin and the HspB2 genes. Sequences of the intergenic regions between mammalian and avian αB-crystallin and the HspB2 genes were aligned using ClustalW and adjusted manually using Genedoc (see supplementary material for the complete alignment). Here only the alignments of the LSR1 and 2, the αBE-4 element, and the 5' UTR's of the αB-crystallin and HspB2 genes are shown; the full alignment is shown in the supplementary material. Chicken and duck sequences were omitted from the LSR1 and αBE-4 alignments, since the LSR1 is not present (see fig. 2) and the αBE-elements contain insertions (see supplementary material). Position numbers are given for the mouse sequences and are relative to αB-crystallin transcription start site (+1). Asterisks (*) indicate newly determined sequences.

mediates insulator activity (for review, see Burgess-Beusse et al. 2002; Kuhn and Geyer 2003), yielded a number of matches but none in a conserved region.

The phylogenetic footprinting together with direct inspection of the complete sequence alignment (supplementary material) revealed some notable changes in (putative) regulatory elements: the complete loss of LSR1 in chicken (and duck; Fig. 2 and supplementary material), the sequence divergence of the 5' part of LSR1 in platypus (Fig. 3), the deletion of the Pax-6 site in the platypus LSR2 (Fig. 3) and the deletion of the MRF binding sequence in the blind mole rat (Hough et al. 2002; supplementary material). The avian sequences have insertions in the elements in the αBE region, separating conserved sequence blocks (supplementary material). In addition, the avian sequences lack HSE's, in agreement with the lack of heat shock response of the duck αB-crystallin gene (Wistow and Graham 1995). In platypus the HSE in the so-called αB-4 element in the αBE region is absent (Fig. 3), while mutations in the manatee and opossum sequences possibly inactivate the HSE in this element. All mammals have a HSE consensus sequence in LSR2 (Fig. 3). The sequence alignment also shows a conservation of the +20 to +40 region of the HspB2 gene (Fig. 3). This could be indicative of a DPE element, as suggested by Swamynathan and Piatigorsky (2002), although the DPE consensus sequence (GNNN[A/G][A/T][C/T][G/A/C]; Kadonaga 2002) is not easily discerned. There is a striking conservation of the 5' non-coding region of the αB-crystallin gene relative to that of the HspB2 gene in mammals (Fig. 3).

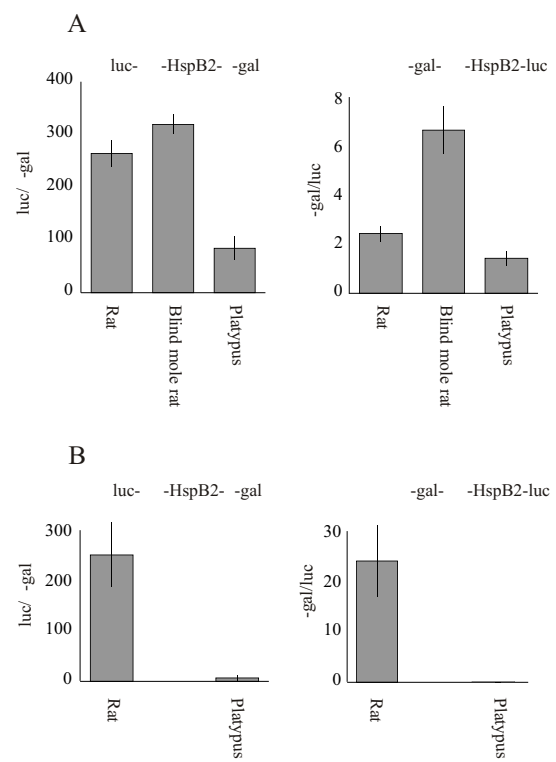


Figure 4 The activity of the αB-crystallin promoter relative to that of the HspB2 promoter of the rat, blind mole rat and platypus in C2 or lens cells. **A**: C2 cells were transfected with constructs containing the intergenic region of rat, blind mole rat or platypus. In these experiments transfection efficiencies were monitored by co-transfection of an EGFP expression construct. Visual inspection showed no major differences in the efficiency of transfection of the various constructs. αB-crystallin promoter activities were calculated relative to HspB2 promoter activities for both luc-αB-HspB2-β-gal and β-gal-αB-HspB2-luc constructs. Error bars indicate the standard deviation. **B**: Differentiated lens explants (lens fiber cells) were transfected with constructs containing the intergenic region of rat or platypus. αB-crystallin promoter activities were calculated relative to HspB2 promoter activities for both luc-αB-HspB2-β-gal and β-gal-αB-HspB2-luc constructs. Error bars indicate the standard deviation.

The Platypus αB-Crystallin Promoter Is Less Active Relative to the HspB2 Promoter Than the αB-Crystallin Promoters of Rat and Blind Mole Rat

The data presented above (Table 1, Figs. 2 and 3) show that the length of the intergenic region between the αB-crystallin and the HspB2 genes is conserved in mammals and that the known regulatory elements of the mouse αB-crystallin promoter are conserved as well, with the exception of platypus. To determine the functional significance of the sequence divergence of the platypus intergenic region, the activity of the platypus αB-crystallin and HspB2 promoters in lens and muscle cells was measured. To that end, we cloned the intergenic region of platypus in a bidirectional reporter construct containing the luciferase coding region on one side of the intergenic region and β-galactosidase coding region on the other side. For comparison, we used the rat intergenic region cloned into the bidirectional reporter construct, while the blind mole rat intergenic region in the bidirectional reporter construct was used to test whether our assay systems could detect the switch from lens to muscle expression of the blind mole rat αB-crystallin promoter described by Hough et al. (2002). The intergenic regions were inserted in both directions between the luciferase and β-galactosidase reporter genes to make sure that differences in activity of αB-crystallin and HspB2 promoters are not due to differences in reporter gene activity and sensitivity of the assays. Constructs are called luc-αB-HspB2-β-gal when the αB-crystallin promoter drives luciferase expression and the HspB2 promoter drives β-galactosidase expression; the reverse construct is called β-gal-αB-HspB2-luc. The relative activity of the αB-crystallin promoter with respect to that of the HspB2 promoter, or vice versa, can then be determined from the ratio of luciferase to β-galactosidase activity.

The activity of the promoter constructs in C2 myoblasts, indicative of expression in muscle, is shown in fig. 4A. For unknown reasons there was a difference between the relative promoter activities when either luciferase or β-galactosidase was used as reporter gene, making comparisons more difficult. However, a clear trend was seen. The platypus αB-crystallin promoter is between 50 and 75% (depending on the reporter gene) weaker with respect to the platypus HspB2 promoter than the corresponding rat promoters. For the blind mole rat, the ratio of the activity of the αB-crystallin promoter to that of the HspB2 promoter was two-fold higher than that of the corresponding rat promoters when β-galactosidase was driven by the αB-crystallin promoter. The difference was marginal when the αB-crystallin promoter drives the luciferase gene. These data suggest, in agreement with the work of Hough et al. (2002), that the blind mole rat αB-crystallin promoter is more active in muscle cells than the rat αB-crystallin promoter.

To test the activity of the αB-crystallin and HspB2 promoters in lens cells, the bidirectional constructs were transfected into rat lens fiber cells, obtained by in vitro differentiation of explanted newborn rat lens epithelial cells. The rat and blind mole rat αB-crystallin promoters were very active in this system, as expected. Less activity was obtained from the platypus αB-crystallin promoter. When the relative activity of the αB-crystallin promoter with respect to the HspB2 promoter is calculated, then the expression directed by the rat αB-crystallin promoter relative to HspB2 promoter is about 30 times higher than that directed by the platypus αB-crystallin promoter relative to the platypus HspB2 promoter (Fig. 4B). For the blind mole rat, the activity of the reporter genes driven by the HspB2 promoter in lens is not significantly above background. Thus, no relative activities for the blind mole rat promoters could be calculated.

The Platypus αB-Crystallin Promoter is Activated by a Heat Shock

Since the platypus intergenic region lacks the HSE in the αB-4 element (Fig. 3), we tested whether there are differences in heat shock response between rat, blind mole rat and platypus αB-crystallin and HspB2 promoters. The bidirectional reporter constructs were transfected into C2 cells, these cells were heat shocked 48 h after transfection and assayed for luciferase and β-galactosidase activity

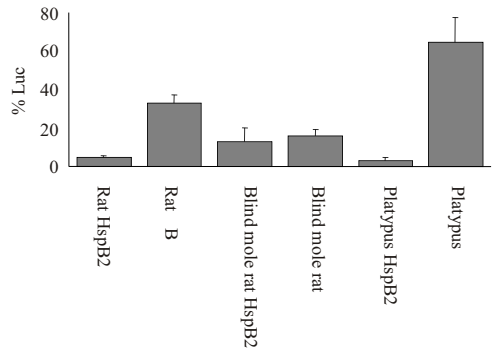


Figure 5 Heat shock response of the rat, blind mole rat and platypus αB-crystallin and HspB2 promoters in C2 cells. C2 cells transfected with the constructs containing the intergenic region of rat, blind mole rat or platypus were heat shocked 48 h after transfection for 30 min at 45ºC and harvested after 6 h of recovery at 37ºC; control cells were not heat shocked. The luciferase values were corrected for differences in transfection efficiency using the β-galactosidase values. The luciferase activity (Luc) measured after heat shock and recovery is expressed relative to that of non-heat shocked cells, which was set at 100%. Error bars indicate the standard deviation.

after 6 h of recovery at 37°C. An experimental problem in these experiments is that the promoters are already active before the heat shock is applied and a distinction must thus be made between the amount of reporter gene product (protein or mRNA) present before the heat shock and the additional amount made in response to the heat shock. To circumvent this problem we argued as follows. Luciferase is heat labile, hence after a heat shock only newly synthesized luciferase (and not luciferase present before heat shock) will be measured. However, the pre-existing luciferase mRNA is still present. To determine how much luciferase is made after heat shock from pre-existing mRNA, we used the rat HspB2 promoter as reference as this promoter has already been shown to be the non-heat shock inducible (Iwaki et al. 1997). When this promoter drives luciferase expression (the rat β-gal-αB-HspB2-luc construct), the luciferase activity after 6 hours of recovery from a heat shock was about 5% of that found in non-heat shocked cells (Fig. 5). Hence recovery of 5% of the initial luciferase activity was taken as indicative of a lack of heat shock response. In contrast to luciferase, β-galactosidase is not heat labile. Any additional increase in β-galactosidase activity due to the heat shock will be marginal compared to the activity already present in the cell and the level of β-galactosidase activity was thus taken as a control for the transfection efficiency.

Figure 5 shows the relative amount of luciferase after a heat shock compared to that without a heat shock for the bidirectional promoter constructs from the three species. For the rat, the amount of luciferase activity from the rat luc-αB-HspB2-β-gal construct after heat shock was 33% of the control, indicating a heat shock response. For platypus the relative amount of luciferase activity obtained from the αB-crystallin promoter in the construct luc-αB-HspB2-β-gal was twice as high (64%), indicative of a strong heat shock response. The relative amount of luciferase activity recovered after heat shock from the platypus HspB2 promoter (the β-gal-αB-HspB2-luc) was even less than that from the rat HspB2 promoter and the platypus HspB2 promoter is thus unlikely to be heat shock inducible. Finally, the blind mole rat αB-crystallin promoter was relatively less active after a heat shock than the rat or platypus αB-crystallin promoters, while the blind mole rat HspB2 promoter seems to be more active. A possible explanation for the latter finding is that in the blind mole rat the HSE's act towards both the αB-crystallin and the HspB2 promoter thereby decreasing the effect on the αB-crystallin promoter.

Deletion of the Rat Intergenic Region Decreases the Relative Activity of the αB-Crystallin Promoter and Makes the HspB2 Promoter Slightly Heat Shock Responsive

The data presented in Figure 5 show that in vitro, as in vivo, the HSE's present in the rat or platypus intergenic region act unidirectionally towards the αB-crystallin promoter. To determine whether the rat HspB2 promoter becomes heat shock sensitive if the HSE's are moved closer to this promoter, two deletion constructs were made (Fig. 6). In the Δ630 construct, 290 bp spanning the region

between the E-boxes at the HspB2 side of the promoter and the upstream HSE were deleted, moving this element close to the HspB2 promoter. In the Δ850 construct the region between the E-boxes and the LSR1 was deleted, thus placing the HSE in LSR2 close to the HspB2 promoter. In non heat shocked C2 cells (Fig. 6A), the Δ850 deletion adversely affected the activity of both promoters (data not shown) with the αB-crystallin promoter relatively more inhibited than the HspB2 promoter (Fig. 6A). The Δ630 deletion had little effect on the relative activity of the αB-crystallin promoter when that promoter was driving the β-gal reporter gene, while it appeared to have a strong negative effect when the αB-crystallin promoter was driving the luciferase reporter gene, making it difficult to interpret the effect of this deletion on promoter activity (Fig. 6A). After heat shock (Fig. 6B), the HspB2 promoter in either one of the two deletion constructs was more active than in the wild type construct. The activity of the αB-crystallin promoter after heat shock was not affected by the Δ630 deletion but was diminished by the Δ850 deletion (Fig. 6B).

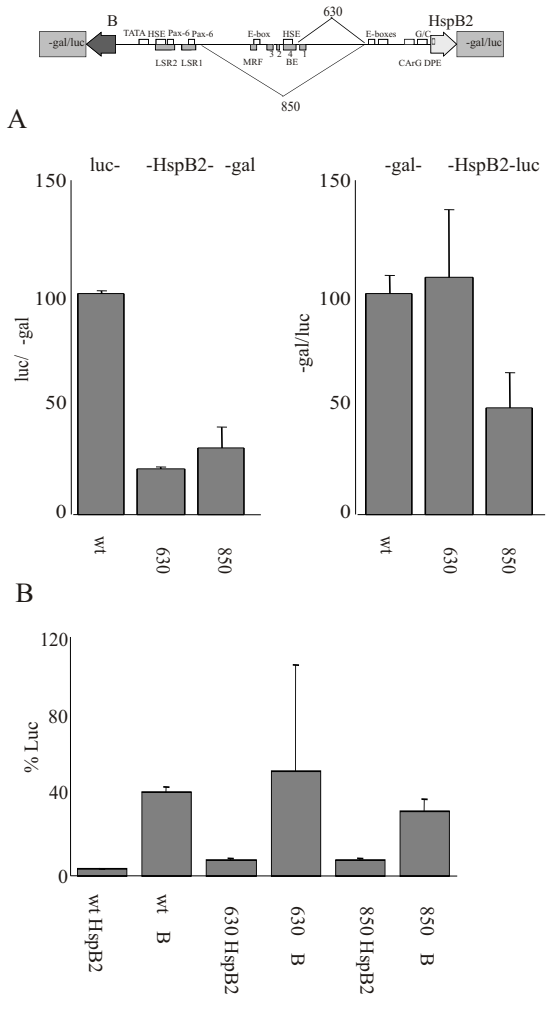


Figure 6 Effect of deleting part of the intergenic region on the activity and heat shock response of the rat αB-crystallin and HspB2 promoters in C2 cells. **A**: Deletion constructs of the rat intergenic region (schematic representation) in the bidirectional reporter vector were transfected into C2 cells. The figure shows the ratio of activity of the αB-crystallin and HspB2 promoters as calculated from the levels of luciferase and β-galactosidase relative to that obtained from the full length constructs. **B**: C2 cells transfected with the full length or deletion constructs of the rat intergenic region were heat shocked 48 h after transfection for 30 min at 45ºC and harvested after 6 h of recovery at 37ºC, control cells were not heat shocked. The luciferase values were corrected for differences in transfection efficiency using the β-galactosidase values. The luciferase activity (Luc) measured after heat shock and recovery is expressed relative to that of non-heat shocked cells, which was set at 100%. Error bars indicate the standard deviation.

The αB-Crystallin and HspB2 Promoters Are Active in the Lens of *Xenopus laevis* Transgenic Animals.

In *in vitro* differentiated lens fiber cells the platypus αB-crystallin promoter was only poorly active (Fig. 4B). To rule out that the activity of this promoter was only due to leakiness of the transfection system and not to lens-specific recognition, we turned to an *in vivo* system based upon an animal which is evolutionarily equidistant from rat, blind mol rat and platypus, namely *Xenopus laevis*. We have previously shown that a rat γ-crystallin promoter is recognized as a lens-specific promoter in *Xenopus laevis* (Brakenhoff et al. 1991). Constructs containing the whole intergenic region and in which either the αB-crystallin or the HspB2 promoter drives an EGFP reporter gene were injected into *Xenopus* oocytes. Expression was monitored during development. When EGFP expression was regulated by the rat γD-crystallin promoter, expression is mainly seen in lens, but a slight background staining was also found during the first 8 days of development. In contrast, EGFP expression driven by either the HspB2 or the αB-crystallin promoter shows specific staining of the somites as well as lens after 2 days (Fig. 7). After 4-8 days not only muscle and lens but also heart and some brain and neuronal tissue can be seen to be fluorescent (note that the expression level -

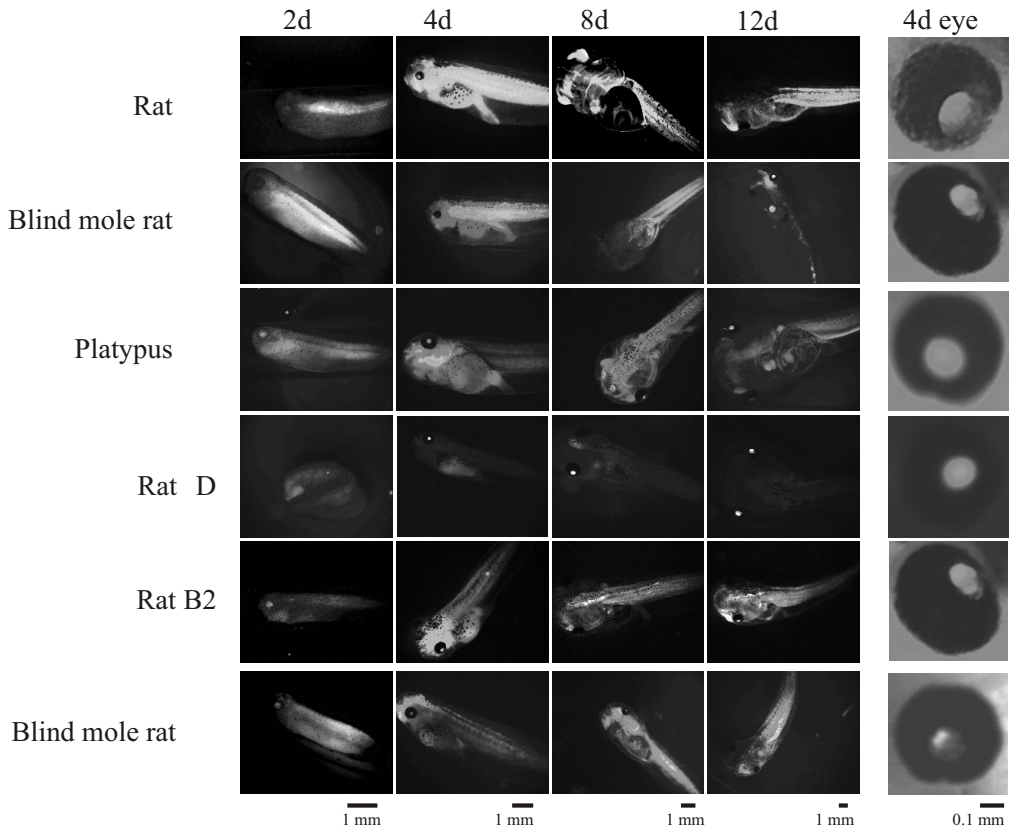


Figure 7 Expression of EGFP directed by the αB-crystallin and HspB2 promoters in transgenic *Xenopus laevis*. The rat, blind mole rat and platypus intergenic regions cloned in either direction (in the case of rat and blind mole rat) or a single direction (in the case of platypus) in front of the EGFP reporter gene were used to generate transgenic *X. laevis* tadpoles. Transgenic animals with the rat γD-crystallin promoter driving EGFP expression are shown as a control. Transgenic animals were photographed after 2, 4, 8 and 12-15 days of development. An enlargement of the eye is shown of animals at the 12-15 days stage of development. Representative photographs are shown.

but not the expression pattern - varied between embryos and that the pictures shown in Figure 7 are qualitative not quantitative). The most striking observation was that in these *Xenopus* larvae there was little difference in the pattern of EGFP expression from the αB-crystallin or from the HspB2-crystallin promoter, in particular, clear lens staining was seen as well when EGFP expression was driven by the HspB2 promoter. These data suggest that in *Xenopus laevis* the HspB2 promoter is also recognized as a lens-specific promoter since we only observed staining of the lens in *Xenopus* transgenes when lens-specific (e.g. crystallin) promoters are used, not when muscle- or brain-specific promoters are driving EGFP expression (data not shown).

Discussion

We have shown here that the head-to-head orientation of the αB-crystallin/HspB2 gene pair is present in all mammalian lineages as well as in birds. Assuming that this gene pair originated from an inverted duplication of an ancestral sHsp gene, then our results show that the close linkage has been retained through mammalian evolution. It is commonly suggested that maintenance of a close apposition of head-to-head promoter regions is due to the selective pressure imposed by the sharing of promoter elements (see for example Shinya and Shimada 1994; Labrador and Corces 2002; Yoshida et al. 2002; Meyer et al. 2003; Otte, Schwaab and Luers 2003; Shin, Kim and Paek 2003; Takai and Jones 2003; Zhang et. al. 2003). Given the difference in expression pattern between the αB-crystallin and the HspB2 genes, one would then have to postulate that the expression in muscle requires shared elements, while the lens specific regions and the HSE's are directed at the αB-crystallin promoter only. Swamynathan and Piatigorsky (2002) have shown that deletion of the upstream enhancer element (the αBE region and the MRF), which is involved in the expression of the αB-crystallin gene in muscle, decreased the activity of the HspB2 promoter in C2C12 cells as well, although the inhibitory effect was much less than that for the αB-crystallin promoter. Similarly, we found that deletion of the enhancer region (the Δ850 deletion) inhibited both the αB-crystallin and the HspB2 promoters, with the αB-crystallin promoter being the most affected. These data suggest that at least some elements within the complex enhancer region work towards the HspB2 promoter. It is not clear whether these elements are actually shared by the two promoters as Swamynathan and Piatigorsky (2002) demonstrated that inverting the upstream enhancer region did not enhance the expression of the HspB2 promoter, while we found that merely shortening the distance between the upstream enhancer and the HspB2 promoter did not increase the activity of the HspB2 promoter (Fig. 6 and data not shown). Apparently most of the elements within this complex enhancer region work only towards the αB-crystallin promoter, irrespective of distance or orientation. It is thus unlikely that a single element enhances expression of both promoters. Rather, elements acting on the HspB2 promoter may be interspersed with those acting on the αB-crystallin promoter, where occupancy of one element working towards one promoter could promote occupancy of a second element working towards the other promoter. Intermingling of elements and synergistic binding of transcription factors to the different elements could be one explanation for the selective pressure to maintain the complex enhancer region and the head-to-head orientation. Alternatively, the selective pressure could be due to as yet unknown regulatory phenomena. An obvious possibility is that one or both of the two conserved regions between the upstream enhancer region and the HspB2 promoter detected by phylogenetic footprinting is required for optimal function of both the αB-crystallin and the HspB2 promoter. Neither of these two regions is absolutely required for promoter activity as deletion of these regions in the Δ630 or Δ850 deletion constructs did not silence either the αB-crystallin or the HspB2 promoter. However, factors binding to these regions could play a crucial role in the transcriptional response to stress or apoptotic insults, an obvious role for members of the REL/NF-κB/IκB family of transcription factors (Foo and Nolan 1999), or in setting the proper level of expression during development, one of the functions of the Oct transcription factors (Phillips and Luisi 2000). Clearly, the role, if any, of these two conserved elements needs to be elucidated and the functional elements of the

HspB2 promoter need to be mapped before a model of the regulatory functions embedded in the intergenic region between the αB-crystallin and HspB2 genes can be formulated.

The LSR's and the HSE's present in the intergenic region between the αB-crystallin and the HspB2 gene work towards the αB-crystallin gene only, as only this gene is expressed in the lens and only this gene is heat shock inducible. Yet the textbook definition of an enhancer is that its stimulatory effect on the rate of transcription initiation is independent of the location or orientation of the enhancer. One possibility is that the HspB2 promoter is silenced in the lens by an epigenetic mechanism. DNA methylation has been suggested to be responsible for the inactivity of γ-crystallin promoters in the lens epithelial cells (Peek et al. 1991), while differential histone acetylation controls the alternative activation of the murine bidirectional TK and KF promoters (Schuettengruber et al. 2003). Alternatively, the transcription factors which mediate the enhancing effect may require a TATA box promoter and be incompatible with the GC rich promoter region of the HspB2 promoter. In the case of the HSE's this is unlikely as the bidirectional promoter region of the Hsp60 and Hsp10 genes lacks TATA boxes and contains only a single HSE that acts on both promoters (Hansen et al. 2003). Finally, as suggested by Swamynathan and Piatigorsky (2002), the enhancing effect may be blocked by an insulator. If this insulating effect is mediated by CTCF, a common blocker of enhancer activity in mammals (Bell, West and Felsenfeld 1999), then the location of that insulator must be variable as none of our phylogenetic footprints contained the CTCF recognition sequence. Our experiments do not address the question whether CTCF is involved since CTCF insulator activity is mediated by chromatin structure (Burgess-Beusse et al. 2002; Labrador and Corces 2002; Kuhn and Geyer 2003) and is unlikely to be detected in transient transfection assays. In the *Xenopus* transgenes, the constructs are integrated in the genome, but, at least during embryogenesis, CTCF expression in *Xenopus laevis* is restricted and absent in lens fiber or muscle cells (Burke et al. 2002).

In a comparison of 51 rodent and human promoter regions, Dermitzakis and Clark (2002) found that about 30-40% of the transcription factor binding sites in the human promoter regions were not functional in the orthologous rodent promoter regions. Such a divergence between species is unlikely in the case of the αB-crystallin promoter. All elements important for expression of this promoter in either the lens or muscle as determined for the murine promoter (Dubin et al. 1991; Gopal-Srivastava and Piatigorsky 1993; Gopal-Srivastava and Piatigorsky 1994; Srinivasan and Bhat 1994; Gopal-Srivastava, Haynes and Piatigorsky 1995; Gopal-Srivastava, Cvekl and Piatigorsky 1996, 1998), are conserved in the human sequence as well as in other mammals, with the exception of platypus. The platypus promoter has an intriguing pattern of divergence: the LSR2 is present, except for the Pax-6 site, while of the LSR1 only the Pax-6 site is found. If Pax-6 interacts with the other factors binding to the LSR2, one would expect that loss of the Pax-6 site would relieve the selective pressure to maintain the binding site of other factors and that LSR2 would be lost, yet in the platypus promoter lacks most of the LSR1. As both the LSRs bind RAR/RXR and Pax-6 (Gopal-Srivastava and Piatigorsky 1994; Gopal-Srivastava, Cvekl and Piatigorsky 1996, 1998), these two regions could have originated from a duplication of an ancestral LSR. Initially then, two redundant complexes would have been present: a complex of Pax-6 with the factors interacting with the LSR1 and a complex of Pax-6 with the same factors binding to the LSR2. Loss of the Pax-6 site from LSR2 could have resulted in an interaction of the Pax-6 bound to LSR1 with the factors binding to LSR2. The remainder of LSR1 would then no longer be functional and that sequence could diverge freely. In the avian promoter region only the LSR2 is found; the LSR1 is missing. The platypus, chicken and duck promoter sequences do illustrate that a single LSR complex suffices for lens expression of the αB-crystallin promoter. It has been shown experimentally that the LSR2 can direct expression of the mouse αB-promoter to the lens, although the level of expression is low (Gopal-Srivastava, Cvekl and Piatigorsky 1996). The platypus promoter also lacks the upstream HSE but does have the proximal HSE which explains why the promoter is heat shock responsive. The avian αB-crystallin promoters lack both HSE's and at least the duck αB-crystallin

promoter has been reported to be insensitive to a heat shock (Wistow and Graham 1995). Although it is difficult to trace the evolution of promoter regions (for discussion and review, see Wray et al. 2003), our results argue that the primordial mammalian αB-crystallin promoter had two LSR's and two HSE's and that the loss of one of the two LSR's and one of the two HSE's is secondary.

Acknowledgements

We thank Erik Jansen for technical support with the *Xenopus* transgenesis. This investigation was supported by the Research Council for Earth and Life Sciences (ALW) with financial aid from the Netherlands Organization for Scientific Research (NWO) and by the European Commission (TMR ERB-FMRX-CT98-0221 and BMH4-CT98-3895).

References

Adachi, N., and M.R. Lieber. 2002. Bidirectional gene organization: a common architectural feature of the human genome. *Cell* **109**:807-809.

Bell, A. C., A. G. West, and G. Felsenfeld. 1999. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98:**387-396.

Brakenhoff, R. H., R. C. Ruuls, E. H. Jacobs, J. G. G. Schoenmakers, and N. H. Lubsen. 1991. Transgenic *Xenopus laevis* tadpoles: a transient in vivo model system for the manipulation of lens function and lens development. *Nucleic. Acids. Res.* **19:**1279-1284.

Burgess-Beusse, B., C. Farrell, M. Gaszner, M. Litt, V. Mutskov, F. Recillas-Targa, M. Simpson, A. West, and G. Felsenfeld. 2002. The insulation of genes from external enhancers and silencing chromatin. *Proc. Natl. Acad. Sci. USA* **99:**16433-16437.

Burke, L. J., T. Hollemann, T. Pieler, and R. Renkawitz. 2002. Molecular cloning and expression of the chromatin insulator protein CTCF in *Xenopus laevis*. Mech. Dev. **113:**95-98.

Chamberlain, C. G., and J. W. McAvoy. 1987. Evidence that fibroblast growth factor promotes lens fibre differentiation. *Curr. Eye Res.* **6:**1165-1169.

Dermitzakis, E. T., and A. G. Clark. 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* **19:**1114-1121.

Dubin, R. A., R. Gopal-Srivastava, E. F. Wawrousek, and J. Piatigorsky. 1991. Expression of the murine alphaB-crystallin gene in lens and skeletal muscle: identification of a muscle-preferred enhancer. *Mol. Cell Biol.* **11:**4340-4349.

Foo, S. Y., and G. P. Nolan. 1999. NF-κB to the rescue: RELs, apoptosis and cellular transformation. *Trends Genet.* **15:**229-235.

Gopal-Srivastava, R., and J. Piatigorsky. 1993. The murine alphaB-crystallin/small heat shock protein enhancer: identification of alphaBE-1, alphaBE-2, alphaBE-3, and MRF control elements. *Mol. Cell Biol.* **13:**7144-7152.

Gopal-Srivastava, R., and J. Piatigorsky. 1994. Identification of a lens-specific regulatory region (LSR) of the murine alphaB-crystallin gene. *Nucleic Acids Res.* **22:**1281-1286.

Gopal-Srivastava, R., J. I. Haynes, and J. Piatigorsky. 1995. Regulation of the murine alphaB-crystallin/small heat shock protein gene in cardiac muscle. *Mol. Cell Biol.* **15:**7081-7090.

Gopal-Srivastava, R., A. Cvekl, and J. Piatigorsky. 1996. Pax-6 and alphaB-crystallin/small heat shock protein gene regulation in the murine lens. Interaction with the lens-specific regions, LSR1 and LSR2. *J. Biol. Chem.* **271:**23029-23036.

Gopal-Srivastava, R., A. Cvekl, and J. Piatigorsky. 1998. Involvement of retinoic acid/retinoid receptors in the regulation of murine alphaB-crystallin/small heat shock protein gene expression in the lens. *J. Biol. Chem.* **273:**17954-17961.

Hansen, J. J., P. Bross, M. Westergaard, M. N. Nielsen, H. Eiberg, A. D. Borglum, J. Mogensen, K. Kristiansen, L. Bolund, and N. Gregersen. 2003. Genomic structure of the human mitochondrial chaperonin genes: HSP60 and HSP10 are localised head to head on chromosome 2 separated by a bidirectional promoter. *Hum. Genet.* **112:**71-77.

Hough, R. B., A. Avivi, J. Davis, A. Joel, E. Nevo, and J. Piatigorsky. 2002. Adaptive evolution of small heat shock protein/alpha B-crystallin promoter activity of the blind subterranean mole rat, *Spalax ehrenbergi*. *Proc. Natl. Acad. Sci. USA* **99**:8145-8150.

Ishihara, K., and H. Sasaki. 2002. An evolutionarily conserved putative insulator element near the 3' boundary of the imprinted Igf2/H19 domain. *Hum. Mol. Genet.* **11:**1627-1636.

Iwaki, A., T. Nagano, M. Nakagawa, T. Iwaki, and Y. Fukumaki. 1997. Identification and characterization of the gene encoding a new member of the alpha-crystallin/small hsp family, closely linked to the alphaB-crystallin gene in a head-to-head manner. *Genomics* **45:**386-394.

Jansen, E. J., T. M. Holling, F. van Herp, and G. J. Martens. 2002. Transgene-driven protein expression specific to the intermediate pituitary melanotrope cells of *Xenopus laevis*. *FEBS Lett.* **516:**201-207.

Kadonaga, J.T. 2002. The DPE, a core promoter element for transcription by RNA polymerase *II. Exp. Mol. Med.* **34:**259-264.

Kappé, G., E. Franck, P. Verschuure, W. C. Boelens, J. A. M. Leunissen, and W. W. de Jong. 2003. The human genome encodes 10 alpha-crystallin-related small heat shock proteins: HspB1-10. *Cell Stress Chaperones* **8:**53-61.

Klok, E., N. H. Lubsen, C. G. Chamberlain, and J. W. McAvoy. 1998. Induction and maintenance of differentiation of rat lens epithelium by FGF-2, insulin and IGF-1. *Exp. Eye Res.* **67:**425-431.

Kuhn, E. J., and P. K. Geyer. 2003. Genomic insulators: connecting properties to mechanism. *Curr. Opin. Cell Biol.* **15:**259-265.

Labrador, M., and V. G. Corces. 2002. Setting the boundaries of chromatin domains and nuclear organization. *Cell* **111:**151-154.

Meyer, R. G., M. L. Meyer-Ficca, E. L. Jacobson, and M. K. Jacobson. 2003. Human poly (ADP-ribose) glycohydrolase (PARG) gene and the common promoter sequence it shares with inner mitochondrial membrane translocase 23 (TIM23). *Gene* **314:**181-190.

Ohlsson, R., R. Renkawitz, and V. Lobanenkov. 2001. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet.* **17:**520-527.

Otte, D. M., U. Schwaab, and G. H. Luers. 2003. The Pxmp2 and PoleI genes are linked by a bidirectional promoter in an evolutionary conserved fashion. *Gene* **313:**119-126.

Peek, R., R. W. Niessen, J. G. Schoenmakers, and N. H. Lubsen. 1991. DNA methylation as a regulatory mechanism in rat gamma-crystallin gene expression. *Nucleic Acids Res.* **19:**77-83.

Phillips, K., and B. Luisi. 2000. The virtuoso of versatility: POU proteins that flex to fit. *J. Mol. Biol.* **302:**1023-1039.

Schuettengruber, B., A. Doetzlhofer, K. Kroboth, E. Wintersberger, and C. Seiser. 2003. Alternate activation of two divergently transcribed mouse genes from a bidirectional promoter is linked to changes in histone modification. *J. Biol. Chem.* **278:**1784-1793.

Shaw, P. E., and J. Saxton. 2003. Ternary complex factors: prime nuclear targets for mitogen-activated protein kinases. *Int. J. Biochem. Cell Biol.* **35:**1210-1226.

Shin, R., M. J. Kim, and K. H. Paek. 2003. The CaTin1 (Capsicum annuum TMV-induced clone 1) and CaTin1-2 genes are linked head-to-head and share a bidirectional promoter. *Plant Cell Physiol.* **44:**549-554.

Shinya, E., and T. Shimada. 1994. Identification of two initiator elements in the bidirectional promoter of the human dihydrofolate reductase and mismatch repair protein1 genes. *Nucleic Acids Res.* **22:**2143-2149.

Srinivasan, A. N., and S. P. Bhat. 1994. Complete structure and expression of the rat alphaB-crystallin gene. *DNA Cell Biol.* **13:**651-661.

Suzuki, A., Y. Sugiyama, Y. Hayashi, N. Nyu-i, M. Yoshida, I. Nonaka, S. Ishiura, K. Arahata, and S. Ohno. 1998. MKBP, a novel member of the small heat shock protein family, binds and activates the myotonic dystrophy protein kinase. *J. Cell Biol.* **140:**1113-1124.

Swamynathan, S. K., and J. Piatigorsky. 2002. Orientation-dependent influence of an intergenic enhancer on the promoter activity of the divergently transcribed mouse Shsp/alphaB-crystallin and Mkbp/HspB2 genes. *J. Biol. Chem.* **277:**49700-49706.

Takai, D., and P. A. Jones. 2003. The origins of bi-directional promoters – computational analyses of intergenic distances in the human genome. *Mol. Biol. Evol.* Advance Access Epub. Dec. 4.

Wasserman, W. W., M. Palumbo, W. Thompson, J. W. Fickett, and C. E. Lawrence. 2000. Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.* **26:**225-228.

Wistow, G., and C. Graham. 1995. The duck gene for alphaB-crystallin shows evolutionary conservation of discrete promoter elements but lacks heat and osmotic shock response. Biochim. *Biophys. Acta* **1263:**105-113.

Wray, G. A., M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman, and L. A. Romano. 2003. The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20:**1377-1419.

Yoshida S., H. Harada, H. Nagai, K. Fukino, A. Teramoto, and M. Emi. 2002. Head-to-head juxtaposition of Fas-associated phosphatase-1 (FAP-1) and c-Jun NH2-terminal kinase (JNK3) genes: genomic structure and seven polymorphisms of the FAP-1 gene. *J. Hum. Genet.* **47:**614-619.

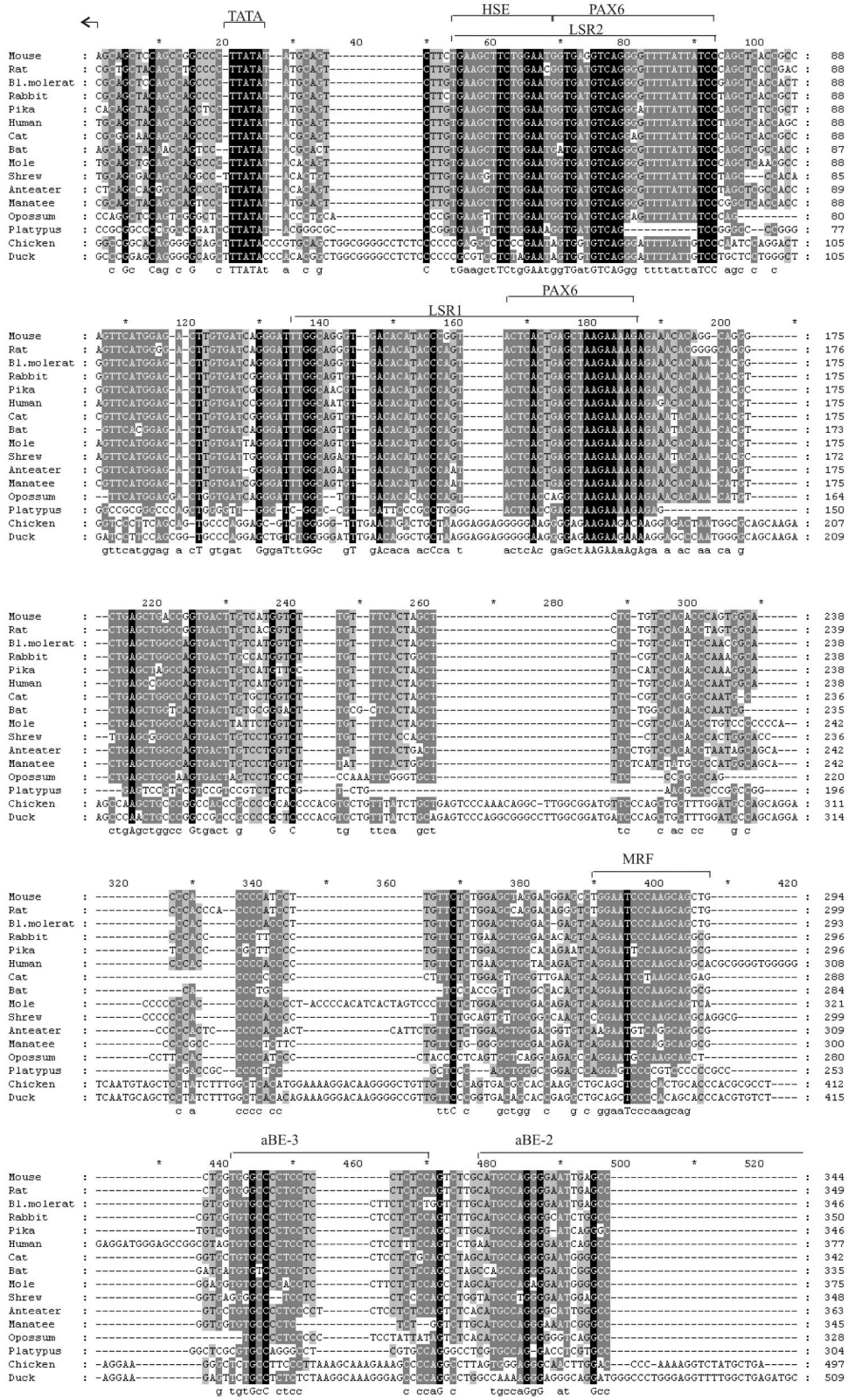
Zhang, L. F., J. H. Ding, B. Z. Yang, G. C. He, and C. Roe. 2003. Characterization of the bidirectional promoter region between the human genes encoding VLCAD and PSD-95. *Genomics* **82:**660-668.

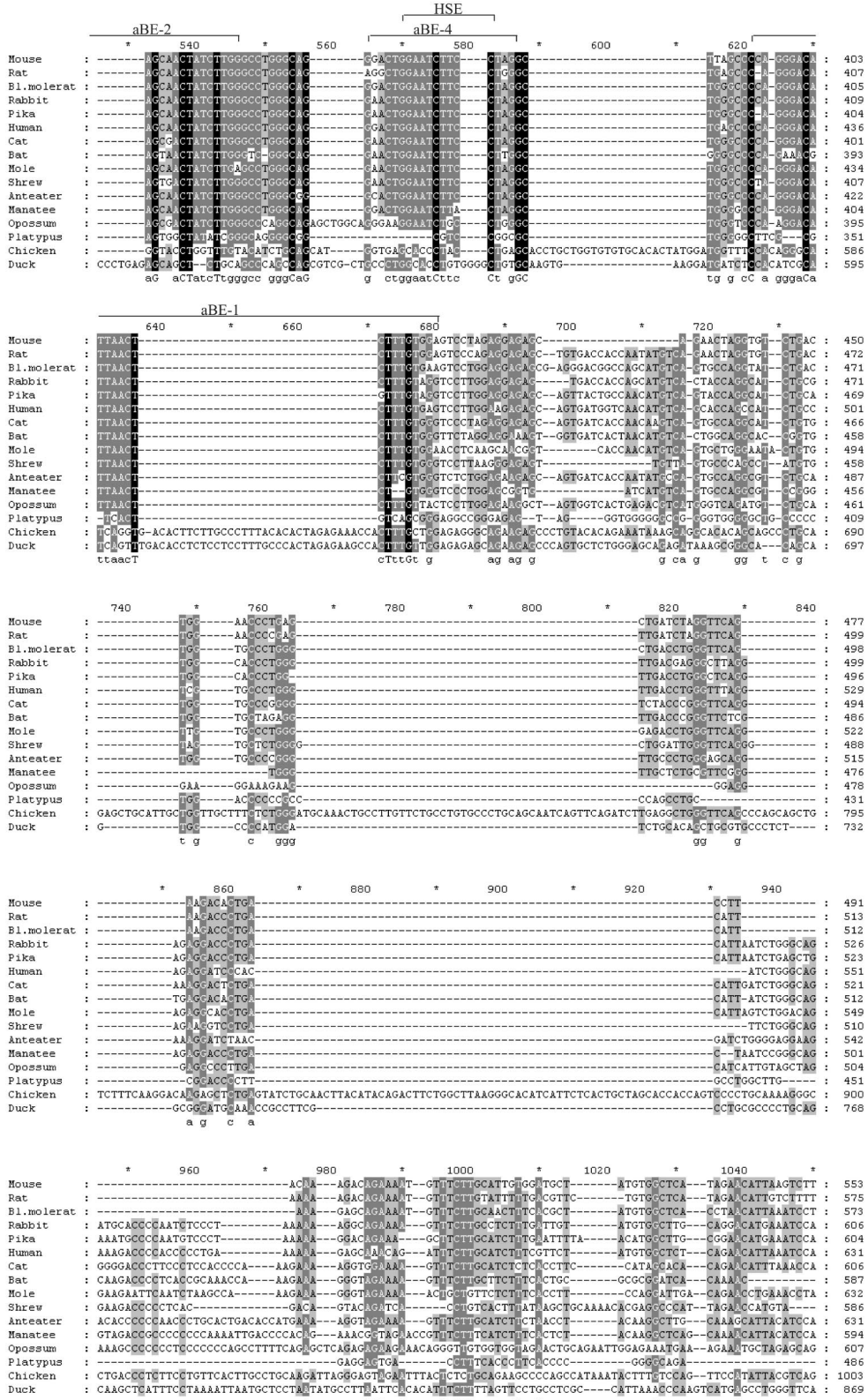
Zhu, J., J. L. Liu, and C. E. Lawrence. 1998. Bayesian adaptive sequence alignment algorithms. *Bioinformatics* **14:**25-39.

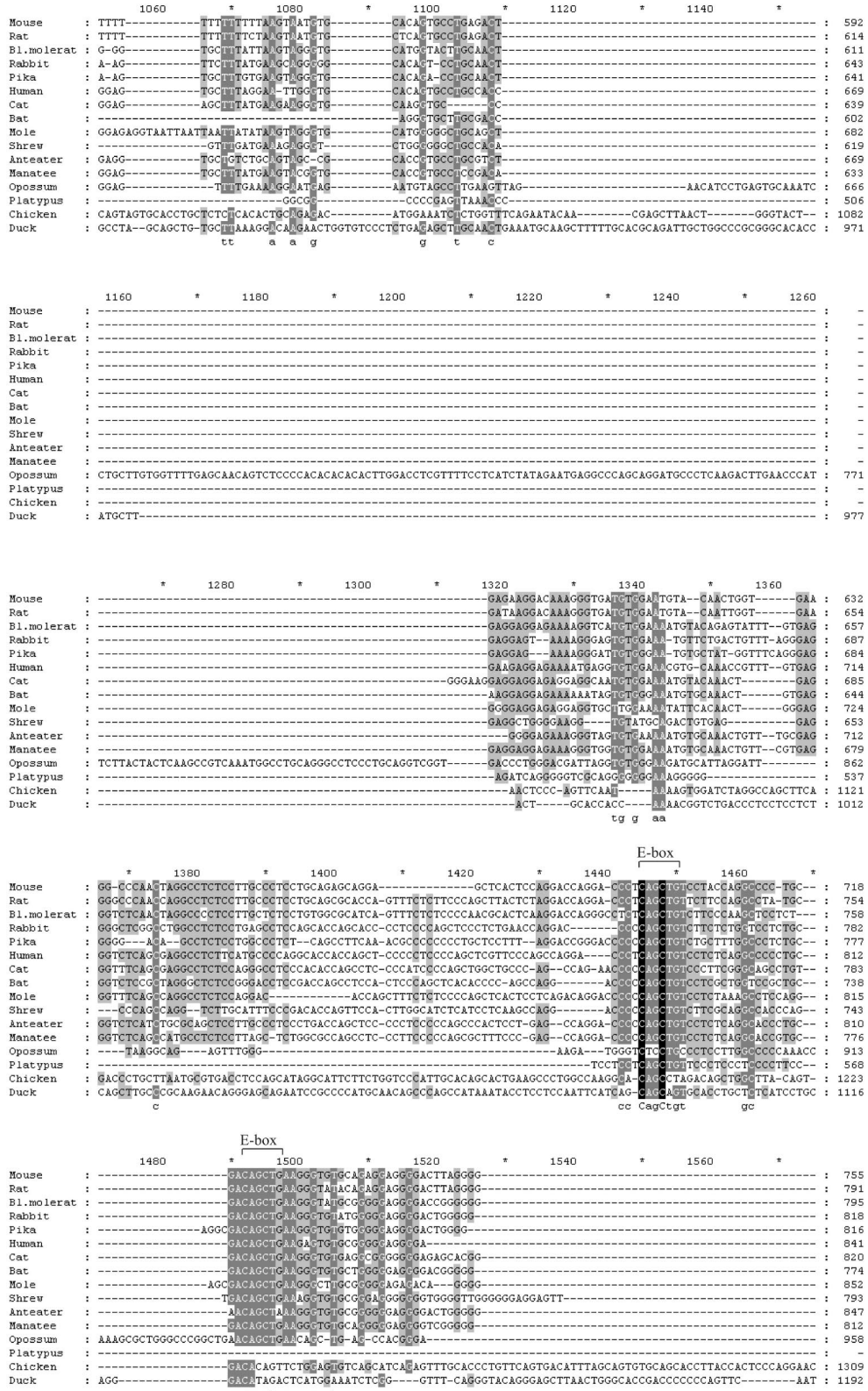
Supplementary Material

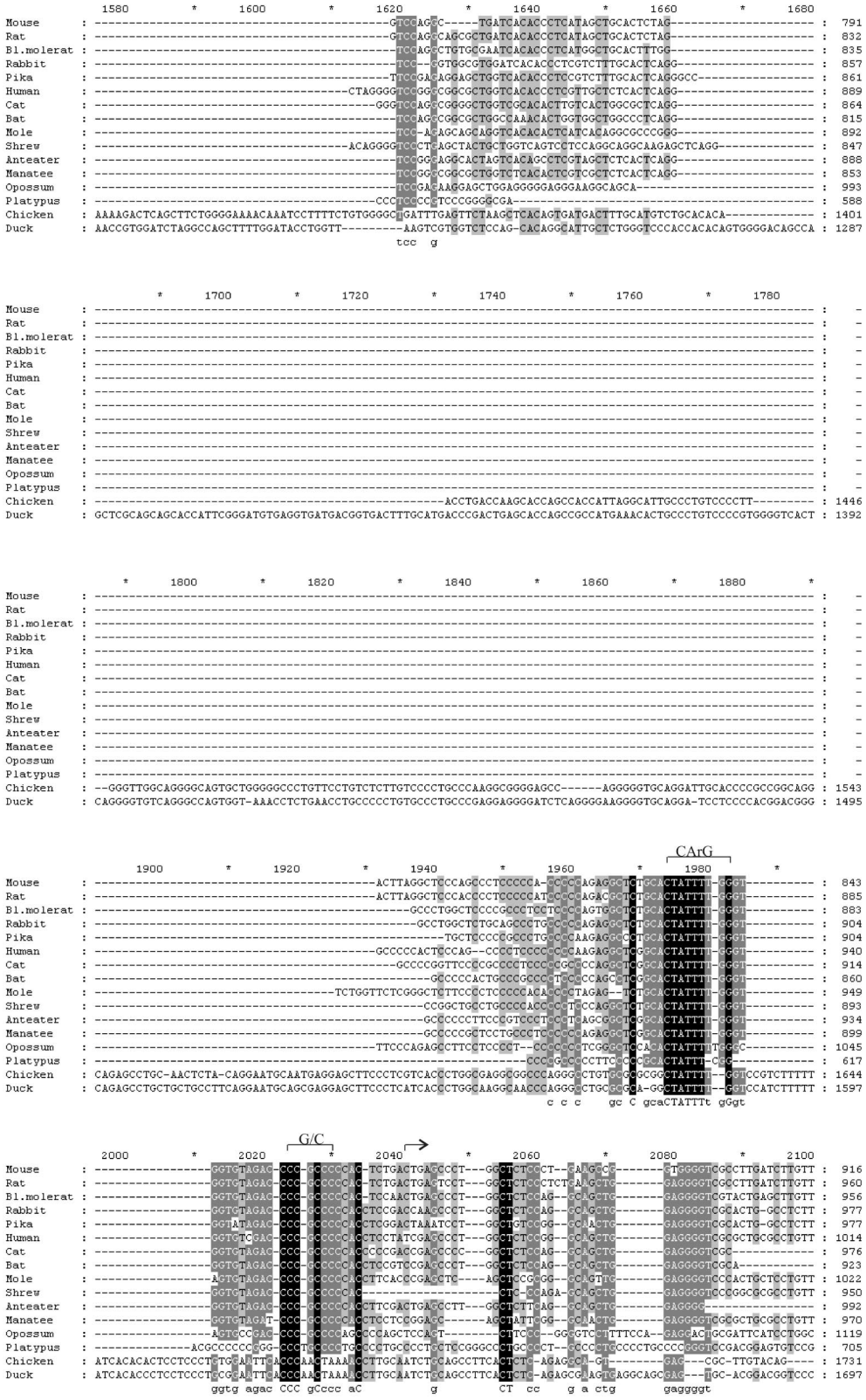
Intergenic Region between αB-Crystallin and HspB2 Genes

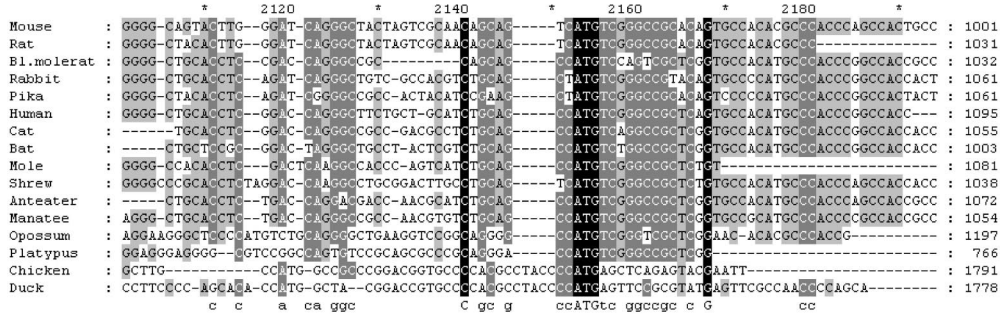
Transcription start sites are indicated with an arrow.
The known regulatory sites of the mouse αB-crystallin promoter are indicated in the figure as follows:
TATA: TATA-box; HSE's: heat shock elements (consensus sequence: inverted repeat of nGAAm);
LSR: Lens specific regions (contain Pax-6 and RAR/RXR binding sites);
MRF: binding site for the muscle response factor, binding to this element increases transcription in muscle
αBE: αB-crystallin elements 1-4, enhance promoter activity in lens, muscle and heart.
The putative regulatory elements of the HspB2 promoter have been identified only on the basis of sequence analysis: E-box: binding site for MyoD family members; CArG-box: binding site for SRF; G/C: G/C-rich promoter region.




HSE
aBE-2
aBE-4
aBE-1


E-box
E-box


CArG
G/C



Chapter 7

Molecular Evolution of the Mammalian Prion Protein

Teun van Rheede, Marcel M. W. Smolenaars, Ole Madsen, and Wilfried W. de Jong

Department of Biochemistry, NCMLS, University of Nijmegen, The Netherlands

Abstract

Prion protein (PrP) sequences are until now available for only six of the 18 orders of placental mammals. A broader comparison of mammalian prions might help to understand the enigmatic functional and pathogenic properties of this protein. We therefore determined PrP coding sequences in 26 mammalian species to include all placental orders and major subordinal groups. Glycosylation sites, cysteines forming a disulfide bridge, and a hydrophobic transmembrane region are perfectly conserved. Also, the sequences responsible for secondary structure elements, for N- and C-terminal processing of the precursor protein, and for attachment of the glycosyl-phosphatidylinositol membrane anchor are well conserved. The N-terminal region of PrP generally contains five or six repeats of the sequence P(Q/H)GGG(G/-)WGQ, but alleles with two, four, and seven repeats were observed in some species. This suggests, together with the pattern of amino acid replacements in these repeats, the regular occurrence of repeat expansion and contraction. Histidines implicated in copper ion binding and a proline involved in 4-hydroxylation are lacking in some species, which questions their importance for normal functioning of cellular PrP. The finding in certain species of two or seven repeats, and of amino acid substitutions that have been related to human prion diseases, challenges the relevance of such mutations for prion pathology. The gene tree deduced from the PrP sequences largely agrees with the species tree, indicating that no major deviations occurred in the evolution of the prion gene in different placental lineages. In one species, the anteater, a prion pseudogene was present in addition to the active gene.

Molecular Biology and Evolution Vol. 20 (01), pp. 111-121, 2003

Introduction

The prion protein (PrP) is associated with the various forms of transmissible spongiform encephalopathies (TSEs) in mammals. Because of its unique features as a pathogenic protein, PrP is probably one of the most intensively studied proteins (reviewed in Prusiner 1998; Harris 1999; Hope 2000; Brown 2001; Collinge 2001; Rudd et al. 2001). Yet, remarkably little is known about its normal cellular function and about the precise manner in which it exerts its pathogenicity. The human prion gene *PRNP* encodes a 253-residue precursor protein (see Fig. 1) in an intronless open reading frame. It is expressed in most tissues, but highest levels are found in the central nervous system, notably associated with synaptic membranes. After translocation across the endoplasmic reticulum membrane, the N-terminal signal peptide is cleaved off. Subsequently, a hydrophobic peptide at the C-terminus is removed in a transamidation reaction which attaches a glycosyl-phosphatidylinositol (GPI) moiety to the prion protein. The GPI anchor attaches the protein to the outer membrane surface of the cell. In addition to the fully translocated form, ^{sec}PrP, two transmembrane variants of PrP have been described, ^{Ctm}PrP and ^{Ntm}PrP, in which a highly conserved hydrophobic sequence (Fig. 1) spans the lipid bilayer in opposite directions (Hegde et al. 1998). NMR measurements have established the conformation of recombinant PrP of mouse (Riek et al. 1996), hamster (James et al. 1997), human (Hosszu et al. 1999), and bovine (Lopez Garcia et al. 2000), which have closely similar global folds. The N-terminal region is largely unstructured and flexible, but residues 37–53 have the potential to form an extended poly(L-proline) II (PPII) helix structure, forming a hydroxylation site at Pro44 (Gill et al. 2000). The N-terminus further comprises a segment of five or six repeats which is implicated in copper binding (Brown et al. 1997). The C-terminal region forms a more rigid globular domain, containing a bundle of three α-helices and a short, two-stranded, antiparallel β-sheet. This domain is stabilized by a disulfide bridge and comprises two variably occupied N-linked glycosylation sites.

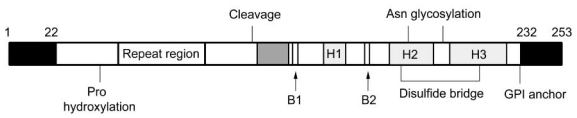


Figure 1 Schematic diagram of human PrP showing signal peptides (black), hydrophobic region (gray), α-helical regions (denoted with H1-3) and ß-strands (denoted with B1-2). Indicated are sites for glycosylation (Asn181, Asn197), hydroxylation (Pro44), and cleavage (Lys112/His113). Cysteines 179 and 214 form a disulfide bridge.

The function of the normal cellular isoform of the prion protein (PrP^c) remains enigmatic. PrP^c-deficient mice develop normally but display minor defects attributable to a higher sensitivity to various forms of stress (Raeber et al. 1998; Brown, Nicholas, and Canevari 2002). PrP^c appears to protect against programmed cell death (Kuwahara et al. 1999) and Bax-mediated apoptosis (Bounhar et al. 2001). PrP^c is a copper-binding protein that may have superoxide dismutase activity. It thus could protect against oxidative damage and contribute to synaptic homeostasis (Brown et al. 1999; Brown 2001). Exposure to Cu^{2+} ions promotes the endocytosis of PrP^c (Sumudhu, Perera, and Hooper 2001). Furthermore, there is evidence that PrP^c is a cell-surface receptor for signal transduction, coupled to the tyrosine kinase Fyn (Mouillet- Richard et al. 2000). PrP^c cycles rapidly between the cell surface and early endosomes via clathrin-coated vesicles, as do many cell-surface receptors. PrP^c functioning may be modulated by removal of the N-terminal domain by proteolysis between Lys112 and His113 (Harris 1999).

A conformational change in PrP^c gives rise to the pathogenic form PrP^{sc}. This transition involves a dramatic increase in β-sheet content from 3% to ~40%, and a decrease in α-helical structure from 40% to ~30% (Cohen and Prusiner 1998). PrP^{sc} is relatively resistant to chemical and heat treatment. Protease digestion leaves a resistant core comprising residues 90–231. PrP^{sc} catalyzes further misfolding of PrP^c, thus leading to a self-amplifying cycle and the formation of insoluble, extracellular aggregates. Inherited, sporadic, and infectious forms of prion diseases exist. Inherited forms are associated with mutations in the prion gene that enhance the transition from PrP^c to PrP^{sc}. In sporadic cases, PrP^{sc} may derive from spontaneous misfolding of PrP^c. Infectivity occurs when the pathological transformation of the host's PrP^c is induced by PrP^{sc} particles transmitted from individuals of the same or different species (Prusiner 1998). It is important to note, however, that PrP^{sc} by itself is not directly neurotoxic (Hill et al. 2000) and that not all prion-diseased brains contain PrP^{sc} (see Stewart and Harris 2001).

Interspecies infectivity of TSEs varies greatly (Prusiner and Scott 1997). Sequence differences between PrP of donor and recipient species play a role, but interspecies susceptibility is not simply determined by overall sequence similarity (Goldmann et al. 1996). More important seem the prion strains within a species, which are isoenergetic conformers of PrP^{sc}, characterized by variations in clinical presentation and protease resistance (Cohen and Prusiner 1998). Strain variation is encoded by different PrP^{sc} conformations and ratios of the three PrP glycoforms (diglycosylated, monoglycosylated, and unglycosylated) and further influenced by PrP sequence polymorphisms and metal binding (Collinge 2001; Priola and Lawson 2001). However, the precise molecular basis of strain variation remains unknown.

Many residues and regions in the prion protein have been implicated in functioning, pathogenicity, and species barrier. Sequence comparison of mammalian prion proteins may help to evaluate such proposals and gain insight in the molecular evolution of PrP. Previous studies have revealed conserved and more variable regions in mammalian PrP, as well as variation in repeat number (Schätzl et al. 1995; Wopfner et al. 1999). It appeared that substitutions implicated in prion diseases occur in the least variable regions and that residues important for the species barrier occur in a restricted region (Krakauer, Zanotto, and Pagel 1998). However, at present, taxon representation of mammalian prion sequences is heavily biased. Sequences are available for many primates, artiodactyls, and a subset of rodents, as well as for rabbit, some carnivores, and perissodactyls. These represent only six of the 18 placental mammalian orders. Moreover, these six orders belong to only two of the four major clades within the Eutheria (Murphy et al. 2001). In this study, we extend the taxon sampling with 26 species to include all 12 previously unsampled orders. Combined with the recent insights in mammalian ordinal relationships (Murphy et al. 2001), it is now possible to better reconstruct the molecular evolution of the mammalian prion.

Materials and Methods

Taxon Sampling

In this study, we included species representing all 18 eutherian orders and some of the major subgroups in speciose orders such as rodents and bats. New *PRNP* gene sequences were obtained for the following orders and species. Afrosoricida: *Amblysomus hottentotus* (Hottentot golden mole), *Chrysochloris stuhlmanni* (Stuhlmann's golden mole, N-terminus only), *Tenrec ecaudatus* (common tenrec); Cetartiodactyla: *Hippopotamus amphibius* (hippopotamus), *Physeter macrocephalus* (sperm whale); Chiroptera: *Cynopterus sphinx* (Indian short-nosed fruit bat), *Macrotus californicus* (California leaf-nosed bat), *Myotis daubentoni* (Daubenton's bat); Dermoptera: *Cynocephalus variegatus* (flying lemur); Eulipotyphla: *Erinaceus europaeus* (Western European hedgehog), *Hylomys suillus* (lesser gymnure), *Sorex cinereus* (masked shrew), *Talpa europaea* (European mole); Hyracoidea: *Procavia capensis* (rock hyrax); Lagomorpha: *Ochotona princeps* (American pika); Macroscelidea: *Macroscelides proboscideus* (shorteared elephant shrew); Perissodactyla: *Diceros bicornis* (black rhino), *Equus equus* (horse); Pholidota: *Manis sp.* (pangolin); Proboscidea: *Elephas maximus* (Asian elephant); Rodentia: *Cavia*

Table 1 Primers used in the amplification and sequencing of the PrP gene

Primer	Sequence (5'-3')[a]	Position[b]
For1	TGA AGT GAC GTG GGC CTC TGY AAR AA	45–70
Seqfor	TGA AGT GAC GTG GGC CTC	45–62
Rev1	CTA TCC CAC TAT GAG RAA RAT NAR RA	736–761
Seqrev	CTA TCC CAC TAT GAG RAA	744–761
S1for	TGC AAG AAG CGN CCR AAR CC	63–82
S2for	GTG GGG GGI CTT GGY GGN TA	363–382
S1rev	GTC TCG GTG AAG TTC TCB CCY T	580–601
S2rev	GCI CCT GCC ACA TGC TTC ATR TTG	320–343
Intrev	CTY CCM CCA GTR TTC CAN CCN CC	84–106
Bio-PrP-rev[c]	CGC TCC CCA GCA TGT AAC CRC CRA RNC CNC C	366–396
PrPflank	TCT TYA TTT TKC AGA BMA GYC RTY A	225–1

[a]B: not A; I: Inosine; H: not G; K: G or T; M: A or C; N: A, G, T or C; R: A or G; Y: C or T.
[b]Position relative to the start of the coding sequence of the human PrP (accession number M13899).
[c]5′-Biotinylated primer.

porcellus (guinea pig), *Sciurus vulgaris* (European red squirrel), *Spalax ehrenbergi* (Ehrenberg's mole rat); Scandentia: *Tupaia tana* (tree shrew); Sirenia: *Trichechus manatus* (manatee); Tubulidentata: *Orycteropus afer* (aardvark); Xenarthra: *Cyclopes didactylus* (silky anteater). Sequences were submitted to GenBank with the accession numbers AY133034–AY133063.

From GenBank, we extracted *PRNP* sequences of representatives of two other eutherian orders: Carnivora (*Mustela sp.*, mink, S46825) and Primates (*Homo sapiens*, human, M13899; *Saimiri sciureus*, squirrel monkey, U08310.1) and of additional Cetartiodactyla (*Bos taurus*, cow, AJ298878; *Sus scrofa*, pig, L07623; *Camelus dromedarius*, camel, Y09760; *Ovis aries*, sheep, M31313.1), Lagomorpha (*Oryctolagus cuniculus*, rabbit, U28334), and Rodentia (*Mus musculus*, mouse, M13685), as well as a marsupial outgroup (*Trichosurus vulpecula*, brush-tailed possum, L38993). The available *PRNP* sequences of perissodactyls and dolphin (Wopfner et al. 1999) do not represent the complete mature protein and were therefore not used. A prion gene sequence of the guinea pig is present in the database (AF139166) but contains a conspicuous deletion in the N-terminus. We therefore determined an independent guinea pig sequence (acc. nr. AY133039) and found the same deletion. For the dog (*Canis familiaris*), highly dissimilar *PRNP* sequences are present in GenBank, of which AF042843 and AF022714 group with artiodactyls in phylogenetic analyses, whereas a partial dog sequence (AF113937) and close relatives of the dog (dingo, AF113937; gray wolfe, AF113939) group with other carnivores. Similarly, two different cat (*Felis catus*) *PRNP* sequences are available, AF003087 grouping with artiodactyls and Y13698 grouping with carnivores again. This casts doubt on the reliability of available dog and cat sequences.We therefore choose the mink prion sequence as a genuine carnivore representative in our analysis. This sequence is corroborated by phylogenetic analysis, grouping with wolf, dingo, and the closely related ferret and polecat sequences.

PCR, Cloning, and Sequencing

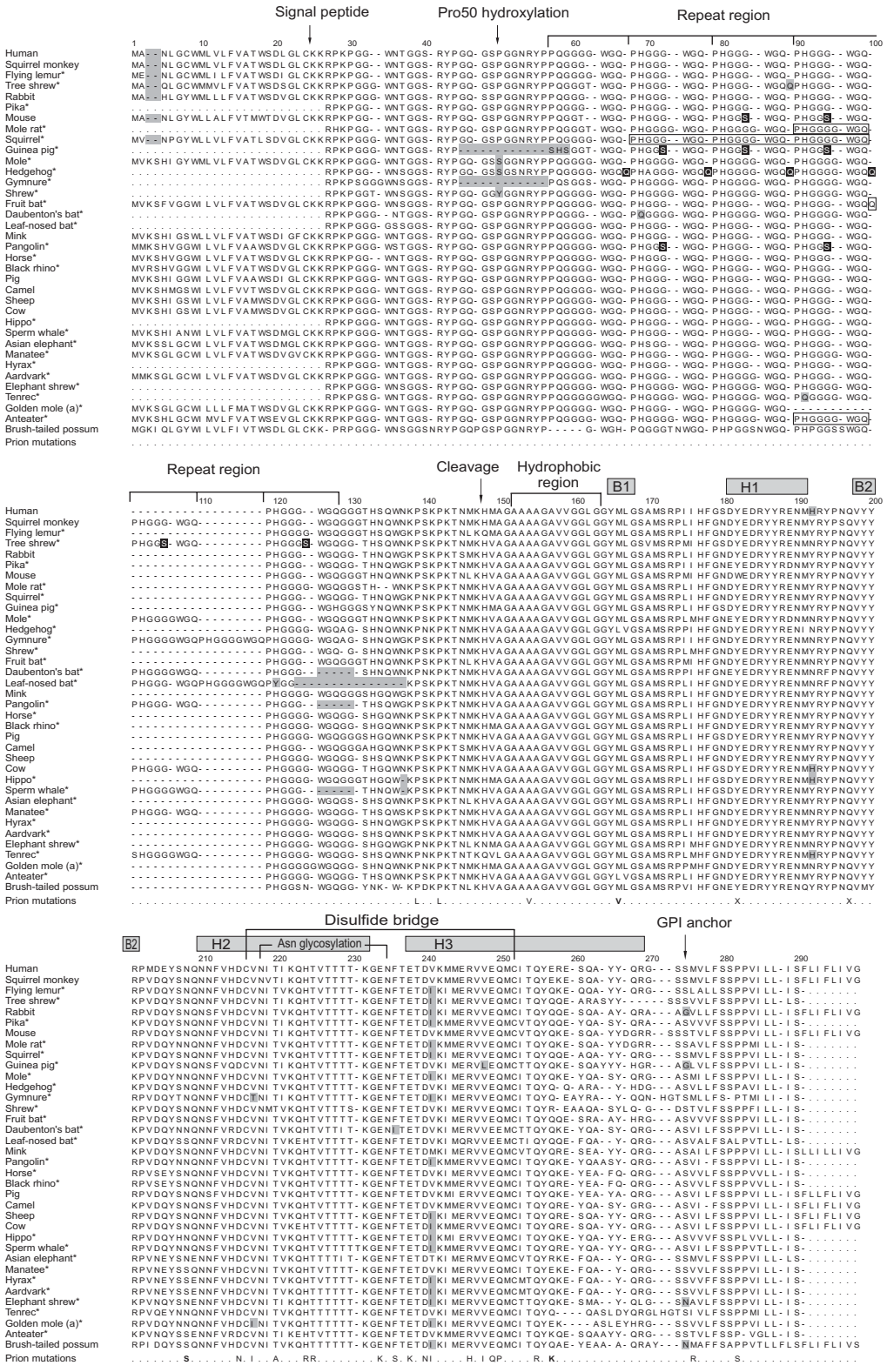
Amplification of a ±700-bp fragment of the *PRNP* gene was performed with primers based on known sequences coding for the N- and C-terminal signal peptides (Table 1). This yields the open-reading frame of the complete mature protein (positions 23–231 in Fig. 1), apart from the first two amino acids. PCR reactions contained approximately 100 ng genomic DNA, 375 mM dNTPs (Boehringer Mannheim), 20–100 pmol of each primer, and 0.5 µl Taq Expand polymerase (Expand HF system, Roche Diagnostics) in a final volume of 50 µl. The PCR program was 95°C for 4 min, followed by 35 cycles at 94°C for 60 s, annealing at 56°–60°C for 60 s, and extension at 68°C for 90 s. Gel-extracted PCR fragments (Amersham Pharmacia gel extraction kit) were sequenced

directly or cloned in pGEM-T easy vector (Promega). Sequencing reactions were performed using Big Dye fluorescent technology and run on an ABI 3700 96-capillary sequencer. Sequences were obtained from at least two independent PCR reactions, and all bases were sequenced at least once on both strands.

Sequences coding for the N-terminal signal peptide were determined with a PCR technique designed to amplify a fragment of unknown flanking sequence (Sørensen et al. 1993). Partially randomized primers are used along with a specific biotinylated primer. Subsequent purification of the biotinylated PCR product with streptavidine beads (Dynal) and a second, nested PCR increase the specifity. In our case, the biotinylated primer bio-PrPrev and the nested primer S2rev were used (Table 1). Applying this method, 59 coding and noncoding sequences were obtained for *Cynopterus sphinx* and *Elephas maximus*. This sequence information was used to design a highly degenerate primer (PrPflank, table 1), ending with the adenine of the start codon. This primer was used to amplify N-terminal signal peptide sequences of 12 other species (see Fig. 2 for names). The PCR program was 95°C for two min, followed by 35 cycles at 94°C for 30 s, annealing at 50°C for 30 s, and extension at 68°C for 40 s.

Phylogenetic Analysis

DNA sequences were analyzed using the Staden package programs PreGAP4 and GAP4 (http://www.mrclmb. cam.ac.uk/pubseq/). Nucleotide and amino acid alignments were produced using ClustalW and adjusted manually. Positions that were ambiguous in the alignment were excluded from phylogenetic analysis; the tree in figure 3 is based on nucleotide sequences (570 bp) corresponding to amino acid positions 27–67, 132–263, and 274–290 in figure 2. We applied maximum likelihood (ML) and Bayesian posterior probability analyses to reconstruct phylogenetic trees. We used Modeltest 3.06 (Posada and Crandall 2001) to determine which model of sequence evolution had the best fit to the data under the maximum likelihood assumption. The best model was a general time reversible model with gamma distribution (eight categories) and proportion of invariable sites (GTR+G8+I). This model was used in all analyses. Model parameters for ML were estimated on an NJ tree and subsequently refined in two consecutive rounds of heuristic ML tree searches with the previously found tree as starting tree. The best model parameters were used in a nonparametrically bootstrapped ML search. To search for the best ML tree, five heuristic tree searches were performed: four with different random starting trees and one with the "best ML tree" from the search for optimal model parameters as a starting tree. All searches converged to the same tree. Bootstrap analyses included 100 replicates. In all ML analyses, the tree bisection reconnection branch swapping option was used. The analyses were performed with PAUP* 4.0 (Swofford 2002). Bayesian phylogenetic analyses were performed using the program MrBayes 2.1 (Huelsenbeck and Ronquist 2001). The Metropoliscoupled Markov chain Monte Carlo (MCMCMC) sampling approach was used to calculate posterior probabilities. Prior probabilities for all trees were equal and starting trees were random. To check consistency of results, four Markov chains were run simultaneously for 200,000 and 500,000 times. Tree sampling was every 10 generations, and "burn-in" values were determined from the likelihood values.



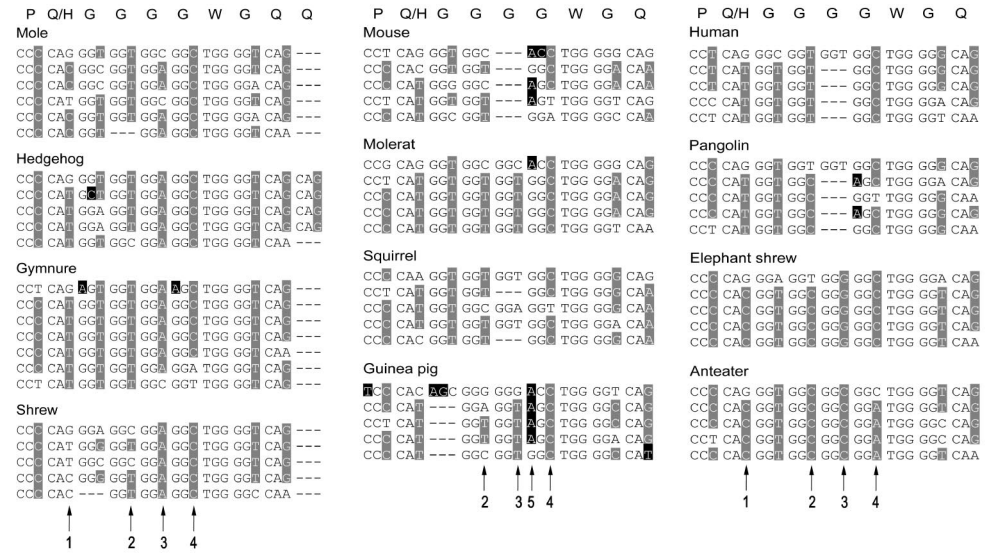


Figure 3 Homogenization of DNA sequences in eutherian prion repeat regions. Sequences of the five to seven repeats in four eulipotyphlans (left), four rodents (middle), and four representatives from other eutherian orders (right) are aligned to show that codons at corresponding positions in a repeat are more homogeneous within than between species. Replacements are colored according to the majority consensus rule. Synonymous substitutions are in gray; nonsynonymous substitutions are in black. Arrows indicate evidence for homogenization, as outlined in the Results.

Results

Sequencing the *PRNP* Gene

PCR primers were designed on the sequences coding for the N- and C-terminal signal peptides of known eutherian prion proteins. This allowed us to amplify in a single reaction the DNA coding for the mature protein of 26 eutherian species. In 23 specimens, a single amplification product was obtained, while three specimens (squirrel, mole rat, and anteater) yielded a double band, due to length polymorphisms in the repeat region. In anteater, a pseudogene was amplified in addition to the normal *PRNP* gene. This pseudogene contains two frame shift mutations, three stop codons, and many nonsynonymous substitutions (data not shown; acc. nr. AF545183). Since published primate and rodent PrP sequences show a deletion of two residues in their N-terminal signal peptide (Schätzl et al. 1995; Wopfner et al. 1999), we further assessed the taxonomic distribution of this deletion. This was done for 14 species, representing the major superordinal clades, using a modified PCR technique (see Materials and Methods). The amino acid sequences deduced from the newly determined *PRNP* genes are aligned in Figure 2, together with sequences selected from the databases, to represent all major eutherian clades.

Figure 2 (left) Alignment of mammalian prion protein sequences. Important structural features are marked above the alignment (cf. Fig. 1). The numbering deviates from that in Figure 1, because of the introduction of gaps (—). Not in all species N- and C-terminal sequences were determined (--). Black shading indicates residues that provide evidence for repeat homogenization. Gray shading denotes sequence characteristics that are discussed in the text. Repeats are flushed left (because repeat homogenization does not allow for meaningful alignment), apart from the last one (to facilitate comparison of the truncated last repeat in bats, pangolin, and sperm whale). Boxed repeat units denote repeat number polymorphisms in mole rat (4 and 5 repeats), squirrel (2 and 5 repeats), and anteater (4 and 5 repeats). Fruit bat has alleles with and without Gln. Mutations related to spongiforme encephalopathies and polymorphisms (in bold) are shown under the alignment. Asterisks (*) indicate newly determined sequences. (a) The sequence of the golden mole N-terminal signal peptide is of *C. stuhlmanni,* the mature protein sequence of *A. hottentotus.*

Characteristics of the Mammalian Prion Protein

Starting from the N-terminus and referring to the position numbering as used in Figure 2, the following features are noteworthy. Two length variants of the N-terminal signal peptide can be observed amongst placental mammals. The longer one is present in most orders, as well as in the outgroup marsupial, and starts with the consensus sequence MVKSH in the placentals. The shorter variant, with the sequence MAN, is found in primates, flying lemur, tree shrew, rabbit, and rodents, which form the recently recognized clade Euarchontoglires (Murphy et al. 2001). The residues flanking the signal peptide cleavage site, between Cys24 and Lys25, are perfectly conserved; the presence of Gly22 and Cys24 at positions -3 and -1 before the cleavage site agrees with the consensus residues for the cleavage enzyme (Udenfriend and Kodukula 1995).

Immediately before the repeat region, similar deletions occur from positions 45 to 56 in guinea pig (but not in other rodents) and in gymnure (but not in the related hedgehog). These deletions are probably caused by independent unequal crossing-over events between the repeated sequence coding for GGSRYP at positions 38–44 and GGNRYP at 51–56. The deletions remove Pro50, of which 4-hydroxylation is thought to be an important functional feature (Gill et al. 2000). Pro50 is also absent in mole, hedgehog, and shrew, being replaced by serine or tyrosine.

The number of repeats varies from two (in one of the squirrel alleles) to seven (gymnure and leaf-nosed bat). The squirrel specimen included in this study was heterozygous for alleles with two and five repeats, and both mole rat and anteater had alleles with four and five repeats. Truncated repeats are present in leaf-nosed and Daubenton's bat, pangolin, and sperm whale (positions 126–130). In the latter three species, the deletion may have been triggered by Gly runs on both sites of the WGQ triplet as present in the last repeat in other placentals. The eutherian repeats rigidly conserve the consensus sequence P(Q/H)GGG(G/-)WGQ. The first repeat always has Q and the following ones have H at position 2, except an incidental repeat in Daubenton's bat and tenrec. Conspicuous are the deviating first repeat in guinea pig with the highly conserved N-terminal PQG replaced by SHS (positions 57–59). The first repeat generally has a GGGG track. GGG is common to most other repeats, but GGGG runs and even GGGGG runs do occur.

Occasional deviations from the consensus repeat sequence are indicative of concerted evolution (indicated with a black background in Fig. 2). The first four repeats of hedgehog PrP have a duplication of the C-terminal Glutamine (Q). The repeats of the closely related gymnure have no such extra Q, suggesting homogenization of the repeats after the divergence of gymnure and hedgehog. An incidental extra Q also occurs in tree shrew and in one allele of the fruit bat. Other replacements occur mostly in the Gly runs and are also suggestive of repeat homogenization (e.g., Gly to Ser in tree shrew, mouse, guinea pig, and pangolin). Homogenization of the repeats is also apparent at the DNA level. In Figure 3 the repeat units within a number of species are aligned to emphasize that codon usage at corresponding positions in the repeats is more similar within than between species. For example, the His residues are generally encoded by CAT, but CAC is used in the mole, elephant shrew, and anteater (Fig. 3 arrow 1). In eulipotyphlans and rodents, the second Gly is generally encoded by GGT, but by GGC in bat, elephant shrew, and anteater (arrow 2). The third Gly in eulipotyphlans is encoded by GGA, but by GGT in rodents, GGG in elephant shrew, and GGC in anteater (arrow 3). Finally, almost all Gly residues preceding the Trp are encoded by GGC, except in anteater, which uses GGA (arrow 4). At the first position of the same codon, mouse and guinea pig show expansion of the nonsynonymous substitution G→A (arrow 5). In addition to homogenization, similarity by descent certainly plays a role in structuring the repeats. For example, insectivore repeats are more similar to one another than to the repeats of more distantly related species (Fig. 3).

Next to the repeat region, the sequence 143–163 (Fig. 2) includes the hydrophobic transmembrane segment. It is highly conserved, and the Ala-Gly-rich region even perfectly so. In the structured

domain comprising the α-helices and β-strands, very limited and almost exclusively conservative replacements are observed. Cys216 and Cys252, involved in the disulfide-bridged helix-loop-helix motif H2-H3, are strictly conserved, as are the Asn-X-Thr motifs required for N-glycosylation of Asn218 and Asn235. Considerable variation occurs in the region immediately preceding the GPI anchor site. This region forms a flexible linkage to the GPI anchor (Riek et al. 1996; Liu et al. 1999). The GPI-attachment Ser (position 275) is replaced by Gly in rabbit and guinea pig and by Asn in elephant shrew. However, the requirements for the transamidase reaction by which the GPI moiety is attached are sufficiently flexible that either these small residues or adjacent ones can serve as attachment sites (Udenfriend and Kodukula 1995). Also the few replacements in the C-terminal signal peptide do not interfere with the required hinge (including the prolines 282 and 283) and stretch of hydrophobic residues.

Gene Tree of Mammalian PrP

Strongly supported discrepancies between a gene tree and the corresponding species tree may point to interesting features of the evolution of the gene in a particular lineage. We therefore performed phylogenetic analyses on the *PRNP* sequences. In preliminary analyses, rooting the tree with the single marsupial sequence rendered some species highly unstable and often located at implausible places. Notably, the position of the root was consistently placed within Afrotheria, probably due to deviating base compositions and resulting long-branch attraction. To minimize these problems, unrooted analyses were performed with exclusion of the marsupial sequence. The ML tree shown in Figure 4 reflects the most constant and prominent findings. The branch lengths reflect accelerated rates of substitutions in some species and clades, such as shrew, erinaceids, tenrec, elephant shrew, and most rodents. The best-supported nodes in the tree mostly correspond with unquestioned sister taxa in the data set (hedgehog and gymnure; Daubenton's and leaf-nosed bats; horse and rhino; sheep and cow; mouse and mole rat; human and squirrel monkey). The prion tree supports the nesting of whale with hippo and ruminants in Cetartiodactyla, a molecularly now well-established relationship (Gatesy and O'Leary 2001). Of phylogenetic interest is the support for the grouping

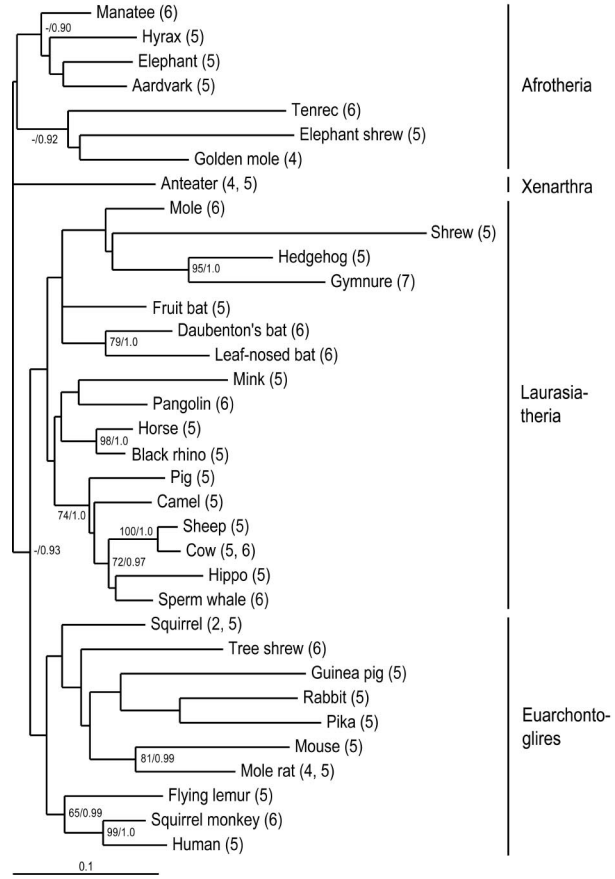


Figure 4 Unrooted maximum likelihood tree based on eutherian prion genes. ML nonparametric bootstrap values and Bayesian posterior probabilities are shown when higher than 50 and 0.90, respectively (ML/Bayesian). In brackets: The repeat numbers observed in the newly determined PrP sequences or reported as most common for the other sequences. The four recently recognized major clades of eutherian mammals (Murphy et al. 2001) are indicated. The bar corresponds with 0.1 (10%) substitutions per site. For further details, see *Materials and Methods*.

of flying lemur with primates, where the largest current concatenated data set suggests that flying lemur is the sister of tree shrew (Murphy et al. 2001). Encouraging is the observation that the prion tree confirms, be it with low support, the recently recognized major superordinal clades Laurasiatheria, Euarchontoglires, and Afrotheria. The only better-supported discrepancy between the prion tree and the species tree is the nesting of aardvark within the paenungulates, although this was seen only in Bayesian analysis. Other oddities, such as the nesting of tree shrew and lagomorphs within the rodents, are only poorly supported.

Comparing the indel signal in the amino acid alignment in Figure 2 with the tree in Figure 4, we note that the short form of the N-terminal PrP signal peptide as observed in Euarchontoglires is concordant with the tree topology. Of some interest, too, is the unique deletion of residue 137 in a well-conserved motif QW×KP (135–139) in hippo and sperm whale. Although molecularly almost unanimously supported, morphological evidence remains ambiguous about a whale-hippo clade (Gatesy and O'Leary 2001). On the other hand, comparing the alignment with the tree also reveals that indels in vulnerable repeat regions (e.g., positions 45–56 in guinea pig and gymnure, and 126–130 in microbats, pangolin, and sperm whale) can occur in parallel.

Discussion

Relevance to Normal Function and Structure of PrP^c

Previous comparative studies established the limited sequence variation of the prion protein between primates, artiodactyls, and rodents (Schätzl et al. 1995; Wopfner et al. 1999), mostly occurring in the least hydrophobic regions (Krakauer, Zanotto, and Pagel 1998). Our sequences broaden the representation of the wild-type variation amongst the eutherian *PRNP* genes. Most of the sequence conservation can readily be understood in relation to the normal structure and functioning of PrP^c. This concerns primarily the sequences that are required for N- and C-terminal processing of the nascent protein and attachment of the GPI anchor. The latter is essential to cluster PrP^c in sphingolipid-sterol microdomains or rafts (Muniz and Riezman 2000). The strict conservation of Lys147 is consistent with a functional role of the cleavage occurring at this site (Harris 1999) (NB: residue numbering throughout the discussion refers to Figure 2). Conservation is also obvious for the residues that are involved in maintaining the tertiary structure of the C-terminal domain. In this domain, the disulfide bond linking Cys216 and Cys252 is essential for the structure of PrP^c; replacement of these residues results in insolubilization (Maiti and Surewicz 2001). Asn218 and Asn235 must be conserved because the large size and dynamic properties of the two N-linked sugars protect large regions of the extracellular PrP surface from proteases and nonspecific protein interactions (Rudd et al. 2001). In addition, the oligosaccharides may direct the folding and routing of nascent PrP^c in the ER (Priola and Lawson 2001). The strict conservation of residues 151–165 suggests an essential function for this hydrophobic region. These residues form the major part of the transmembrane region in ^{Ctm}PrP and ^{Ntm}PrP. In in vitro systems, ^{sec}PrP and ^{Ntm}PrP are equally abundant (40%–50%), while ^{Ctm}PrP forms the remaining 10% (Hegde et al. 1998). It is likely that synthesis of the three topological forms varies in different cell types and may be influenced by sequence differences. Indeed, prion disease–related mutations in or near (but not outside) the transmembrane region enhance the formation of ^{Ctm}PrP (Stewart and Harris 2001). Not all sequence characteristics implicated in functioning of PrP^c are conserved. Hydroxylation of Pro50 was suggested to be an epigenetic control mechanism of normal PrP functioning (Gill et al. 2000). The absence in several species of this residue questions the universal importance of this modification.

The observed variation in repeat number and sequence is of direct relevance to the property of PrP to bind bivalent metal ions, in particular copper (Brown et al. 1997), and its possible role in copper transport and metabolism (Pauly and Harris 1998). PrP can bind up to four Cu^{2+} ions to

imidazol nitrogens of the histidines in the repeat region and likely bind a fifth copper to His133 and His148 (Viles 1999; Kramer et al. 2001). However, only two of these might be physiologically relevant high-affinity binding sites (Jackson et al. 2001). Copper binding is pH-dependent, which makes PrP suitable for internalizing Cu^{2+} ions (Miura et al. 1999; Viles et al. 1999; Jackson et al. 2001). Indeed, exposure of cells to physiologically relevant concentrations of copper leads to rapid endocytosis of PrP^c. This cellular response is abolished when copper binding is hindered by mutagenesis of histidines or deletion of octarepeats (Sumudhu, Perera, and Hooper 2001). Binding of copper ion adds structure to the flexible repeat region (Viles et al. 1999) and is essential for the superoxide dismutase-like activity of PrP (Brown et al. 1999). Finally, copper ions lead to site-specific cleavage at the repeat region of mouse PrP on exposure of cells to reactive oxygen species (McMahon et al. 2001). Since copper binding appears to be an important modulator of the functioning and processing of PrP^c, it is of interest that not all the implicated histidines are conserved (Fig. 2). The number of potential copper-binding histidines in the repeat region can be as low as one, as in the two-repeat squirrel allele. Also the copper-binding site around positions His133 and His148 is not perfectly conserved in all species. These findings raise the question whether and how many copper ions must be bound for the normal functioning of PrP^c.

The wild-type number of repeats is five or six in almost all eutherian species, and a species such as cattle is polymorphic for five or six repeats (Schätzl et al. 1995; Wopfner et al. 1999) (Figs. 2 and 4). Alleles with four repeats are found at frequencies of up to 2% in human populations (Puckett et al. 1991) and equally occur in other primates (Schätzl et al. 1995). The finding in this study of an animal homozygous for four repeats (golden mole) suggests that this is compatible with normal functioning of PrP^c. Alleles with three repeats are common in goat (Goldmann et al. 1998). Even alleles with only two repeats have been reported in lemur (Gilch, Spielhaupter, and Schätzl, 2000) and here in squirrel. This suggests that this low number is evolutionarily viable, at least in heterozygotes. The fact that in our study homozygous individuals with seven repeats (gymnure and leaf-nosed bat) have been reported, demonstrates that this expansion, too, is within the normal range.

It appears that reduction and increase of the number of repeats, between two and seven, does not follow any phylogenetic pattern (see Fig. 4). This actually is to be expected in view of the observed repeat number polymorphisms within various species. Expansion and contraction of repeats clearly is a frequent mutational process in the eutherian prion gene. The mechanisms involved can be unequal crossing-over and replication slippage (Collinge 2001). This will simultaneously lead to homogenization of substitutions in the repeat sequences and to the length variation of the GGG runs in the repeats. Selection likely plays a role in balancing the repeat number: a high enough repeat number is needed for copper ion binding, but too high a repeat number promotes the early onset of prion disease (see below).

Relevance to Prion Diseases and Species Barrier

In relation to the various regions and residues in PrP^c that have been implicated in prion pathology, the observed variety of the eutherian PrP sequences is informative, too. It should be kept in mind, however, that our sequences may not reflect all intraspecies sequence variation because of our limited sampling within species. In human, Creutzfeldt-Jakob disease (CJD), Gerstmann-Sträussler-Scheinker disease (GSS), and kuru form part of the phenotypic spectrum of the prion diseases (Collinge 2001). More than 20 amino acid replacements have been observed in inherited prion diseases (Fig. 2, "prion mutations") (Prusiner 1998; Collinge 2001; see also SWISSPROT entry P04156). All mutations with pathological significance occur either within or adjacent to regions of secondary structure, notably associated with the second and third α-helix (Krakauer, Zanotto, and Pagel 1998) and in most cases appear to destabilize the PrP structure (Prusiner 1998; Liemann and Glockshuber 1999). Consequently, replacements that are associated with human prion diseases are only rarely observed in our eutherian sequences. The CJD-related mutation V217I is present in

golden mole, and V241I is even found in most of the eutherian sequences. It is unlikely that these conservative replacements cause prion disease by themselves. Also some other replacements are observed at positions implicated in human prion disease: V217T is present in gymnure, F236I in Daubenton's bat, V241M in mink, V241T in elephant, V248L in guinea pig, and Q249E in microbats. Although the C-terminal signal peptide is quite variable, the disease-related mutations M276R and P282S are not observed in any species.

Some amino acid polymorphisms in PrP of human, sheep, and mouse appear to influence the onset and phenotype of prion disease (Prusiner 1998; Collinge 2001). Notably, position 166 (129 in human) is polymorphic for Met or Val in human PrP and modulates protease sensitivity of PrPsc (Parchi et al. 2000). Heterozygosity at this site has been proposed to have a protective effect against sporadic and acquired prion diseases and in some of the inherited forms (Collinge 2001). All cases to date of vCJD, the novel human variant caused by the BSE prion strain from cattle, are homozygous for Met166. The fact that all other mammalian PrP sequences were found to have Met at this position (apart from Leu in hedgehog and anteater) suggests that the assumed protective effect of codon 166 heterozygosity is not a general feature. The other polymorphic sites known in man are N208S and E257K, and these residues occur at these positions in one or more other species.

Various regions in PrP^{c} have been proposed as critical to prion formation, all located in that part of the prion protein that is structured in NMR analysis and conserved in eutherians. Residues 127–157 play a major role in the PrP^{c}/PrP^{sc} interface (Cohen and Prusiner 1998), but also the surface regions 156–175, ~202–211, and 246–264 may be involved in initial binding of PrP^{c} to PrP^{sc} (Horiuchi et al. 2000). An antibody directed against residues 169–193 blocks this interaction (Peretz et al. 2001). Interestingly, the isolated peptide 164–201 can adopt two isoenergetic conformations, with all β or αβ structures, an essential feature of the conformational change of PrP^{c} into PrP^{sc} (Derreumaux 2001). Residues 205, 209, 254, and 258 (Q167, Q171, T214, and Q218 in mouse) PrP^{c} have been postulated to bind an auxiliary molecule, protein X, thought to facilitate formation of PrP^{sc} (Kaneko et al. 1997). None of these residues is perfectly conserved in our sequences. The species specificity of protein X might contribute to the species barrier by promoting or slowing down prion formation (Prusiner 1998).

The role of the repeat region in the transition of PrP^{c} to PrP^{sc} is enigmatic. Transition can still proceed after deletion of all repeats, although the repeat region modulates prion replication and pathogenicity (Flechsig et al. 2000). Extension of the normal number of five repeats with one to nine copies has been observed in human prion disease kindreds (Collinge 2001). The higher numbers of repeats are associated with earlier onset of the disease. Reduction of the repeat number to four does not lead to prion disease, but heterozygosity for three repeats was reported in an elderly patient suffering from a rapidly progressive dementia consistent with CJD (Beck et al. 2001). On the other hand, in goats it was observed that animals with three repeats succumbed after unusual long periods when challenged with PrP^{sc} (Goldmann et al. 1998). Heterozygosity for two and five repeats was found in two specimens of different lemur species, which are known to be particularly susceptible for TSE (Gilch, Spielhaupter, and Schätzl 2000). Alleles with two repeats likely are common in lemurs and squirrel, and therefore homozygotes may also occur. However, it is not known whether homozygotes are viable. Our finding of animals that are homozygous for PrP alleles with four and seven repeats demonstrates that this in itself cannot be deleterious. Although repeat numbers of two to seven are thus evolutionarily viable, it is possible that in humans such repeat numbers contribute to higher prion disease susceptibility at older, postreproductive age.

Finally, what can our sequence comparisons tell about the TSE species barrier? As mentioned, prion strain variation is probably the most important factor. Inoculated prions preferentially convert PrP^{c} into one of the thermodynamically favored PrP^{sc} conformers. Species transmission barriers may be determined by the degree of overlap between the subset of PrP^{sc} conformers allowed

by the host's PrP with that presented by the donor PrP^{sc} (Hill et al. 2000). Strain variation is only partially determined by sequence differences. Riek et al. (1996) pinpointed residues 175, 180, 182, 192, and 204 as important for the mouse-human barrier, but also residues 221, 223, 243, and 245 have been proposed to form an epitope involved in controlling the species barrier (Scott et al. 1997). Several of these residues are perfectly conserved and therefore will not contribute to the species barrier. In other instances, only two or three character states are observed (e.g., positions 175, 180, and 203 in Fig. 2), which are distributed without obvious phylogenetic pattern. When these sites would be involved in the species barrier, this barrier is expected to be independent of species relationships. For example, the mouse residue Tyr192 (155 in mouse), which is Asn in hamster, seems important for the hamster-mouse species barrier (Priola, Chabry, and Chan 2001). At this position an His is not only present in human and cow but also in hippo and tenrec. Species barriers thus could be present between closely related species and could be broken again in distant ones. Also, the fact that the overall topology of the prion gene tree (Fig. 4) agrees with the species tree suggests that no dramatic sequence changes have occurred to avoid cross-species TSE infectivity. In conclusion, the TSE species barrier remains elusive for the time being.

Acknowledgments

This work was supported by grants from the Netherlands Organization for Scientific Research (NWO-ALW) and the European Commission.

References

Beck, J. A., S. Mead, T. A. Campbell, A. Dickinson, D. P. Wientjens, E. A. Croes, C. M. Van Duijn, and J. Collinge. 2001. Two-octapeptide repeat deletion of prion protein associated with rapidly progressive dementia. *Neurology* **57**:354–356.

Bounhar, Y., Y. Zhang, C. G. Goodyer, and A. LeBlanc. 2001. Prion protein protects human neurons against Bax-mediated apoptosis. *J. Biol. Chem.* **276**:39145–39149.

Brown, D. R. 2001. Prion and prejudice: normal protein and the synapse. *Trends. Neurosci.* **24**:85–90.

Brown, D. R., R. S. Nicholas, and L. Canevari. 2002. Lack of prion protein expression results in a neuronal phenotype sensitive to stress. *J. Neurosci. Res.* **67**:211–224.

Brown, D. R., K. Qin, J. W. Herms et al. (13 co-authors). 1997. The cellular prion protein binds copper in vivo. *Nature* **390**:684–687.

Brown, D. R., B. S. Wong, F. Hafiz, C. Clive, S. J. Haswell, and I. M. Jones. 1999. Normal prion protein has an activity like that of superoxide dismutase. *Biochem. J.* **344**:1–5.

Cohen, F. E., and S. B. Prusiner. 1998. Pathologic conformations of prion proteins. *Annu. Rev. Biochem.* **67**:793–819.

Collinge, J. 2001. Prion diseases of humans and animals: their causes and molecular basis. *Annu. Rev. Neurosci.* **24**:519–550.

Derreumaux, P. 2001. Evidence that the 127–164 region of prion proteins has two equi-energetic conformations with beta or alpha features. *Biophys. J.* **81**:1657–1665.

Flechsig, E., D. Shmerling, I. Hegyi, A. J. Raeber, M. Fischer, A. Cozzio, C. von Mering, A. Aguzzi, and C. Weissmann. 2000. Prion protein devoid of the octapeptide repeat region restores susceptibility to scrapie in PrP knockout mice. *Neuron* **27**:399–408.

Gatesy, J., and M. A. O'Leary. 2001. Deciphering whale origins with molecules and fossils. *Trends Ecol. Evol.* **16**:562–570.

Gilch, S., C. Spielhaupter, and H. M. Schätzl. 2000. Shortest known prion protein allele in highly BSE-susceptible lemurs. *Biol. Chem.* **381**:521–523.

Gill, A. C., M. A. Ritchie, L. G. Hunt, S. E. Steane, K. G. Davies, S. P. Bocking, A. G. Rhie, A. D. Bennett, and J. Hope. 2000. Post-translational hydroxylation at the N-terminus of the prion protein reveals presence of PPII structure in vivo. *EMBO J.* **19**:5324–5331.

Goldmann, W., A. Chong, J. Foster, J. Hope, and N. Hunter. 1998. The shortest known prion protein gene allele occurs in goats, has only three octapeptide repeats and is nonpathogenic. *J. Gen. Virol.* **79**:3173–3176.

Goldmann, W., N. Hunter, R. Somerville, and J. Hope. 1996. Prion phylogeny revisited. *Nature* **382**:32–33.

Harris, D. A. 1999. Cell biological studies of the prion protein. *Curr. Issues. Mol. Biol.* **1**:65–75.

Hegde, R. S., J. A. Mastrianni, M. R. Scott, K. A. DeFea, P. Tremblay, M. Torchia, S. J. DeArmond, S. B. Prusiner, and V. R. Lingappa. 1998. A transmembrane form of the prion protein in neurodegenerative disease. *Science* **279**:827–834.

Hill, A. F., S. Joiner, J. Linehan, M. Desbruslais, P. L. Lantos, and J. Collinge. 2000. Species-barrier-independent prion Molecular Evolution of the Mammalian Prion Protein replication in apparently resistant species. *Proc. Natl. Acad. Sci. USA* **97**:10248–10253.

Hope, J. 2000. Prions and neurodegenerative diseases. *Curr. Opin. Genet. Dev.* **10**:568–574.

Horiuchi, M., S. A. Priola, J. Chabry, and B. Caughey. 2000. Interactions between heterologous forms of prion protein: binding, inhibition of conversion, and species barriers. *Proc. Natl. Acad. Sci. USA* **97**:5836–5841.

Hosszu, L. L., N. J. Baxter, G. S. Jackson, A. Power, A. R. Clarke, J. P. Waltho, C. J. Craven, and J. Collinge. 1999. Structural mobility of the human prion protein probed by backbone hydrogen exchange. *Nat. Struct. Biol.* **6**:740–743.

Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754–755.

Jackson, G. S., I. Murray, L. L. Hosszu, N. Gibbs, J. P. Waltho, A. R. Clarke, and J. Collinge. 2001. Location and properties of metal-binding sites on the human prion protein. *Proc. Natl. Acad. Sci. USA* **98**:8531–8535.

James, T. L., H. Liu, N. B. Ulyanov et al. (11 co-authors). 1997. Solution structure of a 142-residue recombinant prion protein corresponding to the infectious fragment of the scrapie isoform. *Proc. Natl. Acad. Sci. USA* **94**:10086–10091.

Kaneko, K., L. Zulianello, M. Scott, C. M. Cooper, A. C. Wallace, T. L. James, F. E. Cohen, and S. B. Prusiner. 1997. Evidence for protein X binding to a discontinuous epitope on the cellular prion protein during scrapie prion propagation. *Proc. Natl. Acad. Sci. USA* **94**:10069–10074.

Krakauer, D. C., P. M. Zanotto, and M. Pagel. 1998. Prion's progress: patterns and rates of molecular evolution in relation to spongiform disease. *J. Mol. Evol.* **47**:133–145.

Kramer, M. L., H. D. Kratzin, B. Schmidt, A. Romer, O. Windl, S. Liemann, S. Hornemann, and H. Kretzschmar. 2001. Prion protein binds copper within the physiological concentration range. *J. Biol. Chem.* **276**:16711–16719.

Kuwahara, C., A. M. Takeuchi, T. Nishimura, K. Haraguchi, A. Kubosaki, Y. Matsumoto, K. Saeki, T. Yokoyama, S. Itohara, and T. Onodera. 1999. Prions prevent neuronal cell-line death. *Nature* **400**:225–226.

Liemann, S., and R. Glockshuber. 1999. Influence of amino acid substitutions related to inherited human prion diseases on the thermodynamic stability of the cellular prion protein. *Biochemistry* **38**:3258–3267.

Liu, H., S. Farr-Jones, N. B. Ulyanov, M. Llinas, S. Marqusee, D. Groth, F. E. Cohen, S. B. Prusiner, and T. L. James. 1999. Solution structure of Syrian hamster prion protein rPrP(90-231). *Biochemistry* **38**:5362–5377.

Lopez Garcia, F., R. Zahn, R. Riek, and K. Wuthrich. 2000. NMR structure of the bovine prion protein. *Proc. Natl. Acad. Sci. USA* **97**:8334–8339.

Maiti, N. R. and W. K. Surewicz. 2001. The role of disulfide bridge in the folding and stability of the recombinant human prion protein. *J. Biol. Chem.* **276**:2427–2431.

McMahon, H. E., A. Mange, N. Nishida, C. Creminon, D. Casanova, and S. Lehmann. 2001. Cleavage of the amino terminus of the prion protein by reactive oxygen species. *J. Biol. Chem.* **276**:2286–2291.

Miura, T., A. Hori-i, H. Mototani, and H. Takeuchi. 1999. Raman spectroscopic study on the copper(II) binding mode of prion octapeptide and its pH dependence. *Biochemistry* **38**:11560–11569.

Mouillet-Richard, S., M. Ermonval, C. Chebassier, J. L. Laplanche, S. Lehmann, J. M. Launay, and O. Kellermann. 2000. Signal transduction through prion protein. *Science* **289**:1925–1928.

Muniz, M. and H. Riezman. 2000. Intracellular transport of GPIanchored proteins. *EMBO J.* **19**:10–15.

Murphy, W. J., E. Eizirik, S. J. O'Brien et al. (11 co-authors). 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**: 2348–2351.

Parchi, P., W. Zou, W. Wang et al. (13 co-authors). 2000. Genetic influence on the structural variations of the abnormal prion protein. *Proc. Natl. Acad. Sci. USA* **97**:10168–10172.

Pauly, P. C., and D. A. Harris. 1998. Copper stimulates endocytosis of the prion protein. *J. Biol. Chem.* **273**:33107–33110.

Peretz, D., M. R. Scott, D. Groth, R. A. Williamson, D. R. Burton, F. E. Cohen, and S. B. Prusiner. 2001. Strainspecified relative conformational stability of the scrapie prion protein. *Protein. Sci.* **10**:854–863.

Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**:817–818.

Priola, S. A., J. Chabry, and K. Chan. 2001. Efficient conversion of normal prion protein (PrP) by abnormal hamster PrP is determined by homology at amino acid residue 155. *J. Virol.* **75**:4673–4680.

Priola, S. A., and V. A. Lawson. 2001. Glycosylation influences cross-species formation of protease-resistant prion protein. *EMBO J.* **20**:6692–6699.

Prusiner, S. B. 1998. Prions. *Proc. Natl. Acad. Sci. USA* **95**: 13363–13383.

Prusiner, S. B., and M. R. Scott. 1997. Genetics of prions. *Annu. Rev. Genet.* **31**:139–175.

Puckett, C., P. Concannon, C. Casey, and L. Hood. 1991. Genomic structure of the human prion protein gene. *Am. J. Hum. Genet.* **49**:320–329.

Raeber, A. J., S. Brandner, M. A. Klein, Y. Benninger, C. Musahl, R. Frigg, C. Roecki, M. B. Fischer, C. Weissmann, and A. Aguzzi. 1998. Transgenic and knockout mice in research on prion diseases. *Brain. Pathol.* **8**:715–733.

Riek, R., S. Hornemann, G. Wider, M. Billeter, R. Glockshuber, and K. Wuthrich. 1996. NMR structure of the mouse prion protein domain PrP(121–321). *Nature* **382**:180–182.

Rudd, P. M., M. R. Wormald, D. R. Wing, S. B. Prusiner, and R. A. Dwek. 2001. Prion glycoprotein: structure, dynamics, and roles for the sugars. *Biochemistry* **40**:3759–3766.

Schätzl, H. M., M. Da Costa, L. Taylor, F. E. Cohen, and S. B. Prusiner. 1995. Prion protein gene variation among primates. *J. Mol. Biol.* **245**:362–374.

Scott, M. R., J. Safar, G. Telling, O. Nguyen, D. Groth, M. Torchia, R. Koehler, P. Tremblay, D. Walther, F. E. Cohen, S. J. DeArmond, and S. B. Prusiner. 1997. Identification of a prion protein epitope modulating transmission of bovine spongiform encephalopathy prions to transgenic mice. *Proc. Natl. Acad. Sci. USA* **94**:14279–14284.

Sørensen, A. B., M. Duch, P. Jorgensen, and F. S. Pedersen. 1993. Amplification and sequence analysis of DNA flanking integrated proviruses by a simple two-step polymerase chain reaction method. *J. Virol.* **67**:7118–7124.

Stewart, R. S., and D. A. Harris. 2001. Most pathogenic mutations do not alter the membrane topology of the prion protein. *J. Biol. Chem.* **276**:2212–2220.

Sumudhu, W., Perera, W. S. and N. M. Hooper. 2001. Ablation of the metal ion-induced endocytosis of the prion protein by disease-associated mutation of the octarepeat region. *Curr. Biol.* **11**:519–523.

Swofford, D. L. 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Udenfriend, S., and K. Kodukula. 1995. How glycosylphosphatidylinositol- anchored membrane proteins are made. *Annu. Rev. Biochem.* **64**:563–591.

Viles, J. H., F. E. Cohen, S. B. Prusiner, D. B. Goodin, P. E. Wright, and H. J. Dyson. 1999. Copper binding to the prion protein: structural implications of four identical cooperative binding sites. *Proc. Natl. Acad. Sci. USA* **96**:2042– 2047.

Wopfner, F., G. Weidenhofer, R. Schneider, A. von Brunn, S. Gilch, T. F. Schwarz, T. Werner, and H. M. Schätzl. 1999. Analysis of 27 mammalian and 9 avian PrPs reveals high conservation of flexible regions of the prion protein. *J. Mol. Biol.* **289**:1163–1178.

Chapter 8

General Discussion and Summary

Genes and proteins from different organisms tell us about the evolution of these macromolecules as well as the history of these organisms. Both aspects are dealt with in this thesis. Chapter 1 emphasizes the interdependence between molecular evolution and molecular phylogenetics. A thorough understanding of mutational and evolutionary processes is required to design the mathematical models used in phylogeny reconstruction. In turn, phylogenetic information is needed to interpret the evolution of genes and proteins.

Mammalian Molecular Phylogeny

Chapter 2 gives a prime example of how molecular evolution can elude phylogenetic inference. Taxonomists divide mammals into three subclasses. The smallest subclass is formed by the egg-laying monotremes, of which the Australian platypus and echidna are the only surviving representatives. Marsupials, the second subclass, occur in Australia and South America and are characterized by their pouch, in which the young develop after a very short internal gestation. The third and predominant subclass is formed by the eutherian mammals, in which the placenta allows a prolonged gestation period, resulting in the birth of well developed young. Morphology provides strong evidence that Monotremata are the oldest offshoot of the mammals, which subsequently divided into Marsupialia and Eutheria (Fig. 1, page 20). The latter two are jointly grouped as Theria. It therefore was a great surprise that comparison of the complete mitochondrial genome of the platypus with that of marsupials and placental mammals clearly told a different story: it supported the grouping of monotremes with marsupials, the Marsupionta hypothesis (Janke et al. 1996; Penny and Hasegawa 1997; Janke et al. 2002). However, results from later sequence studies based on limited sampling of nuclear genes were inconsistent as to the phylogenetic position of monotremes. We therefore created and analysed two large datasets of nuclear sequences, both newly determined and from the databases, of which the results are reported in Chapter 2.

Dataset I consisted of sequences from 5 nuclear genes, with a total length of 2793 bp, for 2 monotreme species, 6 marsupials, 9 eutherians and 4 outgroup species (a bird and 3 reptiles). Dataset II comprised a much longer combined sequence - 11544 bp or 4050 amino acids - derived from 9 genes and 4 proteins, but only for 5 taxa (a monotreme, a marsupial, a rodent, human and chicken). Phylogenetic analyses of both datasets were performed by parsimony, minimum evolution, maximum likelihood and Bayesian methods. Depending on method and model of evolution, dataset I gave the highest bootstrap support for the classical grouping of marsupials with eutheria (Theria), while the support for a monotreme-marsupial grouping (Marsupionta) was negligible (Fig. 2 and Table 5, pp. 26 and 27). Interestingly, the third alternative – to group monotremes with eutherians, a relationship which has never been proposed – received intermediate support in some analyses. Differences in base composition between investigated groups can affect phylogenetic outcome. In dataset I, the monotreme genes were very GC-rich while those for marsupials were GC-poor. Correcting for this base compositional bias considerably increased the support for Theria. The longer dataset II exclusively supported Theria with the highest possible values in all analyses. Moreover, on basis of dataset II the alternative possibilities for mammalian relationships (Marsupionta or a Monotremata-Eutheria clade) could both be rejected with statistical significance.

Insertions and deletions in protein-coding sequences have recently been used as markers to address several phylogenetic problems concerning for example the relationships in teleost fish (Venkatesh et al. 2001). The mutational events giving rise to insertions and deletions (indels) are qualitatively different from base substitutions. As such, they provide a source of phylogenetic information which is independent of sequence data (Rokas and Holland 2000). We therefore searched the aligned sequences of both datasets for insertions or deletions that could be diagnostic for resolving the trichotomy of the mammalian subclasses. Only for Theria was supporting evidence found, consisting of two single amino acid deletions and one insertion in two different proteins (Fig. 3, page 29). No indels supporting the two possible alternative relationships were observed. The morphological evidence for Theria is thus fully supported by our molecular data from nuclear genes. It has been shown that the discrepancy between the positioning of the monotremes on basis of nuclear and mitochondrial evidence is caused by the base compositional bias in the mitochondrial genomes of monotremes and marsupials (Phillips and Penny 2003). Our results confirm that mitochondrial data in deeper vertebrate phylogeny should be considered with caution (Curole and Kocher 1999).

Comparative sequence data can also be used to estimate the times of divergence of homologous genes, and thus of the organisms from which the genes originate (Hedges 2002). Since monotremes have a very poor fossil record, little is known about the time by which they diverged from the other mammals. In Chapter 2 the sequences of dataset II were therefore also used to estimate the time of divergence between Monotremata and Theria, taking the fossil-based divergence time of the lineages leading to birds and mammals (310 million years ago; Mya) as a calibration point. Our molecular clock analyses indicate that marsupials diverged from eutherians in the Middle or Late Jurassic (173-151 Mya) and that monotremes diverged from therians in the Late Triassic or Early Jurassic (223-198 Mya). Paleontological estimates for the marsupial-placental divergence have ranged from ca. 125 to 97 Mya, but the most recent fossil evidence would bring this split back to at least 167 Mya (Woodburne et al. 2003), which is compatible with our molecular estimate of 173-151 Mya. Other recent molecular datings push the marsupial-eutherian split even further back in time, proposing a divergence bewteen 190-182 Mya (Woodburne et al. 2003). Whatever the final answer, it is clear that the actual divergence of marsupials and eutherians, and hence that of monotremes and therians, occurred much earlier than thought until now on basis of paleontological evidence.

Chapter 3 presents another example of the phylogenetic utility of indels in protein-coding DNA sequences. In this case it concerns the recent debate on whether rodents are basal in the placental mammalian tree, as supported by mitochondrial sequence data (e.g., Reyes, Pesole and Saccone 2000; Arnason et al. 2002) and some analyses of nuclear genes (Misawa and Janke 2003) or rather nested higher in the tree, as indicated by more extensive nuclear sequence analyses (Madsen et al. 2001; Murphy et al. 2001). These latter nuclear analyses place rodents in a clade, named Euarchontoglires, which also includes primates. Such a close relationship of rodents and primates is also difficult to reconcile with morphological insights (Liu et al. 2001; Novacek 2001). While studying genes involved in various neurodegenerative disorders, we noticed two deletions that might be informative with respect to this problem. One is a large deletion in exon 8 of the gene for spinocerebellar ataxia 1 (*SCA1*), resulting in an 18-residue deletion in the encoded protein. The other is a six base-pair deletion at the 5' end of the intronless coding region of the prion protein gene (*PRNP*). The presence or absence of these deletions was determined in representatives of all 18 orders of placental mammals and in marsupials as an outgroup. It was found that both deletions only occur in the 5 orders which together form the Euarchontoglires, and are absent in all other 13 placental orders as well as in the outgroup marsupials (Fig. 1, page 38).

The most logical interpretation of this finding obviously is that the two deletions originated independently in the *SCA1* and *PRNP* genes of the last common ancestor of Euarchontoglires, and thus supports their monophyly. If the morphological or mitogenomic trees were true, both

deletions should have originated at least twice and in parallel in exactly the same lineages. Although reversal of a deletion is in principle impossible, a repeated origin cannot totally be excluded. Instances of such homoplasy, especially in regions with sequence repeats, have been documented (e.g., de Jong et al. 2003). However, it is the congruence of two independent deletions in the same orders that are also grouped by independent sequence evidence (Madsen et al. 2001; Murphy et al. 2001) that makes them compelling indicators for the monophyly of Euarchontoglires.

Molecular Evolution of Vertebrate Eye Lens Proteins

The vertebrate eye lens has a very high concentration of proteins, of which the majority derives from the families of α- and β/γ-crystallins, the so called ubiquitous lens proteins (Wistow 1993, 1995). The β/γ-crystallins form a largely lens specific family of structural proteins (Lubsen et al. 1988), while the two α-crystallins belong to the family of small heat shock proteins (sHsps) (de Jong et al. 1998). Because of their chaperone-like properties, characteristic for sHsps, the α-crystallins might protect against the aggregation of other lens proteins and thus maintain transparency of the lens (Horwitz 2003). The proportions and subunit compositions of these ubiquitous crystallins vary considerably between species, and additional crystallins – so called taxon-specific - are found in specific vertebrate lineages. It is assumed that the ubiquitous crystallins were recruited to function as lens proteins when the vertebrate eye lens came into existence, i.e. in the last common ancestor of the vertebrates. The taxon-specific crystallins were recruited in specific lineages later in evolution, mostly by overexpression in the lens of enzymes that have a house keeping function in other cell types (Piatigorsky and Wistow 1991; Piatigorsky 1998). Comparing the crystallins that are expressed in various vertebrate groups allows a reconstruction of the evolution and origins of these proteins. It may reveal differences that reflect adaptations to changing structural or functional requirements of the lens, since vertebrate lenses vary in texture, plasticity and protein concentration. Other changes in crystallin composition may reflect evolutionarily neutral events that are allowed as long as they are compatible with the functioning of the lens. As such, the study of crystallin evolution is informative for understanding protein evolution in general.

Analysing the crystallins of the duck-billed platypus, *Ornithorhynchus anatinus*, was thought to be useful because such data would fill the gap between the well studied crystallins of man, mouse and other placental mammals on one hand and those from chicken and other birds on the other. Chapters 4 and 5 report the results of these studies of platypus lens proteins.

In Chapter 4 the focus is on the α-crystallins of the platypus. The sequences of αA- and αB-crystallin were deduced from cDNA obtained from the platypus eye lens. Both proteins were generally well conserved, as is usual for α-crystallins, but the platypus αB-crystallin showed a unique tandem duplication of a heptapeptide FPTSFPA in its N-terminal region (Figs. 1 and 3, pp. 47 and 48). Interestingly, this repeat was found to be further extended by an FPTFPA insert in αB-crystallin of the echidna *Tachyglossus aculeatus*, of which the sequence was deduced from genomic DNA. These duplications apparently arose from replication slippages in a CT-rich repetitive sequence in the gene, resulting in a remarkable six times repeated FP(A/T) motif in echidna αB-crystallin. Regions around this repeat have been implicated in subunit interactions and chaperone activity of α-crystallins and other sHsps (e.g., Pasta et al. 2003). It thus would be worthwhile to study the functional effects of the repeats in echidna αB-crystallin, the more so as αB-crystallin increasingly appears to fulfil important functions outside the lens, notably in apoptosis, stress tolerance and cytoskeletal structuring (Arrigo and Müller 2002; van Montfort et al. 2002).

An alternative splice variant of αA-crystallin, called αA^{ins}-crystallin, occurs at levels of up to 20% of total αA-crystallin in the lenses of several placental mammals. It has an insert of 23 amino acid residues, encoded by an optional exon in the middle of the first intron of the αA-crystallin

gene (King and Piatigorsky 1983). Alternative splicing is a common mechanism to increase the diversity of gene products, and thus is often a useful evolutionary achievement. However, in the case of αA^{ins}-crystallin, no obvious advantage for the origin and maintenance of this splice variant could be determined (Hendriks et al. 1990; Smulders et al. 1995; van Dijk et al. 2001). Chapter 4 demonstrates the presence in the platypus lens of the αA^{ins}-crystallin mRNA. The 23-residue insert peptide is well conserved as compared with that in other placentals and in the kangaroo (Fig. 2, page 47). Any indication of the presence of the alternatively spliced exon αA^{ins} could however not be found in the αA-crystallin genes of a crocodile and a lizard (newly determined for this study), nor in those genes from duck or chicken (in the database). This suggests that the optional exon for αA^{ins}-crystallin originated in the last common ancestor of mammals, but the way it originated and why it is conserved remains completely enigmatic.

Since α-crystallin sequences have been useful in the past to unravel various problems in deeper placental mammal and avian phylogeny, there was the hope that platypus α-crystallins might also be informative about the phylogenetic position of monotremes. However, from the phylogenetic analyses in Chapter 4 it appears that the phylogenetic signal in platypus and echidna α-crystallin sequences is insufficient to distinguish between the two competing proposals for monotreme relationships, the Theria and Marsupionta hypotheses. The analyses in Chapter 4 also include the newly determined αA-crystallin sequence of a curious reptile, the tuatara *Sphenodon punctatus*. It appears that, amongst the other reptiles included in this analysis (gecko, crocodile, turtle) tuatara groups most closely with the gecko, which belongs to the lizards. This would be in agreement with recent evidence from complete mitochondrial genome analyses which place the tuatara as the sister group of lizards and snakes (Squamata) (Rest et al. 2003).

Chapter 5 extends the earlier evidence that the evolution of the eye provides a unique example of the acquirement of new protein structures and functions, either by "gene sharing" (i.e., a gene acquiring a dual function) or after gene duplication (reviewed in Wistow 1995; Piatigorsky 1998). About ten types of taxon-specific crystallins are presently known to occur in various vertebrate lineages. These taxon-specific crystallins are related to, or identical with, common metabolic enzymes like lactate dehydrogenase B, α-enolase, alcohol dehydrogenase and aldose/aldehyde reductases. A single exception is iota-crystallin, which is a cellular retinol-binding protein (CRBP I), occurring in the lenses of some diurnal gecko species (Werten et al. 2000). In Chapter 5 we now describe the presence of an unusual ~41-kD protein that makes up 16-18% of the total protein in the platypus eye lens. Its cDNA sequence was determined, which identified the protein as muscle-type lactate dehydrogenase (LDH-A). This is the first observation of LDH-A as a crystallin, and we designate it as upsilon-(υ)-crystallin. Interestingly, the related heart-type LDH-B was already known to occur as an abundant lens protein, named ε-crystallin, in many birds and crocodiles (Wistow et al. 1987). Thus, two members of the *ldh* gene family have independently been recruited as crystallins in different higher vertebrate lineages, suggesting that they are particularly suited for this purpose in terms of gene regulatory or protein structural properties. To assess whether platypus LDH-A/υ-crystallin has been under different selective constraints as compared to other vertebrate LDH-A sequences, we also reconstructed the vertebrate *ldh-a* gene phylogeny (Fig. 3, page 57). However, no conspicuous rate deviations or amino acid replacements, that might reflect a change in selective constraint, were observed.

Chapter 6 is inspired by the observation that the αB-crystallin gene is located only about 0.9 kb apart and in a head-to-head manner from the gene for HspB2 - a related sHsp - in the human, mouse and rat genomes (Iwaki et al. 1997). While αB-crystallin is abundantly expressed in lens and muscle and upregulated in response to heat shock, HspB2 is abundant only in muscle and not upregulated by a heat shock. It has recently been established that bidirectional gene pairs, located so close that promoter regions overlap, are surprisingly common in mammals (Adachi and

Lieber 2002; Takai and Jones 2003). Tracing such gene pairs that are evolutionary conserved is of considerable interest as it will help delineate the regulatory modules used in the eukaryotic genomes. In addition, comparison of the intergenic region between a conserved gene pair will provide insight into the evolution of eukaryotic promoter regions. With the elucidation of the sequence of eukaryotic genomes, it is becoming evident that changes in gene regulation rather than gene number are the driving force for phenotypic evolution (for recent review, see Wray et al. 2003).

In Chapter 6 we therefore examined the linkage of the αB-crystallin and HspB2 genes in the major mammalian lineages and in chicken and duck. The intergenic distance in mammals was found to range from 645 bp (platypus) to 1069 bp (opossum), with an average of about 900 bp. In chicken and duck the intergenic distance is considerably larger, around 1.6 kb. Phylogenetic footprinting (Fig. 2, page 68) and direct inspection of the sequence alignment (Fig. 3, page 70) showed conservation of sequence elements close to the HspB2 promoter and identified two additional conserved regions further upstream. All known regulatory elements of the mouse αB-crystallin promoter are conserved in studied mammals, except in platypus and birds. The lens-specific-region 1 (LSR1) as well as the heat shock elements (HSE's) are missing in the avian intergenic region; platypus has only the Pax-6 site of LSR1 and lacks both the Pax-6 site in LSR2 and one of the two HSE's. Our results argue that the primordial mammalian αB-crystallin promoter had two LSR's and two HSE's and that the loss of one of the Pax-6 sites and one of the HSE's in platypus is secondary.

To determine the functional significance of the sequence divergence of the platypus intergenic region, the activities of the platypus, blind mole rat and rat αB-crystallin and HspB2 promoters in lens and muscle cells were compared in transfection experiments. It appeared that the platypus αB-crystallin promoter retained heat shock responsiveness and lens expression. It also directed lens expression in *Xenopus laevis* transgenes, as did the HspB2 promoter of rat or blind mole rat. This complex enhancer region in the middle of the intergenic region has been previously suggested to work only towards the αB-crystallin promoter, irrespective of distance or orientation (Swamynathan and Piatigorsky 2002). Yet deletion of this region affected the activity of both the rat αB-crystallin and HspB2 promoters, suggesting that at least some elements within the complex enhancer region work towards the HspB2 promoter. Elements acting on the HspB2 promoter may be interspersed with those acting on the αB-crystallin promoter, where occupancy of an element working towards one promoter could promote occupancy of a second element working towards another promoter. Intermingling of elements and synergistic binding of transcription factors to the different elements could be an explanation for the selective pressure to maintain the complex enhancer region and the head-to-head orientation.

The lens proteins αA- and αB-crystallin form together with eight related proteins the sHsp family in man and other mammals (Kappé et al. 2003). Considering the considerable sequence differences between these 10 sHsps, one may infer that the gene duplications responsible for their origin already occurred before the divergence of the earliest vertebrates. In that case one would expect that orthologs of each of them could be present in all vertebrate classes, and perhaps even in their closest relatives, the cephalochordates. It therefore was interesting to establish whether a newly sequenced sHsp from amphioxus (*Branchyostoma lanceolatum*) (T. van Rheede, unpublished data), the first known cephalochordate sHsp, could be identified as an ortholog of one of the mammalian sHsps. An exhaustive search of the databases, including the genomes of *Fugu*, zebrafish and the urochordate *Ciona intestinalis*, retrieved all entries that could be recognized as chordate sHsps (Franck et al., submitted). Phylogenetic analyses of these sHsps revealed the presence of at least 15 paralogous sHsps in vertebrates. As expected, orthologs of most of the 10 mammalian sHsps – including αA- and αB-crystallin - were present in all investigated vertebrate classes. However, the

amphioxus and *Ciona* sHsps did not appear to be the orthologs of any specific vertebrate sHsp. This can be explained by the rounds of gene and genome duplications that have occurred in early vertebrate evolution, after their separation from the lower chordates.

Molecular Evolution of the Prion Protein

The prion protein (PrP) is included in this thesis for two reasons. First of all it is an evolutionary attractive subject because of its unique features as a pathogenic protein that is associated with transmissible spongiform encephalopathies (TSEs) (reviewed in Prusiner 1998; Collinge 2001). Second, as shown already in Chapter 3, it contains relevant information about mammalian phylogenetic relationships. Remarkably little is known about the normal cellular function of PrP and the manner in which it exerts its pathogenicity. The human prion gene *PRNP* encodes a 253-residue precursor protein in an intronless open reading frame. It is expressed in most tissues, but highest levels are found in the central nervous system. After processing, a glycosyl-phosphatidylinositol (GPI) anchor attaches the protein to the outer membrane surface of the cell. NMR measurements have established the conformation of recombinant PrP of human and other mammals. The N-terminal region is unstructured and flexible, and comprises a segment of five or six repeats which is implicated in copper-binding. The C-terminal region forms a more rigid globular domain, stabilized by a disulfide bridge and comprising two variably occupied N-linked glycosylation sites.

The function of the normal cellular isoform of the prion protein (PrP^c) remains enigmatic (Mouillet-Richard et al. 2000; Bounhar et al. 2001; Brown 2001). PrP^c-deficient mice develop normally. PrP^c is a copper-binding protein that appears to protect against programmed cell death and Bax-mediated apoptosis. There is some evidence that PrP^c is a cell surface receptor for signal transduction, coupled to the tyrosine kinase Fyn.

A conformational change in PrP^c gives rise to the pathogenic form PrP^{sc}. This transition involves a dramatic increase in β-sheet content, and a decrease in α-helical structure (Prusiner 1998). PrP^{sc} catalyzes further misfolding of PrP^c, thus leading to a self-amplifying cycle and the formation of insoluble, extracellular aggregates. Inherited, sporadic and infectious forms of prion diseases exist. Inherited forms are associated with mutations in the prion gene that enhance the transition from PrP^c to PrP^{sc}. In sporadic cases PrP^{sc} may derive from spontaneous misfolding of PrP^c. Infectivity occurs when the pathological transformation of the host's PrP^c is induced by PrP^{sc} particles transmitted from individuals of the same or different species.

Interspecies infectivity of TSEs varies greatly (Prusiner and Scott 1997). Sequence differences between PrP of donor and recipient species play a role, but interspecies susceptibility is not simply determined by overall sequence similarity. More important seem to be the prion strains within a species, which are isoenergetic conformers of PrP^{sc}, characterized by variations in clinical presentation and protease resistance. Strain variation is encoded by different PrP^{sc} conformations and ratios of the three PrP glycoforms (di-, mono- and unglycosylated), and further influenced by PrP sequence polymorphisms and metal binding (Collinge 2001). However, the precise molecular basis of strain variation remains unknown.

Many residues and regions in the prion protein have been implicated in functioning, pathogenicity and species barrier. Sequence comparison of mammalian prion proteins may help to evaluate such proposals and gain insight in the molecular evolution of PrP. However, PrP sequences were until now available for only six of the 18 orders of placental mammals, being mainly primates, artiodactyls (like cow, sheep and goat) and a subset of rodents (Schätzl et al. 1995; Wopfner et al. 1999). A broader comparison of mammalian prions might help to understand the enigmatic functional and pathogenic properties of this protein.

As described in Chapter 7, we therefore determined PrP coding sequences from 26 mammalian species to include all placental orders and major subordinal groups. It was found that in mammalian PrP the glycosylation sites, the cysteines forming a disulfide bridge, and a hydrophobic transmembrane region, are perfectly conserved. Also the sequences responsible for secondary structure elements, for N- and C-terminal processing of the precursor protein, and for attachment of the GPI membrane anchor, are well conserved. The N-terminal region of PrP generally contains five or six repeats of the sequence P(Q/H)GGG(G/-)WGQ, but alleles with two, four and seven repeats were observed in some species. This suggests, together with the pattern of amino acid replacements in these repeats, the regular occurrence of repeat expansion and contraction. Histidines implicated in copper ion binding, and a proline involved in 4-hydroxylation are lacking in some species, which questions their importance for normal functioning of cellular PrP. The finding in certain species of two or seven repeats, and of amino acid substitutions that have been related to human prion diseases, challenges the relevance of such mutations for prion pathology.

The PrP sequences presented in Chapter 7 were also used to construct a *PRNP* gene tree (Fig. 4, page 95). The obtained phylogeny largely agrees with the species tree, which indicates that no major deviations occurred during the evolution of the prion gene in different placental lineages. In one species, the anteater, a prion pseudogene was found to be present in addition to the active gene.

A striking finding in Chapter 7 is the apparent heterozygosity in a squirrel, *Sciurus vulgaris*, for *PRNP* alleles with 5 and 2 repeats. Heterozygosity for 2 and 5 repeats has also been reported in lemurs, being primates which are particularly susceptible for TSE (Gilch, Spielhaupter and Schätzl 2000). The repeat region is needed for copper ion binding, and modulates prion replication and pathogenicity (Flechsig et al. 2000; Collinge 2001). In humans, reduction of the repeat number from 5 to 4 does not lead to prion disease. In fact, alleles with 4 repeats are found at frequencies of up to 2% in human populations (Puckett et al. 1991). However, heterozygosity for 3 repeats has been associated with a rapidly progressive dementia consistent with Creutzfeldt-Jakob disease (CJD) (Beck et al. 2001). Alleles with 2 repeats have not been found in human, and are expected to be deleterious. However, if alleles with 2 repeats occur in squirrel and lemurs, homozygotes may be expected too, unless this condition is lethal.

As a follow up of Chapter 7, the *PRNP* gene sequences in a larger number of squirrels has meanwhile been determined (squirrel ears kindly supplied by Stichting Eekhoornopvang Nederland, De Meern; T. van Rheede, T. Kortum, O. Madsen, W.W. de Jong, in preparation). It was found that all squirrels in which the *PRNP* gene could be amplified have a sequence with 2 and one with 5 repeats. Since not all of these squirrels can be heterozygotes, it must be concluded that squirrels have a duplicated set of *PRNP* genes. The duplication must have occurred relatively recent in the squirrel ancestry, since only a few base substitutions distinguish the short and the long *PRNP* genes. It further appears that the *PRNP* gene with two repeats is not expressed as a protein: western blotting of squirrel brain extract with prion-specific antibodies (kindly provided by Dr. J. Langeveld, CIDC, Lelystad) only showed a single PrP band of the same size as found in normal human and rat brain. It must be concluded that after the gene duplication, one of the copies has lost 3 repeats by unequal crossing over and replication slippage (Collinge 2001). This gene has apparently become silenced, which unfortunately makes it impossible to find out whether the expression of a PrP protein with two repeats is indeed deleterious.

References

Adachi, N., and M.R. Lieber. 2002. Bidirectional gene organization: a common architectural feature of the human genome. *Cell* **109**:807-809.

Arnason U., J.A. Adegoke, K. Bodin, E.W. Born, Y.B. Esa, A. Gullberg, M. Nilsson, R.V. Short, X. Xu and A. Janke. 2002. Mammalian mitogenomic relationships and the root of the eutherian tree. *Proc. Natl. Acad. Sci. USA* **99**:8151-8156.

Arrigo, A.P., and W.E.G. Müller (eds). 2002. Small Stress Proteins, Springer Verlag, Berlin

Beck, J.A., S. Mead, T.A. Campbell, A. Dickinson, D.P. Wientjens, E.A. Croes, C.M. Van Duijn, and J. Collinge. 2001. Two-octapeptide repeat deletion of prion protein associated with rapidly progressive dementia. *Neurology* **57**:354-356.

Bounhar, Y., Y. Zhang, C.G. Goodyer, and A. LeBlanc. 2001. Prion protein protects human neurons against Bax-mediated apoptosis. *J. Biol. Chem.* **276**:39145-39149.

Brown, D.R. 2001. Prion and prejudice: normal protein and the synapse. *Trends. Neurosci.* **24**:85-90.

Collinge, J. 2001. Prion diseases of humans and animals: their causes and molecular basis. *Annu. Rev. Neurosci.* **24**:519-550.

Curole, J.P., and T.D. Kocher. 1999. Mitogenomics: digging deeper with complete mitochondrial genomes. *Trends Ecol. Evol.* 1**4**:394-398

de Jong, W.W., G.-J. Caspers, and J.A.M. Leunissen. 1998. Genealogy of the α-crystallin/small heat-shock protein superfamily. *Int. J. Biol. Macromol.* **22**:151-162.

de Jong, W.W., M.A.M. van Dijk, C. Poux, G. Kappé, T. van Rheede, and O. Madsen. 2003. Indels in protein-coding sequences of Euarchontoglires constrain the rooting of the eutherian tree. *Mol. Phylogen. Evol.* **28**: 328-340.

Flechsig, E., D. Shmerling, I. Hegyi, A.J. Raeber, M. Fischer, A. Cozzio, C. von Mering, A. Aguzzi, and C. Weissmann. 2000. Prion protein devoid of the octapeptide repeat region restores susceptibility to scrapie in PrP knockout mice. *Neuron* **27**:399-408.

Franck, E., O. Madsen, T. van Rheede, G. Ricard, M.A. Huynen, and W.W. de Jong. 2004. Evolutionary diversity of vertebrate small heat shock proteins.(submitted)

Gilch, S., C. Spielhaupter, and H.M. Schätzl. 2000. Shortest known prion protein allele in highly BSE-susceptible lemurs. *Biol. Chem.* **381**:521-523.

Hedges, S.B. 2002. The origin and evolution of model organisms. *Nat. Rev. Genet.* **3**:838-849.

Hendriks, W., H. Weetink, C.E.M. Voorter, J. Sanders, H. Bloemendal, and W.W. de Jong. 1990. The alternative splicing product αA^{ins}-crystallin is structurally equivalent to αA and αB subunits in the rat α-crystallin aggregate. *Biochim. Biophys. Acta* **1037**:58-65.

Horwitz, J. 2003. Alpha-crystallin. *Exp. Eye Res.* **76**:145-53.

Iwaki A., T. Nagano, M. Nakagawa, T. Iwaki, and Y. Fukumaki. 1997. Identification and characterization of the gene encoding a new member of the α-crystallin/small hsp family, closely linked to the αB-crystallin gene in a head-to-head manner. *Genomics* **45**:386-394

Janke, A., N.J. Gemmell, G. Feldmaier-Fuchs, A. von Haeseler, and S. Pääbo. 1996. The mitochondrial genome of a monotreme--the platypus (*Ornithorhynchus anatinus*). *J. Mol. Evol.* **42**:153-159.

Janke, A., O. Magnell, G. Wieczorek, M. Westerman, and U. Arnason. 2002. Phylogenetic analysis of 18S rRNA and the mitochondrial genomes of the wombat, *Vombatus ursinus*, and the spiny anteater, *Tachyglossus aculeatus*: increased support for the Marsupionta hypothesis. *J. Mol. Evol.* **54**:71-80.

Kappé, G., E. Franck, P. Verschuure, W.C. Boelens, J.A.M. Leunissen, and W.W. de Jong. 2003. The human genome encodes ten α-crystallin-related small heat shock proteins: HspB1-10. *Cell Stress Chaperon.* 8: 53-61.

King, C.R. and J. Piatigorsky. 1983. Alternative RNA splicing of the murine αA-crystallin gene: protein-coding information within an intron. *Cell* **32**:707-712.

Liu, F.G., M.M. Miyamoto, N.P. Freire, P.Q. Ong, M.R. Tennant, T.S. Young, and K.F. Gugel. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. *Science* **291**:1786-1789.

Lubsen, N.H., H.J. Aarts, and J.G. Schoenmakers. 1988. The evolution of lenticular proteins: the beta- and gamma-crystallin supergene family. *Prog. Biophys. Mol. Biol.* **51**:47-76.

Madsen, O., M. Scally, C.J. Douady, D.J. Kao, R.W. DeBry, R. Adkins, H.M. Amrine, M.J. Stanhope, W.W. de Jong, and M.S. Springer. 2001. Parallel adaptive radiations in two major clades of placental mammals. *Nature* **409**:610-614.

Misawa, K., and A. Janke. 2003. Revisiting the Glires concept--phylogenetic analysis of nuclear sequences. *Mol. Phylogenet. Evol.* **28**:320-327.

Mouillet-Richard, S., M. Ermonval, C. Chebassier, J.L. Laplanche, S. Lehmann, J.M. Launay, and O. Kellermann. 2000. Signal transduction through prion protein. *Science* **289**:1925-1928.

Murphy, W.J., E. Eizirik, W.E. Johnson, Y.P. Zhang, O.A. Ryder, and S.J. O'Brien. 2001. Molecular phylogenetics and the origins of placental mammals. *Nature* **409**:614-618.

Novacek, M.J. 2001. Mammalian phylogeny: genes and supertrees. *Curr. Biol.* **11**:R573-R575.

Pasta, S.Y., B. Raman, T. Ramakrishna, and C.M. Rao. 2003. Role of the Conserved SRLFDQFFG Region of α-Crystallin, a Small Heat Shock Protein. *J. Biol. Chem.* **278**:51159-51166.

Penny D, and M. Hasegawa. 1997. Molecular systematics. The platypus put in its place. *Nature* **387**:549-550.

Phillips, M.J., and D. Penny. 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. *Mol. Phylogenet. Evol.* **28**:171-85.

Piatigorsky, J. 1998. Gene sharing in lens and cornea: facts and implications. *Prog. Retin. Eye Res.* **17**:145-174.

Piatigorsky, J. and G. Wistow. 1991. The recruitment of crystallins: new functions precede gene duplication. *Science* **252**:1078-1079.

Prusiner, S.B. 1998. Prions. *Proc. Natl. Acad. Sci. USA* **95**:13363-13383.

Prusiner, S.B. and M.R. Scott. 1997. Genetics of prions. *Annu. Rev. Genet.* **31**:139-175.

Puckett, C., P. Concannon, C. Casey, and L. Hood. 1991. Genomic structure of the human prion protein gene. *Am. J. Hum. Genet.* **49**:320-329.

Rest, J.S, J.C. Ast, C.C. Austin, P.J. Waddell, E.A. Tibbetts, J.M. Hay, and D.P. Mindell. 2003. Molecular systematics of primary reptilian lineages and the tuatara mitochondrial genome. *Mol. Phylogenet. Evol.* **29**:289-297.

Reyes, A., G. Pesole, and C. Saccone. 2000. Long-branch attraction phenomenon and the impact of among-site rate variation on rodent phylogeny. *Gene* **259**:177-187.

Rokas, A., and P.W. Holland. 2000. Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol.* **15**:454-459.

Schätzl, H.M., M. Da Costa, L. Taylor, F.E. Cohen, and S.B. Prusiner. 1995. Prion protein gene variation among primates. *J. Mol. Biol.* **245**:362-374.

Smulders, R.H., I.G. van Geel, W.L. Gerards, H. Bloemendal, and W.W. de Jong. 1995. Reduced chaperone-like activity of alpha A(ins)-crystallin, an alternative splicing product containing a large insert peptide. *J. Biol. Chem.* **270**:13916-13924.

Swamynathan, S.K., and J. Piatigorsky. 2002. Orientation-dependent influence of an intergenic enhancer on the promoter activity of the divergently transcribed mouse Shsp/alphaB-crystallin and Mkbp/HspB2 genes. *J. Biol. Chem.* **277**:49700-49706

Takai, D., and P. A. Jones. 2003. The origins of bi-directional promoters – computational analyses of intergenic distances in the human genome. *Mol. Biol. Evol.* Advance Access Epub. Dec. 4.

van Dijk, M.A., M.A. Sweers, and W.W. de Jong. 2001. The evolution of an alternatively spliced exon in the alphaA-crystallin gene. *J. Mol. Evol.* **52**:510-515.

van Montfort, R.L., C. Slingsby, and E. Vierling. 2002. Structure and function of the small heat shock protein/α-crystallin family of molecular chaperones. *Adv. Protein. Chem.* **59**:105-156

Venkatesh, B., M.V. Erdmann, and S. Brenner. 2001. Molecular synapomorphies resolve evolutionary relationships of extant jawed vertebrates. *Proc. Natl. Acad. Sci. USA* **98**:11382-11387.

Werten, P.J., B. Röll, D.M. van Aalten, and W.W. de Jong. 2000. Gecko iota-crystallin: how cellular retinol-binding protein became an eye lens ultraviolet filter. *Proc. Natl. Acad. Sci. USA* **97**:3282-3287.

Wistow, G. 1993. Lens crystallins: gene recruitment and evolutionary dynamism. *Trends Biochem. Sci.* **18**:301-306.

Wistow, G. 1995. Molecular Biology and Evolution of Crystallins: Gene Recruitment and Multifunctional Proteins in the Eye Lens. R. G. Landes Company, Austin, Texas, U.S.A.

Wistow, G.J., J.W. Mulders, and W.W. de Jong. 1987. The enzyme lactate dehydrogenase as a structural protein in avian and crocodilian lenses. *Nature* **326**:622-624.

Woodburne, M.O., T.H. Rich, and M.S. Springer. 2003. The evolution of tribospheny and the antiquity of mammalian clades. *Mol. Phylogenet. Evol.* **28**:360-385.

Wopfner, F., G. Weidenhofer, R. Schneider, A. von Brunn, S. Gilch, T.F. Schwarz, T. Werner, and H.M. Schätzl. 1999. Analysis of 27 mammalian and 9 avian PrPs reveals high conservation of flexible regions of the prion protein. *J. Mol. Biol.* **289**:1163-1178.

Wray, G.A., M.W. Hahn, E. Abouheif, J.P. Balhoff, M. Pizer, M.V. Rockman, and L.A. Romano. 2003. The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20**:1377-1419.

Samenvatting

De genen van een organisme bevatten het verhaal van hun evolutie. Vergelijking van de nucleotidenvolgorden van overeenkomstige genen van verschillende organismen geeft inzicht in de evolutie van die genen én in de verwantschappen tussen de betreffende organismen. Dit proefschrift geeft voorbeelden van beide aspecten: moleculaire evolutie en moleculaire fylogenie. Hoofdstuk 1 legt uit hoe deze twee onderzoeksgebieden verweven en afhankelijk van elkaar zijn. Om optimale modellen te ontwikkelen voor de reconstructie van fylogenetische verwantschappen is een grondig inzicht vereist in moleculaire evolutieprocessen. En om de evolutionaire veranderingen in genen en de daaruit afgeleide eiwitten te interpreteren moeten we de verwantschappen tussen organismen kennen.

Moleculaire zoogdierfylogenie

Hoofdstuk 2 laat zien hoe genen het inzicht in de fylogenie kunnen misleiden, maar ook hoe dit opgelost kan worden. Zoogdieren worden ingedeeld in drie groepen: de meest talrijke placentale (of 'hogere') zoogdieren, de buideldieren en de eierleggende zoogdieren, waarvan vogelbekdier en mierenegel de enige vertegenwoordigers zijn. Overtuigende morfologische kenmerken duiden er op dat de eierleggende zoogdieren als eersten aftakten van de zoogdierboom en dat buideldieren en placentale zoogdieren dus het meest verwant aan elkaar zijn. Op grond van analyses van genen uit het mitochondriële genoom stelden onderzoekers in 1996 echter dat buideldieren meer verwant zouden zijn met de eierleggende dan met de placentale zoogdieren. In hoofdstuk 2 toont onderzoek van 18 genen uit het kerngenoom nu aan dat de morfologische opvatting toch de juiste is. Alle analyses geven aan dat de buideldieren het meest verwant zijn met de placentale zoogdieren en plaatsen de eierleggende zoogdieren als nazaten van de primitieve zoogdierstam (Fig. 2, blz. 26). Deze conclusie wordt ondersteund door specifieke deleties van nucleotiden in twee genen van buideldieren en placentale zoogdieren (Fig. 3, blz. 29). Uit de snelheid waarmee de onderzochte genen geëvolueerd zijn is berekend dat de eierleggende zoogdieren tussen de 223 en 198 miljoen jaar geleden aftakten van de andere zoogdieren en dat de buideldieren en placentale zoogdieren 173-151 miljoen jaar geleden uit elkaar gingen. De misleidende fylogenetische informatie uit de mitochondriële genen van de eierleggende zoogdieren blijkt te wijten aan de afwijkende basensamenstelling van deze genen; wanneer daarmee rekening wordt gehouden in de fylogenetische analyses wordt geen steun meer gevonden voor een verwantschap van eierleggende zoogdieren en buideldieren.

Het fylogenetisch nut van deleties wordt ook geïllustreerd in hoofdstuk 3. Binnen de placentale zoogdieren, waarvan er 18 orden zijn, is omstreden welke orden de naaste verwanten van onze eigen orde, de Primates, zijn en met name ook wat de positie van de knaagdieren in de stamboom is. Twee deleties in verschillende genen laten in dit hoofdstuk overtuigend zien dat de primaten tot een groep behoren die ook de knaagdieren, konijnachtigen, de boomspitsmuizen en de vliegende lemurs omvat (Fig. 1, blz. 38). Dit bevestigt recente bevindingen gebaseerd op grootschalige analyses van genvolgordes en maakt het noodzakelijk om de morfologische opvattingen over de indeling van de zoogdierorden te herzien.

Moleculaire evolutie van ooglenseiwitten

De hoofdstukken 4 en 5 gaan over de genen die coderen voor de eiwitten in de ooglens van het vogelbekdier. De lenseiwitten αA- en αB-crystalline zijn eerder onderzocht bij vele gewervelde dieren en bleken in veel gevallen fylogenetisch informatief. Gehoopt werd dat dit ook het geval zou zijn bij het vogelbekdier. Hoofdstuk 4 vergelijkt de aminozuurvolgorde van de α-crystallines van

het vogelbekdier met die van een aantal andere vertebraten. Helaas bleken de α-crystallines geen duidelijke informatie te verschaffen over de plaats van het vogelbekdier in de stamboom van de zoogdieren. Wel werd er relevante informatie verkregen over een van de andere onderzochte dieren, de raadselachtige brughagedis of tuatara, een 'levend fossiel' uit Nieuw Zeeland. Dit dier bleek op grond van zijn αA-crystalline het nauwst verwant met de andere onderzochte hagedissen en niet een oude aparte tak te vertegenwoordigen.

Hoewel fylogenetisch niet informatief, heeft het vogelbekdier αB-crystalline wel een evolutionair interessante zes maal herhaalde aminozuursequentie Phe-Pro-(Ala/Thr) in een deel van het eiwit dat functioneel van belang geacht wordt. Ook bleek het αA-crystalline gen van het vogelbekdier twee vormen van het eiwit te leveren. Naast het gebruikelijke αA-crystalline van 173 aminozuren was er een vorm waarin 23 extra aminozuren geïnserteerd zijn tussen de posities 63 en 64. Dit verschijnsel, veroorzaakt door zogenaamde alternatieve splicing van het messenger RNA, komt bij veel andere genen voor, maar was wat αA-crystalline betreft tot nog toe alleen bij enkele placentale zoogdieren gevonden. Nader onderzoek toonde aan dat het stukje DNA dat de informatie bevat voor de extra 23 aminozuren terecht gekomen moet zijn in het αA-crystalline gen nadat de zoogdieren waren afgetakt van de andere gewervelde dieren, maar vóórdat de eierleggende en overige zoogdieren uit elkaar gingen. Hoe dit stukje DNA in het gen gekomen is en waarom het gedurende meer dan 200 miljoen jaar gehandhaafd is bij veel zoogdieren – terwijl het geen bekende functie heeft - blijft echter een raadsel.

De ooglenzen van de meeste zoogdieren hebben een ongeveer vergelijkbare samenstelling van een beperkt aantal specifieke eiwitten, crystallines genoemd. Hoofdstuk 5 laat zien dat in de ooglens van het vogelbekdier 16-18% van het eiwit echter bestaat uit een onbekende eiwitcomponent. Bepaling van de bijbehorende DNA volgorde identificeerde dit eiwit als lactaatdehydrogenase A (LDH-A), een enzym dat normaal in kleine hoeveelheden in spierweefsel voorkomt. Het verschijnsel dat bepaalde enzymen in zeer grote hoeveelheden voorkomen als structuureiwitten in de ooglens is bekend van een aantal andere gewervelde dieren, maar komt weinig voor bij zoogdieren. Een nauw verwant enzym, LDH-B, komt als lenseiwit voor bij veel vogels en krokodillen, maar LDH-A was niet eerder als crystalline gevonden. Twee leden van de LDH eiwitfamilie zijn dus onafhankelijk van elkaar in verschillende vertebraten 'gerecruteerd' als lenseiwit, wat er op wijst dat deze eiwitten - en hun genen - geschikte eigenschappen hebben voor dit doel.

Het gen voor αB-crystalline ligt bij de mens kop-aan-kop met het gen van het verwante 'kleine heat-shock eiwit' HspB2, op een afstand van slechts 900 basenparen (bp). Het is recent gebleken dat zulke kop-aan-kop gelegen genenparen, waarbij de regulerende promotergebieden van beide genen overlappen, onverwacht veel voorkomen in het menselijk genoom, tot wel 10% van alle genen. En dat terwijl slechts 1% van het genoom uit eiwit-coderende genen bestaat, die dus alle ruimte hebben om verspreid over de chromosomen voor te komen. Er is nog nauwelijks iets bekend over de evolutionaire oorsprong en functionele gevolgen van dit merkwaardige verschijnsel. In hoofdstuk 6 is daarom onderzocht of deze kop-aan-kop ligging van de genen voor αB-crystalline en HspB2 ook bij andere zoogdieren en bij vogels voorkomt. Dit bleek het geval te zijn, waarbij de afstand tussen de genen variëert van 645 bp bij het vogelbekdier tot 1069 bp bij de opossum en ongeveer 1600 bp bij kip en eend. Vergelijking van de DNA volgordes van de gebieden tussen de twee genen toonde dat sommige regulerende elementen aanwezig zijn bij alle onderzochte soorten, maar dat andere elementen slechts bij bepaalde soorten voorkomen. Zo zijn er bij placentale zoogdieren en buideldieren twee HSE elementen - die er voor zorgen dat het αB-crystalline gen bij hogere temperaturen opgeschakeld wordt - terwijl het vogelbekdier slechts één HSE heeft en vogels zelfs géén. De functionele betekenis van deze verschillen is onderzocht door de activiteit van de intergene gebieden van rat en vogelbekdier te vergelijken, zowel na het inbrengen in gekweekte lens- en spiercellen als tijdens de ontwikkeling van embryos van de klauwpad. Geconcludeerd wordt dat de regelelementen in het intergene gebied samenwerken bij het bepalen van de activiteit van de twee genen en dat daarom de korte afstand en kop-aan-kop ligging gehandhaafd moet blijven.

Moleculaire evolutie van het prioneiwit

Hoofdstuk 7 laat aan de hand van het prioneiwit zien hoe vergelijkend onderzoek veel informatie verschaft over de evolutie van een eiwit en tevens over de verwantschap van de onderzochte zoogdieren. Het prioneiwit is het enige eiwit dat 'besmettelijk' kan zijn; het is de veroorzaker van de gekke koeien ziekte (BSE) en de ziekte van Creutzfeldt-Jacob (CJD) bij de mens. Het prioneiwit komt bij alle zoogdieren voor, voornamelijk in de hersenen, met een nog onbekende functie. De normale cellulaire vorm van het prioneiwit (PrP^{C}) kan overgaan in een pathogene vorm (PrP^{Sc}, van scrapie, zoals de ziekte bij schapen heet) die onoplosbare aggregaten vormt waardoor de hersenen degenereren. De overgang van PrP^{C} in PrP^{Sc} treedt zeer zelden spontaan op, maar wordt bevorderd door bepaalde erfelijke afwijkingen in het prioneiwit en kan ook tot stand komen wanneer PrP^{C} in contact komt met PrP^{Sc}, van dezelfde of van een andere diersoort. Vandaar de angst voor de consumptie van vlees van BSE koeien.

Tot nog toe was de volgorde van het prioneiwit slechts bekend van een beperkt aantal zoogdieren, voornamelijk primaten, hoefdieren en knaagdieren. Een bredere vergelijking van zoogdierprionen zou licht kunnen werpen op de raadselachtige functionele en pathogene eigenschappen van dit eiwit. Voor hoofdstuk 7 werden daarom de genvolgordes bepaald van de prionen van 26 soorten placentale zoogdieren, afkomstig uit alle 18 orden. Een fylogenetische boom, geconstrueerd op grond van de prionsequenties, bleek goed overeen te komen met de thans bekende verwantschappen tussen de betreffende zoogdieren (Fig. 4, blz. 95). Dit wijst er op dat het priongen op een normale en regelmatige wijze evolueert binnen de zoogdieren. Vergelijking van de prionvolgordes toonde dat aminozuren die van structureel belang geacht worden - voor juiste vouwing, modificatie en transport van het eiwit - strikt geconserveerd zijn binnen de zoogdieren. Een aantal aminozuren waarvan gesuggereerd is dat ze functioneel belangrijk zijn (o.a. histidines voor koperbinding en een hydroxyleerbare proline) blijken niet geconserveerd. Ook een aantal zeldzame aminozuurmutaties die bij de mens in verband gebracht zijn met erfelijke vormen van CJD blijken bij andere soorten normaal voor te komen. Deze bevindingen roepen twijfel op over de betekenis van deze aminozuren voor functie en pathogeniciteit.

Bij de meeste soorten bevat het prioneiwit een vijf- of zesmaal herhaalde volgorde van 8 aminozuren. Bij de mens is gerapporteerd dat een erfelijke variant met 4 van deze 'octarepeats' geen gevolgen heeft, maar dat de aanwezigheid van een prion met 2 repeats geassociëerd zou zijn met CJD. In hoofdstuk 7 worden, bij enkele soorten, prionvarianten gevonden met 7 of 4 repeats en de enige onderzochte eekhoorn bleek zelfs een prionvariant met 2 repeats te hebben, naast de normale vorm met 5 repeats. Dit suggereert dat de betreffende eekhoorn heterozygoot was voor allelen met 2 en 5 repeats, en zoals bij de mens, een ernstig risico op 'gekke eekhoornziekte' zou lopen. In een vervolgonderzoek, kort beschreven in hoofdstuk 8, werd DNA uit oortjes van een groter aantal eekhoorns onderzocht. Gevonden werd dat alle eekhoorns zowel de prionvariant met 2 als met 5 repeats bezaten. Dit betekent dat de eekhoorn een duplicatie van het priongen heeft ondergaan, waarna in een van de gencopieën 3 repeats verloren zijn gegaan. De aanwezigheid van het gen met 2 repeats kan kennelijk geen kwaad doordat ook het normale gen met 5 repeats homozygoot aanwezig is. Een prioneiwit met twee repeats is uniek en vraagt om verder onderzoek.

Hoofdstuk 8 tenslotte relateert de bevindingen van het promotieonderzoek aan de huidige stand van zaken in het onderzoeksgebied. Tevens worden in hoofdstuk 8 enkele resultaten besproken die in de voorafgaande hoofdstukken niet aan de orde konden komen. Naast de al genoemde tuatarafylogenie en het eekhoornprion betreft dat de volgordebepaling van enkele kleine heat-shock eiwitten (sHsps), o.a. van het lancetvisje. Samen met sHsps van alle vertebratenklassen en lagere chordaten - verkregen uit de sequentiedatabanken van o.a. de complete genoomvolgordes van *Fugu* (kogelvis) en *Ciona* (zakpijp) - is daar een fylogenetische boom van gemaakt die de evolutie van de sHsp familie binnen de chordaten beschrijft.

List of Publications

1. van Oppen, M.J., Willis, B.L., van Rheede, T. and Miller, D.J. (2002) Spawning times, reproductive compatibilities and genetic structuring in the *Acropora aspera* group: evidence for natural hybridization and semi-permeable species boundaries in corals. *Mol. Ecol.* **11**: 1363-76.

2. Poux, C., van Rheede, T., Madsen, O. and de Jong, W.W. (2002) Sequence gaps join mice and men: phylogenetic evidence from deletions in two proteins. *Mol. Biol. Evol.* **19**: 2035-2037.

3. van Rheede, T., Smolenaars, M.M.W., Madsen, O. and de Jong, W.W. (2003) Molecular evolution of the mammalian prion protein. *Mol. Biol. Evol.* **20**: 111-121.

4. van Rheede, T., Amons, R., Stewart, N. and de Jong, W.W. (2003) Lactate dehydrogenase A as a highly abundant eye lens protein in platypus (*Ornithorhynchus anatinus*): upsilon-(υ)-crystallin. *Mol. Biol. Evol.* **20**: 994-998.

5. de Jong, W.W., van Dijk, M.A.M., Poux, C., Kappé, G., van Rheede, T. and Madsen, O. (2003) Indels in protein-coding sequences of Euarchontoglires constrain the rooting of the eutherian tree. *Mol. Phylogen. Evol.* **28**: 328-340.

6. Franck, E., Madsen, O., van Rheede, T., Ricard, G., Huynen, M.A. and de Jong, W.W. (2004) Evolutionary diversity of vertebrate small heat shock proteins. (submitted)

7. Doerwald, L., van Rheede, T., Dirks, R.P, Madsen, O., Rexwinkel, R., van Genesen, S.T., Martens, G.J., de Jong, W.W. and Lubsen, N.H. (2004) Sequence and functional conservation of the intergenic region between the head-to-head genes encoding the small heat shock proteins αB-crystallin and HspB2 in the mammalian lineage. (submitted)

Curriculum Vitae

Teun van Rheede werd geboren op 23 januari 1973 in Utrecht. Na het behalen van zijn VWO diploma aan het Herman Jordan Lyceum in Zeist, studeerde hij van 1992 tot 1998 Biologie aan de Universiteit van Amsterdam. Tijdens zijn eerste stage, bij de vakgroep Moleculaire Microbiologie van de UvA, bestudeerde hij antilichamen tegen ferredoxine $NADP^+$ oxidoreductase van de cyanobacterie *Synechocystis* onder begeleiding van Jasper van Thor en Hans van der Spek. Zijn tweede stageonderzoek verrichte hij aan de James Cook University, Townsville, Australië. Hij onderzocht daar de genetische variatie en hybridisatie in het koraalgeslacht *Acropora* onder begeleiding van Madeleine van Oppen. Hij schreef scripties over "Structuur-functie relaties in het eiwit ferredoxine $NADP^+$ oxidoreductase" en over "De evolutie van de genetische code". Tijdens zijn studie volgde hij extra vakken als algemene psychologie, neuropsychologie en statistiek en vervulde hij veel student-assistentschappen (o.a. biochemie, computerpractica, biomathematica, bioinformatica).

Naast zijn studie was hij o.a. voorzitter van de studievereniging voor biologen "Congo", lid van verschillende commissies binnen deze vereniging (van kook- tot introductiecommissie), redacteur van het informatiebulletin van de Faculteit Biologie "Nieuwsflits" en van het maandblad van Congo "Gymnorhina", lid van het Studenten Advies College dat het College van Bestuur en de Universiteitsraad van de UvA adviseerde, en bestuurslid van de Commissie Beta-bedrijven contactdagen aan de UvA en de Vrije Universiteit Amsterdam. Daarnaast was hij rondleider bij de Hortus Botanicus van Amsterdam en interviewer over cannabisgebruik ten behoeve van een onderzoek van de Faculteit Sociale Geografie van de UvA.

Vanaf 1 januari 1998 was hij als onderzoeker-in-opleiding (0,8 fte) werkzaam bij de afdeling Biochemie (FNWI) van de Katholieke Universiteit Nijmegen en verrichtte daar het in dit proefschrift beschreven onderzoek. In die tijd nam hij deel aan workshops, symposia en congressen op het gebied van moleculaire fylogenie en evolutie in Parijs, Montpellier, Ischia, Hoor (Zweden), Faro (Portugal), Keulen, IJsland en Sorrento. Hij was o.a. lid van de Onderwijscommissie van de onderzoekschool Institute of Cellular Signalling van de KUN. Naast zijn promotiewerk volgde hij een opleiding mesologie te Amsterdam. In november 2002 vertrok hij naar Nieuw Zeeland voor een onderzoeksstage in de groep van Dr. David Penny aan de Massey University in Palmerston North. Hij moest daarvan voortijdig terugkeren.

Teun overleed op 21 mei 2003 te Amersfoort.

Dankwoord

Volgens goed gebruik: het feit dat je dit proefschrift open voor je hebt liggen, betekent waarschijnlijk dat je op wat voor manier dan ook hebt bijgedragen aan het tot stand komen ervan. Daarom: bedankt!

Wilfried, als mijn promotor en directe begeleider is er heel wat waar ik je voor wil bedanken. De grote vrijheid te doen waar ik zin in had, altijd tijd voor overleg van wetenschappelijke, persoonlijke of lab-technische aard. De prettige omgang in het lab, tijdens congressen en netwerkbijeenkomsten. De inwijding in de omgangsvormen die gelden binnen de wetenschap die ik kreeg en die je zelf zo correct beheerst. Het (over)enthousiasme dat we delen voor alles dat met (moleculaire) evolutie te maken heeft, bij jou hoogstens wat getemperd door ervaring hoeveel hooi er op een vork kan. Maar bovenal, en dat is bijzonder in de wereld van de wetenschap, je menselijkheid. Met enige regelmaat is het terug te vinden in dankwoorden van proefschriften dat de promotor een beleefd geformuleerde sneer meekrijgt ('wetenschappelijk gezien was je in het eerste jaar een goede begeleider'). Waar er ruimte is om ziek te zijn, waar je vier dagen per week kan werken, waar geïnformeerd wordt hoe het op persoonlijk vlak met je gaat: daar telt mee wat goed is voor een wetenschapper als mens, en dat is prettig werken.

In het lab: m'n tafelgenoten/U-maatjes Bas, Guido en Sándor. Buiten het lab waren dezelfde mensen goed voor vele etentjes, avondjes bioscoop en de kroegbezoekjes aan „De Blauwe Hand" en de terrasjes die schijnbaar onmisbaar zijn om een promotietijd tot een goed einde te brengen. Ole, de constante waarde in het lab wanneer het aankomt op fylogenetische methode en PCR. In m'n tijd in Nieuw-Zeeland en de eerste pijn, heb ik vaak gedacht aan hoe je vertelde over het Russische ruimtehondje Leica en dat het dus toch altijd nog erger kon... De komst van de groep Lubsen naar het Trigon was voor mij een welkome aanvulling op de biochemiegroep. Zowel wat betreft het aanleren van traditioneel uitgevoerde moleculaire technieken, begeleiding bij lens en hsp-vragen (Lettie en Siebe bedankt) als natuurlijk de gezelligheidsimpuls. 'Nijmeegse' Linda, samen brachten we heel wat micropauzes door en ik ben blij dat je - voor mij eigenlijk onverwacht - uitgroeide tot een goede vriendin.

Zelf leren is één, voor mij volgt bij voorkeur zo snel mogelijk lesgeven en overdragen aan anderen. In Nijmegen was het niet makkelijk om dat van de grond te krijgen, dus ik was erg blij dat 'mijn' studenten Marcel Smolenaars (prionen) en Trijntje Bastiaans (amniotenfylogenie) op mijn stage-postertjes reageerden. Twee zeer prettige en productieve samenwerkingen; bedankt voor jullie bijdragen! Tijdens een vakantie van Ole klopte Kai Ament af en toe bij mij aan, onder andere resulterend in trips naar Keulen en Parijs. Om het lijstje compleet te maken: Vivi werkte aan de mysterieuze prik lens cDNA bank, Remco Rexwinkel hielp in deeltijd (12%) met de laatste eindjes van het evolutionaire aspect van de intergene regio en de lancetvis hsp.

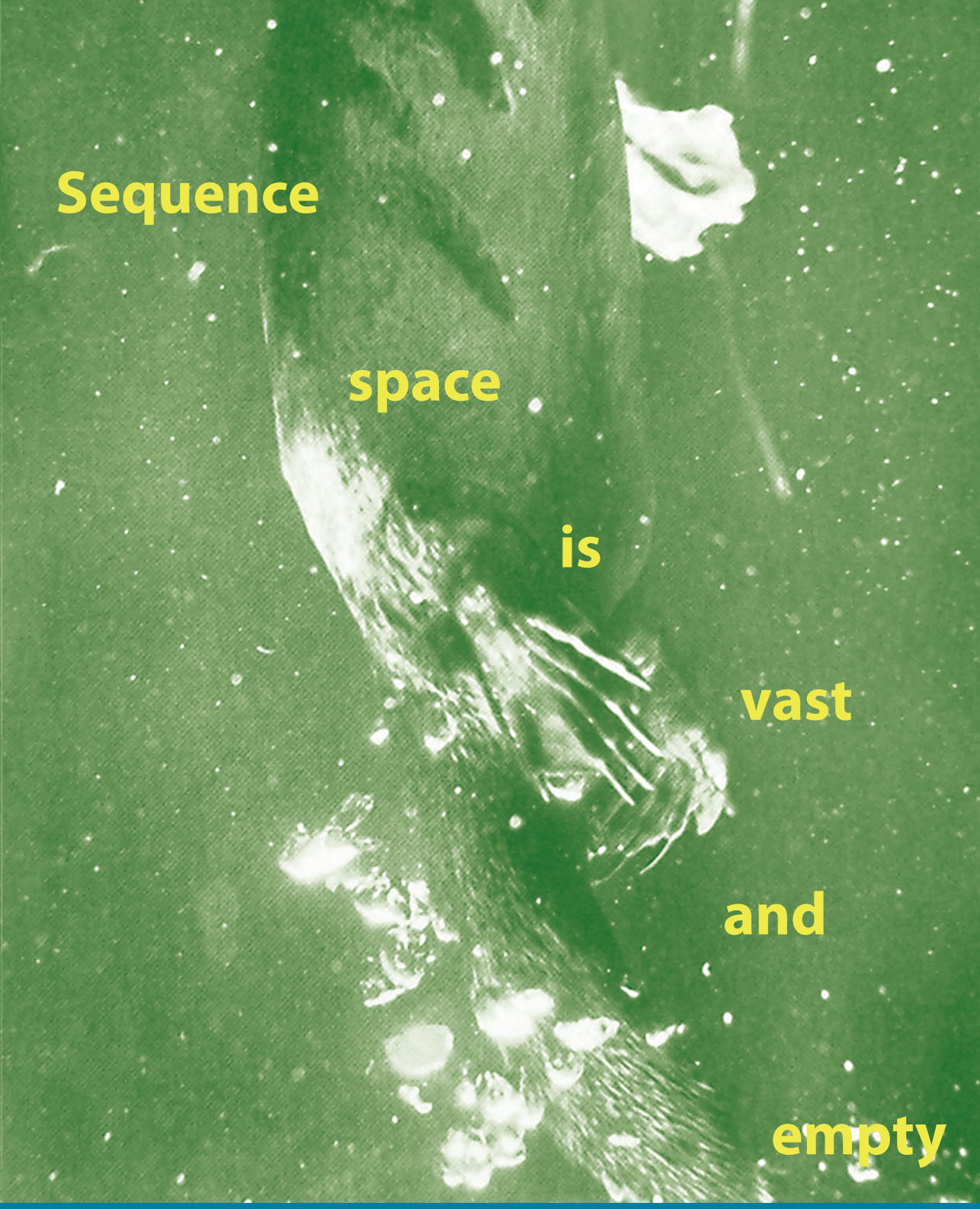
Het doen van onderzoek betekent voor mij werk, maar ook datgene doen wat ik leuk vind en waar ik goed in ben. Honderden enthousiaste plannen maken en er een klein deel van uitvoeren, de intellectuele uitdaging, wetenschappelijke discussies over alles wat wel en niet met evolutie heeft te maken. Naast dat leven in het lab zijn er andere mensen en activiteiten die het OIO-bestaan complementeren.

Cat, als enige directe link naar Amsterdam in Nijmegen, ik heb genoten van de avonden die we samen doorbrachten met film, roddelen en gesprekken over alles van het leven. Harriët, een echte Nijmeegse toevoeging, ik ben blij je ontmoet te hebben op de ijsbaan. Mensen met wie ik enkele voor mij bijzondere, momenten deelde: Christa, Hilde, Marlous, ik ben blij dat ik jullie tegenkwam.

De goede verstaander begrijpt het belang dat ik hecht aan reizen. Hier noem ik Eveline als m'n reisgenootje tijdens een paar van de topvakanties van m'n OIO-periode. Oostenrijk en Griekenland waren onvergetelijk (stel de verwezenlijking van je dromen niet uit tot na je pensioen; een beter voorbeeld kan ik niet meer voor je worden). Ik was meerdere malen in Egypte in de laatste jaren. Van woestijn tot koraalrif, tempel tot pyramide. De chaos, het temperament van de Egyptenaren. De laatste keren had ik het geleerd en voelde ik me er als een vis in het water; ondergedompeld in de stroom van het leven. De laatste keer, het rondje Cairo, Gizeh, Alexandria, Siwa oase, Cairo was een bijna magische ervaring (dankjewel Lin).

Studenten en docenten van de opleiding mesologie die ik drie jaar lang volgde. Toen Nijmegen na driekwart jaar wat benauwd begon te worden was het mijn redding. Veel lol met m'n klasje waar ik zoveel van ben gaan houden. De inzichten in mezelf, de mensen en de wereld om me heen die ik kreeg in de lessen en daartussendoor. Dit alles besproken en geëvalueerd op de terugkerende vrijdagavondetentjes van het goede leven. Naast de brede kennis over zowel reguliere als alternatieve/complementaire geneeswijzen die ik er kreeg, heb ik genoten van de verschillende zienswijzen die ik langzaam aanleerde. Met name de traditonele Chinese geneeswijzen en Ayur Veda. Terugkijkend zie ik een langzame overgang van m'n scherp-omlijnde, wetenschappelijke beta-denken, naar de Chinese denkwijze met z'n eindeloze samenhangen, patronen en afwezigheid van oorzaak-gevolg betrekking. Hoe ik in m'n eerste jaar kritische vragen afvuurde op de TCM-docent om aan te tonen dat dit toch helemaal niet kon, tot in m'n derde jaar, toen ik glimlachend in de klas kon zitten terwijl patronen en relaties zich ontvouwden. Nu bestaan ze vreedzaam naast elkaar, en o wat is het mooi om te zien dat die twee zo uiteenlopende, schijnbaar onverenigbare zienswijzen ieder op hun eigen manier over dezelfde werkelijkheid vertellen.

Het enige dat moeilijker is dan een proefschrift afmaken, is het om een proefschrift niet af te maken. Toen ik tijdens m'n ziekte niet meer verder kon werken aan m'n boekje, was de steun van anderen pas echt hard nodig. Wederom Wilfried: door je inzet om dit proefschrift er te laten komen, kon ik het een stuk makkelijker laten liggen. In die moeilijke periode was het vooral mijn familie die tot grote steun was, Toos vanwege de mooie, verhelderende gesprekken en Linda die was als mijn levenspartner.


Sequence
space
is
vast
and
empty