THE EVOLUTION OF THE VERTEBRATE BETA GLOBIN GENE FAMILY

ELIZABETH GABRIELA AGUILETA ESTRADA

Department of Biology
University College London

2004

Submitted to the University of London
for the degree of Doctor of Philosophy

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

The evolution of the vertebrate beta globin gene family

Elizabeth Gabriela Aguileta Estrada
University College London

D. Phil. Thesis
University of London

This thesis covers different aspects of the evolution of the vertebrate β globin gene family. A wealth of data on globins has been accumulated over decades of work in diverse areas, this information, together with the use of new methods, allowed a comprehensive analysis of β globins. First, a review on the current knowledge of gene family evolution is made and the general objectives of the thesis are stated. This introductory chapter is followed by the careful analysis of the β globin phylogeny comparing different reconstruction methods and discussing the differences between species and gene tree topologies. The molecular evolution of this gene family is investigated using codon models of sequence evolution. Particular emphasis is put on the role of gene conversion and positive selection acting at sites in the genes and along branches in the phylogeny. Also, several models of evolution by gene duplication are tested and results are analysed in the light of the different hypotheses on gene family evolution. The third chapter is devoted to the evolution of the globin protein structure from the analysis of sequence data. The ancestral state reconstruction of structurally relevant amino acids in different globins is conducted and the substitution pathway leading to the observed data is examined. The impact of amino acid changes in the hemoglobin protein is evaluated in terms of structural and functional constraints and the role of positive selection on the protein products of these genes is explored. Also, a possible case of coevolution between residues in the α and β subunits of hemoglobin is proposed. Finally, using new and more sophisticated methods, I estimate dates for gene duplication and gene divergence events in the β globin family. Two different methods of date estimation based on molecular data are compared and evolutionary rate variation in this gene family is tested.

CONTENTS

List of Tables

LIST OF FIGURES

Amadísima Hemoglobina mía:

(...) Qué privilegio, amadísima Hemoglobina mía, hablar con Vos de la luna. No sólo porque sois un quirurgo especializado de la sangre humana, sino porque sois mi médico de la sangre que hizo latir apresuradamente mi corazón y de cuyo impulso nació esta carta que ahora os envío, porque (...) con vos puedo hablar de la circulación de la sangre como con nadie.

Antonio Tabucchi (Se está haciendo cada vez más tarde)

Con todas sus imperfecciones y flaquezas esta tesis es para el bienamado cocinero, su emprendedora suegra y sus alegres cuñados

chanchi

ACKNOWLEDGEMENTS

I want to thank Professor Ziheng Yang for the opportunity he gave me to conduct my doctoral studies under his supervision, for his kind help, patient guidance and for opening a new area of research for me. I am particularly grateful to Dr. Joseph P. Bielawski, for this work would never have been completed without the valuable discussions I entertained with him about my work. Both Ziheng and Joe provided me with insightful comments and questions that guided me throughout my PhD studies. Thank you both for your dedication and support.

I want to thank my two examiners who contributed to improve this thesis. Their comments and suggestions were invaluable and the viva was a very illuminating experience.

I would also like to thank the companionship and friendship of all members of our research group, past and present, who made my stay memorable. Thanks to Stéphane, Maria, Wa, Lounes, Kathy, Paul, Becky, Dave and Samit.

Studying in the UK was possible only thanks to the sponsorship provided by the Mexican Council for Science and Technology (CONACYT). I am indebted to them, literally.

Living in London for over three years allowed me to meet wonderful people from many different countries. All in all, this is the most important part of the experience for me and I will always treasure their friendship. Thanks to kendo-man Fede, to the party animals David, Filipa and Zoe, to all the football people (the list would be long, long); also, to the majísimas Piedad, Marta, Elena y Keeley, to the Quartier Latin: Neith y Laura, Luis y Lupe, Lorenzinho, Marie y Tom, and to my tai chi friends and teachers.

Of all the people that contributed to this adventure from overseas, I am most indebted to the wonderful efforts in transatlantic logistics that my literary agent had to go through for the only reason that she is my mother. Thanks also to all my family who never failed to remember my birthdays and sometimes even sent lovely presents. To the Energizers and the Ramoses for keeping in contact with those loooong telephone calls and huge parcels. And also, in case they get to read this, to my friends back in Mexico who never tired to say they missed us and were, I believe, honest.

And last, thanks most of all to my beloved cook, the amazing Mr. Hernández.

Introduction

The concept of biological evolution has itself evolved from an uncertain beginning marked by heated debates, to the present state of generalised acceptance (albeit, with some heated debate every now and then). From the days of ancient Greece, and even before, in the uncharted history of East and West, people have tried to find an explanation for the variety and change observed in the surrounding world. To some, variation appeared to reflect a constant and fixed organization of the world. There had always been plenty of life forms, each distinct and performing a specific role, and there was no reason to believe they could change. This is a sort of static variety, where once created, organisms went on and on, unchanged and isolated in their uniqueness. Plato's fixed essences come to mind. To others, however, the world was an ever changing scenario full of transformations. Everywhere you looked you could see movement, change, births and deaths, renewal and decay. Some people saw change as the force that brings about multiplicity, which is a more dynamic explanation for variety. Heraclitus is perhaps the most renowned of those who saw change as the great force behind variety.

A long time has passed since the two opposing views of stasis and change were first built as philosophical constructs or explanations of the world. Since then, not only the problem of the origin of multiple life forms has plagued some minds, but also questions about the relation between those life forms, the characteristics they share and those which make them different. The classic example of the 18th century naturalist is Linnaeus, whose "Systema Naturae" (1756) followed well defined rules for classifying and naming living beings. That thorough book already points to an underlying order, an attempt to make sense of variety. However, Linnaeus did not offer an explanation other than God's might and wisdom as sources

Fig. 0.1. A painting of Carl Linnaeus (1707-1778) who published his work in his "Systema Naturae" (1756).

of the order observed in nature. Labelled and stashed into different boxes, species continued to exist, as it were, by somebody's sheer act of will.

As it so often happens, the big conceptual step forward did not occur within biology. While Linnaeus was amazed with the order seen in God's creation, astronomers and geologists were telling a different story of the change and variety they observed in the sky and the Earth. Laplace and Hutton, to name two remarkable examples, started to discuss celestial bodies engaged in transformation, the old age of the Earth and the spectacular forces involved in shaping our planet. The picture they envisioned was far from static. Things were starting to get dynamic.

Evolutionary ideas were, however, rather distrusted by the public. Even in academic circles, some disliked evolutionists. It was in this atmosphere that the most famous biologist of all time, Charles Darwin, champion of evolutionary biology, made his first appearance in the public scene. Natural history was his passion, so we are told, and a trip around the world changed his life, and the way we see life, for ever. If some say that philosophy can be read as a footnote to the works of Plato and Aristotle, others would say that biology is a footnote to Darwin's work.

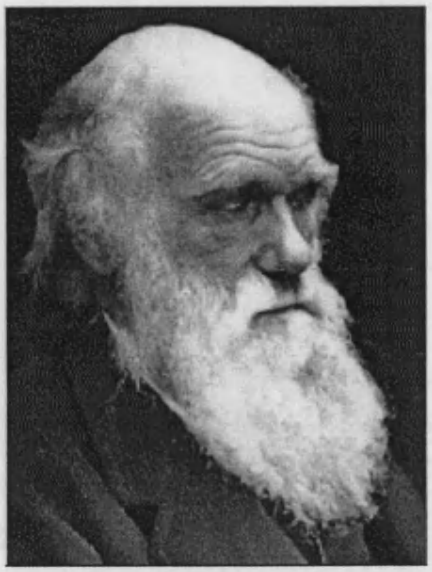


Fig. 0.2. Charles Darwin (1809-1882) in his late years. He published "The Origin of Species" in 1859.

Overstatements like this are, as usual, big oversimplifications but they reflect the impact some body of work has had on human history. With Darwin, and Wallace, came the quantum leap for the concept of evolution. Ernst Mayr summarized Darwin's contribution in five points: (1) evolution as such, (2) common descent, (3) gradualness, (4) speciation, and (5) natural selection. Perhaps natural selection was the most important departure from the conventional thought of species as fixed entities. With these concepts in the pocket, biology took off.

For a long time, the study of evolution has meant the evolution of organisms, of species. It was not until the Evolutionary Synthesis arrived, that genetics, and therefore, molecules, appeared as objects of evolutionary enquiry. Democritus is believed to be the first to think that matter is made of tiny indivisible components, atoms. However, from Democritus to Mendel and his particles of heredity (or genes), there is a long and winding road. Gregor Mendel would have remained an obscure Austrian monk with an inclination for horticulture and peas were it not for his rediscoverers, de Vries, Correns and von Tschermak who resurrected his writings in 1900. Without Mendel's work, Darwin's concept of natural selection would have

Fig. 0.3. Gregor Mendel (1822-1884) in his Augustinian attire. Mendel's work became the foundation for modern genetics

faltered, lacking a mechanistic explanation for the hereditary variations transmitted from parents to progeny.

The next big concept to appear was mutation, one of the sources of variation, developed after 1900. All the pieces of the puzzle started to fit together, although the genetic aspect of mutation (i.e. how do mutations change genes) was elucidated much later, only after the nature of genes themselves was established.

The first generation to bring together genetics and evolutionary biology included personalities like Mayr, Fisher, Wright, Haldane, Dobzhansky, Stebbins, Gaylord Simpson and Huxley. Geneticists, systematists and palaeontologists worked together to achieve a coherent theory that constitutes the foundation of modern biology. All of those pioneers introduced fundamental concepts that are still in use today: phenotype and genotype, mutation, recombination, and random genetic drift. Although much has been revised and extended since then, the synthetic theory established the field of evolutionary genetics and paved the way for the study of population genetics and molecular evolution (Futuyma 1998).

Perhaps the second major revolution in evolutionary biology, after Darwin, came with the discovery of the DNA structure by Watson and Crick in 1953, as well

as the confirmation of its role as the heritable genetic material by Avery. These two breakthroughs accelerated the work in the field to an astonishing extent and provided the physicochemical foundation of evolutionary studies that allowed a very detailed comprehension of the digital nature of information that makes up a living organism and the many sources of mutation, as well as its effects in genes and genomes.

Fig. 0.4. The celebrated Watson and Crick published their work on the structure of DNA in 1953

Molecular evolution as a distinct discipline in biology takes advantage of the work from systematists, geneticists, molecular biologists, and ecologists, as well as chemists, physicists, statisticians, and computer scientists. Those who work in the field are themselves academic mongrels, so to speak. In any case, molecular evolution is a rather young discipline but a vigorously growing one. The first works truly concerned with the natural history of biomolecules were those of the early researchers who analysed proteins electrophoretically to compare information from different organisms. In 1904 Nutall realized that serological cross-reactions were stronger for more closely related species than for more distantly related ones. This precedent gained full acknowledgement about half a century later, when in the 1950s sequencing techniques became available for proteins and then nucleic acids. History

has it that the first protein to be sequenced was insulin and that it was Sanger who patiently developed the method. It soon became apparent that changes among proteins did not occur randomly. Also, it was observed that the majority of changes were tolerated (i.e. the protein still worked) but that a few replacements seemed to have a very important and dramatic effect, as illustrated by the sickle-cell anaemia caused by one change in hemoglobin.

As in every scientific field, progress comes from explosive and often hotly debated new ideas. One of these ideas which has fuelled a good deal of the research done recently in molecular evolution is the suggestion that molecules evolve, accumulate changes, with a constant rate through time. Zuckerkandl and Pauling (1965) were the first to propose the so-called molecular clock hypothesis, which immediately gained advocates and opponents. The mere idea of constancy seemed to defy the most basic principles of evolution, based on change and variation. Palaeontologists were among the first to complain, mostly after the estimation of the divergence dates between apes and humans by Sarich and Wilson in 1967 because this date turned out to be much younger than the fossils suggested. Morris Goodman, a renowned primatologist, was one of the early detractors of the molecular clock hypothesis.

Another landmark in evolutionary studies came from a significant advance in population genetics with the advent of multispecies comparisons of DNA and protein sequences. Kimura (1968) found, through a survey of various protein sequences from the mammalian genome, that the substitution rate observed is too high to be explained by Haldane's "cost of natural selection". He then proposed that most mutations must come at no cost at all and be selectively neutral. A later elaboration of this theory, proposed by Ohta, added the possibility of having slightly deleterious mutations.

According to this "nearly neutral" theory, the effect of selection on the change of allele frequencies is similar to, or weaker than, that of random drift. King and Jukes proposed a similar theory in 1969.

Fig. 0.5. Motoo Kimura (1924-1994) proposed the neutral theory of molecular evolution

Also important was the discovery of a large extent of polymorphism within populations. New hypotheses tried to explain how such variation could be maintained by natural selection. In the opposite extreme, Kimura, King and Jukes, claimed that such polymorphisms were due to most substitutions being neutral and that they were not static but a transient stage in evolution. Kimura's theory is now known as the neutral theory of molecular evolution, which has come to be useful as a null hypothesis, as it allows tests to be easily formulated. The selectionist-neutralist divide had just begun, with each side claiming the predominance of selection or drift in evolutionary processes. Traditionally, systematists and taxonomists side with the selectionists and molecular biologists and biochemists with the neutralists.

With the availability of DNA and protein sequences from different organisms, phylogenetic trees for molecules and organisms began to be used to establish long-term evolutionary histories. Interestingly, these two kinds of trees often do not tell the

same story. The phylogenetic framework in molecular evolution studies has prompted many new questions and the development of new methods to infer trees and analyse the substitution process. Also, with molecular phylogenetics, evolutionary hypotheses can be related to change through time.

The most challenging questions are still related to the old inquiries such as, when did two genes or species diverged, is the evolutionary rate constant or not, which tree-making method is the most accurate, which method is best to determine genetic distance, what substitution model fits the data better, etc. Particularly relevant for this thesis is the study of gene families and specially the theories proposed regarding their origins and evolution, but that is more thoroughly discussed in the following chapters.

Chapter I is an overview of gene families as an object of study in molecular evolution. I focus on what is known about their evolution with a special emphasis on β globins, the case study in this thesis.

Chapter II presents the phylogeny of β globin genes in vertebrates. I discuss speciation, gene duplications and recombination in a phylogenetic framework. I test different models of gene family evolution using a maximum likelihood approach.

Chapter III covers various aspects of the evolution of the globin protein structure. I analyse the relevance of amino acid substitutions in the evolution of hemoglobin interfaces between subunits and the origin of oligomer globins from monomeric structures.

Chapter IV deals with the estimation of dates for gene duplication and gene conversion events in the vertebrate β globin gene family. I use and compare maximum likelihood and Bayesian methods.

Chapter I

-

Gene families: origin, evolution and perspectives

1.1 The organization of the genome in gene families

1.1.1 Gene families and genomes

Given the amount of sequencing projects under way, and the avalanche of sequences already available in databases, we currently have the necessary information to provide us with a panoramic view, not only of genes, but genome organization in general. With all the available information we are beginning to realize that most of the genes we know are part of gene families (Page and Holmes 1998). Today genomes can be seen as an organized ensemble of gene families and regulatory elements working in concert. For a long time, though, genetics was an isolated discipline that studied genes in an equally isolated fashion. A gene performed a given function and was investigated as a unit, a single unconnected entity. With more and more complex inter relationships being discovered among genes, these are now studied in a larger, cellular and genomic context. The elucidation of metabolic pathways and sophisticated gene expression circuitry has also contributed to the present understanding of genes as part of a larger level of organization. Actually, with all the sequencing projects and comparative analysis of genomes, it looks as if the number of genes that are known to be part of a gene family will continue to grow. Also, it has been suggested that gene family copy numbers are underestimated when non-genomic approaches are taken (Charlesworth et al. 2001). From the phylogenetic point of view, studying gene evolution in the context of gene families has represented a step forward in the understanding of complex processes affecting general trends such as evolutionary rates and substitution patterns. This context also provides more robust results, as a larger scope is taken. The study of genes and genomes is a two-way avenue of feedback, as understanding one contributes to understand the other (Charlesworth et al. 2001).

1.1.2 The definition of gene family

In her book "Evolution and variation of multigene families", Tomoko Ohta (1980a) defines multigene families as "a group of genes or nucleotide sequences with the following characteristics: multiplicity, close linkage, sequence homology, and related or overlapping functions". Hood et al. (1975) established three different categories for multigene families, loosely based on the function of the genes: (1) simple-sequence, (2) multiplicational, and (3) informational. This early definition has been updated through the years: for instance, pseudogenes have been included as members of gene families even though in strict terms they do not share function with the rest of the group (Li et al. 1981). Another interesting feature of gene families is the resurrection of dead genes by gene conversion (Martin et al. 1983). Also relevant to the organization of multigene families is that genes can be closely located in the same region of a single chromosome, in which case they are usually regulated in concert, or be located in altogether different chromosomes (Page and Holmes 1998). Even though in principle it is easy to determine if a given gene belongs to a family, some of the basic criteria to determine membership sometimes are not straightforward to establish, for instance, distinguishing between paralogous and orthologous genes can be difficult (Henikoff et al. 1997). It would seem that there is a threshold of sequence divergence beyond which it becomes increasingly more difficult to establish family relationships among genes. In any case, regardless of the definition, gene families are now seen as a prominent feature of genomic organization evolving in multiple and complex ways. Because they are dynamic evolutionary units, gene families vary in size and complexity. Gene number seems to depend to a great extent on the function or functions performed by the family (e.g. ribosomal RNA genes are ubiquitously

required in large amounts in the cells there are thousands of them, whereas some structural genes, such as globins, come only in a dozen or so).

1.1.3 Different classifications of gene families

The organization of genes into gene families has been exploited to construct nucleotide or protein databases, in other words, most databases are built with nucleotide or protein families as a unit on top of which more inclusive categories are added. Examples of family-based databases are SCOP (Murzin et al. 1995), CATH (Orengo et al. 1997), or DALI (Holm and Sanders 1995). Again, just what constitutes a family will vary from database to database. Some of them classify genes into functional categories (i.e. including those genes that perform a similar function); or by the expression groups they are part of (e.g. those genes that are part of operons), or even by homology (i.e. those genes with a common ancestor). Typically, there is a marked difference in what biologists and biochemists consider a family. Criteria for family membership qualification differ according to the interests of the researcher. For a biologist homology is the key requisite to establish membership to a family (Ohta 1980), whereas for biochemists, interested in structure and function more than in origin, structural similarity is enough to determine familiar status (Chothia 1992). Similarity can be seen among sequences and among structures and biochemists are satisfied with 20-35% similarity at the sequence level in order to group different proteins into a family (Orengo et al. 1999). In contrast, for a biologist 30% sequence similarity is no guarantee of common gene origin. This means that even if protein databases are loosely based on homology, their classifications are sometimes too lax to be useful for certain biological inquiries.

1.1.4 Gene families as a unit of evolution

It is not accidental that genes in a genome are organized in gene families. Not only does this organization allow organisms to increase the number of available gene functions and the amount of genes expressed but it also makes feasible a more sophisticated regulation of expression and the fine tuning of complex cellular tasks. In terms of generation of novelty, it appears to be enormously advantageous to innovate from pre-existing material rather than generate genes *de novo* (Ohno 1970). In fact, gene families provide evolution with ample material in the form of redundant gene copies on which to try new possibilities (Walsh 1995). From a given gene, several related but individually diverse functions can be created (Tatusov et al. 1997). The fact that new genes can perform a similar but diverse function presents the organism with a big opportunity to fine tune complex cellular tasks, as each gene in the family can perform a different role in the orchestration of function (Hughes and Nei 1989). Another advantage of gene families is given by the possibility of regulating gene expression according to different needs or in response to different cues, as in the case of genes expressed at different stages of development (Hardison 1998). It has also been pointed out that one of the reasons why gene families are a widespread feature of genomes is that multiple copies of the same gene act as buffers against gene decay by the accumulation of deleterious mutations (Gu et al. 2003). The original gene that undergoes duplication must be a functional gene that has been positively selected, in other words, the structure or function of its products is guaranteed. Also, having multiple genes working in concert may accelerate evolution, as more possibilities are explored simultaneously (Zhang 2003). Gene families constitute themselves a unit of evolution, as suggested by Hood et al. (1975), as evolution acts on them as a whole, not only on the individual genes. All the genes in a gene family can act and evolve in

concert, which again brings new possibilities into play and makes regulation more efficient. One characteristic of gene families is that gene innovation is driven by amplification of weak, ancillary functions rather than strong, established functions (Hooper and Berg 2003).

1.2 The origin of gene families by gene duplication

That most genes are members of gene families also suggests that one of the most important mechanisms in the generation of new genes is gene duplication, as copies of pre-existing genes are made available for evolutionary exploration. Ever since Ohno published his seminal work on gene duplication (1970), this mechanism has been seen as the main source of new genetic material.

1.2.1 Early discoveries

Haldane (1932) and Muller (1935) were perhaps the first to refer to gene duplication as a source of new genes. They recognized the potential of redundant gene duplicates to acquire new functions after accumulating divergent mutations. They reasoned that having more than one copy of a gene would allow the system to retain the original function in one of the genes while the other could explore other alternative functions. This idea would be confirmed later with experimental data, when Bridges (1936) reported the first observation of a gene duplication in the *Bar* locus in *Drosophila*. This locus was the product of the doubling of a chromosomal band that occurred in a mutant that had a considerable reduction in eye size.

After this experimental observation, theories spawned trying to predict the relevance of such a mechanism in the evolution of genomes. In 1951, Stephens published a paper entitled "Possible significance of duplication in evolution". This

was a first attempt to formulate all imaginable evolutionary avenues gene duplication could bring. More than a decade later, in 1967, Ohno recapitulated this line of investigation in his book "Sex Chromosomes and Sex-Linked Genes". This renewed interest proved fruitful and brought much attention to the field. Only two years later, in 1969, Nei published in Nature what is perhaps the first paper whose main subject was gene duplication, called "Gene duplication and nucleotide substitution in evolution". This led to the publication in 1970 of what is considered by many the seminal work of Ohno, "Evolution by Gene Duplication". In this book he proposed that gene duplication is the only way new genes can arise. Even though we now know this is not correct, the conclusions raised by Ohno remain largely valid. This publication immediately caught the attention of biologists and popularised the relevance of gene duplication as the most important source of new genes and functions. All these publications triggered new research on the mechanisms of gene duplication itself and its consequences. However, it was only until the main sequencing projects started that we gained a full realization of the extent of duplication and its role in genome evolution.

1.2.2 The extent of gene duplication in different genomes

There is a considerable variation in the number of gene duplication events reported in different organisms to date (Zhang 2003). Organisms in all three domains of life, Eukarya, Archaea and Bacteria, show extensive gene duplication among genomes (Himmelreich et al. 1996, Klenk at al. 1997, Rubin et al. 2000). One trend that becomes clear is that bacteria show fewer gene duplications (298 duplicate genes in *Mycoplasma pneumoniae*) than either archaebacteria (2436 duplicate genes in *Archaeoglobus fulgidus*) or eukaryotes (40 580 duplicate genes in human), with the

latter having considerably more than the organisms in other life domains (Zhang 2003). One easy interpretation would suggest that the more complex the organism the more likely to show extensive gene duplications. Even though this might be true to some extent, we still don't know exactly why this trend occurs. It could also be that traces of duplications in bacteria and archaebacteria have been erased by evolution. In any case we can realize that a large number of the genes in these organisms originated by gene duplication (Gu et al. 2002). However, the true extent of duplication cannot be known, as gene divergence creates large dissimilarities among old gene duplicates (Ohta 1990). It is likely that estimates of gene duplication are indeed underestimates (Charlesworth et al. 2001).

1.2.3 Different types of gene duplication

There are many types of gene duplication. These differ in the molecular mechanism involved and in the extent of the duplicated material. According to the extent of the duplication there are the following types: (i) partial or internal gene duplication, (ii) complete gene duplication, (iii) partial chromosomal duplication, (iv) complete chromosomal duplication, and (v) polyploidy or genome duplication (Graur and Li 1999). Together, the first four types of duplication are known as regional duplications because they do not involve the entire set of chromosomes.

Gene duplication generates two identical copies; if the two remain identical they are referred to as repeated genes, which can be classified as variant or invariant repeats (Lewin 2000). Invariant repeats usually arise because their protein products are required in large amounts, thus, identical gene copies performing the same function confer an advantage in terms of protein dosage (Page and Holmes 1998). rRNAs and tRNAs are a typical example of genes originated this way (Patthy 1999).

However, not all invariant repeats can be explained by dosage reasons, as in vertebrate genomes some invariant families have been found that have no function (Kondrashov et al. 2002). On the other hand, variant repeats are genes that maintain sequence similarity to a greater or lesser extent but perform different functions (Lewin 2000). Examples abound, as they relate to most members in gene families (Henikoff et al. 1997).

Genome duplication or polyploidy occurs when the cell nucleus contains multiple sets of chromosomes, either because the first meiotic division fails to occur or because chromosome number changes by smaller steps (Griffiths et al. 1999). The usual cause for this to happen is a lack of disjunction between all the daughter chromosomes after DNA replication. Polyploidy occurs far more often in plant genomes than in animal ones, or at least vertebrate genomes (Brown 2002).

The molecular mechanisms actually responsible for duplicating the material also vary. The main mechanism is unequal crossing-over (Ohta 1980) but there are also transpositions (including retropositions), segmental duplications (involving from 1000 to more than 200 000 nucleotides), and gene elongations (Lewin 2000). Also at a different level, duplication of protein-coding genes can involve exons that correspond to protein domains, so they are duplicated as units. This is known as exon or domain duplication (Patthy 1999). Unequal crossing-over occurs during mitosis between the two sister chromatids of a chromosome in a germline cell or during meiosis, between two homologous chromosomes (Griffiths et al. 1999). It is part of what are known as reciprocal recombinations because what happens in one chromosome is reflected on the other one as well. In the particular case of unequal crossing-over this process creates a sequence duplication in one chromatid or chromosome and a corresponding deletion in the other. Fig. 1.1 shows how unequal

crossing-over can generate gene family expansions and how the same process can create deletions or contractions. Genes that are part of gene families and therefore have a high sequence similarity are prone to experience unequal crossing-over, as with highly similar sequences it is difficult to align true homologs during meiosis or mitosis. The process may actually occur repeatedly and result in a homogenisation of gene family members (Ohta 1980).

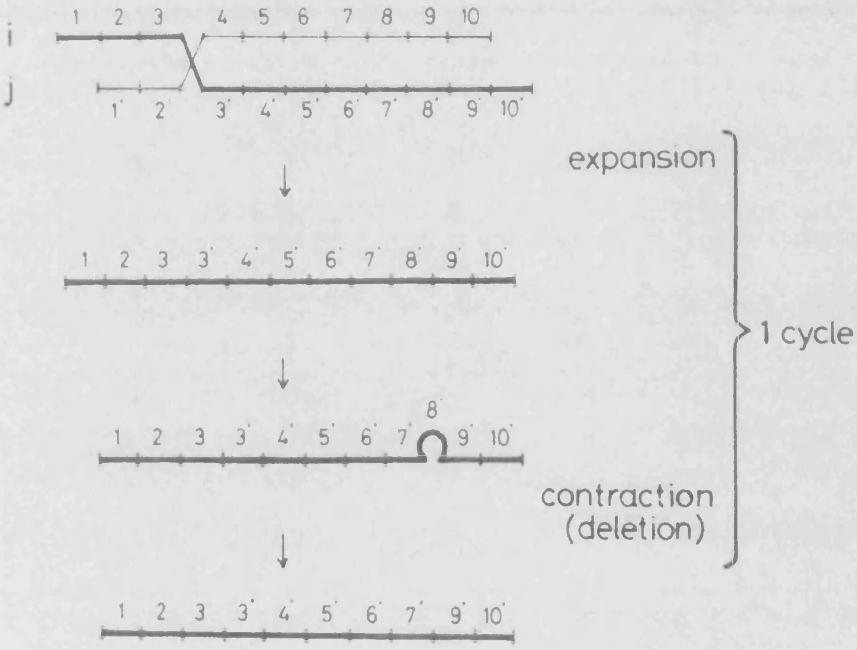


Fig. 1.1 Unequal crossing-over can create gene family expansions and contractions. Image reproduced from Ohta 1980.

1.3 The contribution of gene conversion in the evolution of gene families

Gene conversion is a kind of non-reciprocal recombination that occurs when two sequences interact such that one converts the other (Li 1997). Conversion occurs in different ways: intrachromatid conversion happens between two paralogous sequences on the same chromatid; sister chromatid conversion involves two paralogous sequences from complementary chromatids. Classical conversion occurs when two

alleles at the same locus exchange DNA. Semi classical conversion is an exchange between two paralogous genes from two homologous chromosomes. Non-allelic conversion, that is, exchanges between genes located at different loci, is thought to be the most important type in terms of evolutionary consequences (Page and Holmes 1998). Gene conversion may be biased or unbiased, in other words, it may or may not have a direction. In the case of biased conversion one gene preferentially converts the other (Graur and Li 1999). When this is the case, some authors refer to the converted copy as the slave and the converter as the master.

Gene conversion is a frequent phenomenon with complex and varied consequences. It can bring homogenisation (Jeffreys 1979, Slightom et al. 1980) or generate polymorphisms (Ohta 1998). The extent of the sequences involved in the exchange is variable, it can go from a few base pairs to a few thousand base pairs. The probability of occurrence of gene conversion depends on gene location and similarity among sequences, the more similar the more likely they are to experience conversion (Ohta 1980 a).

Unequal crossing-over and gene conversion transfer DNA sequences between genes, which results in those genes evolving together, that is, in concert (Graur and Li 1999). The process of concerted evolution is particularly relevant in the case of gene families because usually the members of a family have high sequence similarity and are therefore prone to experience unequal crossing-over and gene conversion. Concerted evolution also means that mutations can extend to all genes in a family, even if they are located in different chromosomes. As a result of concerted evolution, there exists the possibility of spreading advantageous mutations to all members in the family (Ohta 1980); also, since there is constant homogenisation of genetic material in the family, the evolution of new genes by divergent mutations is retarded

(Kondrashov et al. 2002), or it may prevent redundant copies from becoming non-functional (Krakauer and Nowak 1999), or even, to resurrect pseudogenes (Martin et al. 1983). It is also important that concerted evolution considerably complicates the study of phylogenetic relationships among genes, as it becomes very difficult to distinguish orthologous from paralogous genes (Henikoff et al. 1997). The evolutionary history of genes that have experienced unequal crossing-over and gene conversion is not described by a single phylogeny, but rather show a mixed history.

1.4 Models of Gene Family Evolution by Gene Duplication

1.4.1 The evolutionary fate of duplicated genes

All the models developed so far to explain gene family evolution by gene duplication are concerned with the fate of genes following gene duplication. The most important issue under discussion seems to be whether the redundancy in terms of function created by duplication can be stably maintained or not in the genome and what alternatives exist for a newly duplicated gene.

Nonfunctionalization.-In the framework of the neutral theory of molecular evolution, the most likely fate for a newly duplicated gene is to lose its original function, in other words, to become a pseudogene (Ohno 1970). This evolutionary fate is referred to as nonfunctionalization in the literature and it occurs as the redundant gene duplicate accumulates deleterious mutations, thus becoming either unexpressed or functionless. This is possible because the other duplicate is available to perform the original function. As we know from Walsh (1995) and Lynch and Conery (2000), pseudogenes usually arise in the following few million years after a duplication provided that the gene is not under selective constraint. They will persist in the

genome until they are deleted or diverge to the point where it is impossible to establish any relationship with the genes in the family they descended from. Pseudogenes are rather frequent in eukaryotic genomes, as an example, in *C. elegans* there have been identified 2168 pseudogenes, which is roughly one pseudogene for every eight functional genes (Harrison et al. 2001). This again reflects the prediction of the neutral theory of molecular evolution. Non-functional genes have been neglected as uninteresting elements in the genome, however, they are involved in many important and attractive phenomena. For instance, pseudogenes can serve as reservoir of genetic material for gene conversion, as it has been demonstrated in the chicken VH1 gene, which encodes the heavy chain variable region of the immunoglobulins (Ota and Nei 1995). In this way, pseudogenes contribute to the diversity of the immunoglobulins, which is directly reflected in the effectiveness of their function. Another interesting example of the dynamic nature of pseudogenes is the possibility of regaining their function, in other words, of being brought back to life. One example is provided by the seminal ribonuclease in cows, this gene has a paralogous gene, the pancreatic ribonuclease, which is expressed in all ruminants. That only in the cow the seminal ribonuclease is functional and in all other ruminants, that share the paralogous pair, this gene is ridden with deleterious mutations or is not expressed, suggests that it was a pseudogene that was resurrected in the cow lineage (Trabesinger-Ruef 1996).

Function retention.- Another alternative fate for newly duplicated genes is to retain their original function (Li 1997). Even if this would seem the most unlikely of fates given the cost to stably maintain two identical copies of the same gene, sometimes, high dosage requirements for a gene product make redundancy advantageous, as more

product is expressed from multiple copies of the same gene. This case is particularly easy to find among strongly expressed genes like rRNAs. Concerted evolution facilitates the retention of function among duplicated genes in a gene family or simply strong purifying selection ensures that genes remain unchanged. Nei et al. (2000) and Piontikivska et al. (2002) have shown that purifying selection plays a more important role in the maintenance of duplicated genes than previously thought. In their analysis they distinguished the effects of gene conversion in the maintenance of gene function from the effects of purifying selection by looking at synonymous (or silent) substitutions. While synonymous differences are not affected by purifying selection, they can be removed by gene conversion, as it operates on stretches of DNA whether these are synonymous or nonsynonymous (amino acid changing).

Subfunctionalization.- Gene functions need not only be lost or maintained, they can also be shared. In population genetic theory it is suggested that two duplicate genes can be maintained if they differ, even minimally, in some aspects of their function (Nowak et al. 1997). The sharing of gene functions can occur when genes descended from the duplication of a gene, whose function is divided among different subunits, each sets to perform one of the functions of the original gene. It thus becomes necessary to maintain the two copies in order to have the complete function of the parental gene. This process has come to be known as subfunctionalization. One of the most interesting ways to achieve this is through the differential expression of duplicate genes. This has been exemplified in numerous occasions for different gene families and is known to be important in the evolution of development (Jensen 1976, Orgel 1977, Hughes 1994, Force et al. 1999). Even though most examples of subfunctionalization come from regulatory sequences and involve partitioning of

function in terms of gene expression, it is possible to find function sharing among coding genes. In this case, the original gene codes for a protein with different functional or structural domains and the duplicate genes each performs the function of one of this domains separately. Though less abundant than the case of subfunctionalization by expression differences, a few cases have been reported to support this type of subfunctionalization (Dermitzakis and Clark 2001, Aguileta et al. submitted). It is also possible to find subfunctionalization at the protein level and it happens when one of the duplicate gene products better performs one of the functions of the original parental protein, thus leading to functional specialization (Hughes 1999).

Neofunctionalization.- The final alternative for newly duplicated genes, and by far the most attractive for biologists, is that of neofunctionalization or the acquisition of new gene functions (Li 1997). Under the neutral theory of molecular evolution, following duplication one of the two copies will retain the original function and the other copy will be selectively free to accumulate mutations, and even though these are mostly deleterious, occasionally advantageous mutations occur and a new function can arise. It is clear that in order to gain a new function, one of the duplicate genes must diverge through mutation, what is less clear is how this is possible in the first place. The first possibility is that the selective constraints that act on one of the copies are relaxed, as there is no need to maintain two identical copies. The alternative is that there is an increased ratio of nonsynonymous changes due to positive selection, which brings changes in the amino acid composition of the protein. In any case, the result is an open window for substitutions, some of which are beneficial and are thus selected for. Support for these two opposing mechanisms comes from examples in different gene

families. Ohta has cited the growth hormone-prolactin gene family case as a paradigm of evolution by gene duplication (1994), favouring relaxed selective constraints, while Goodman is a strong defender of positive selection as the cause of increased nonsynonymous rates of evolution (1981). The debate on the models of gene family evolution by gene duplication centres on how are gene copies maintained and what are the forces driving gene divergence.

1.4.2 Models for the evolution of gene families by gene duplication

I will now discuss the way each model describes the events that occur following duplication, paying special attention to the evolutionary forces they invoke, the predictions they make and the implications assumed.

a) The classical model.- From his studies in cytogenetics and biochemistry Susumu Ohno was among the first authors to recognize the significance of gene duplication as the source of new genes (1970). In his view, a newly duplicated gene had to be selectively neutral or advantageous for otherwise it would be deleted from the population. He viewed the outcome of gene duplications as dependent on a race between neofunctionalization and nonfunctionalization, where the most likely fate for a duplicate gene was to lose its function due to the accumulation of deleterious mutations. This first model assumed the tenets of the neutral theory of molecular evolution. According to Ohno's model, the rounds of mutations had to be recurrent in order for the genes to be either fixed or lost in the population. This view also required an immediate relaxation of selective constraints in one of the copies followed by bursts, as it were, of evolutionary change subsequent to a duplication event.

Tomoko Ohta's work since the late 1970s and all through the1980s and 1990s was largely devoted to the study of gene family evolution. She introduced the concept of nearly-neutral evolution and the mathematical studies of the process of unequal crossing-over and concerted evolution in gene families. Her studies at the population level added significantly to Haldane's and Ohno's first models of gene family evolution by gene duplication. She formulated what is now called the classical model of gene family evolution by gene duplication, although she referred to Haldane's model as the classical one. In Ohta's formulation, newly duplicated genes did not necessarily have to be either neutral or advantageous in order to escape nonfunctionalization and becoming pseudogenes, as Ohno suggested. She proposed that following duplication, new genes could also be slightly deleterious or slightly advantageous, or to use her phraseology, nearly neutral (Ohta 1992). This new aspect of the model allowed selection to play an important role. After genes duplicated, one of the copies was maintained through purifying selection while the other copy was subject to rounds of accumulated mutation which could either be deleterious (the gene became a pseudogene) or advantageous. If the latter was the case, purifying selection stably retained it. Here it is important to note that initially Ohta proposed that duplications are fixed by drift and then advantageous mutations are maintained through purifying selection but later on she recognized the role of positive selection in the generation of diversity by claiming there is an increase in the rate of nonsynonymous substitution immediately following a duplication event. Once a new gene has thus been generated, positive selection is no longer required. However, the issue of whether duplications can spread entirely by the action of drift or need the agency of selection is not entirely settled yet.

b) The DDC model.- Recently, a new model called DDC, for duplication-degeneration-divergence, has challenged the classical model. This model was proposed by Allan Force and his associates (1999). Even though other authors had rebuked certain aspects of the classical model before the DDC, I believe the latter is the most radically different view on the evolution of gene families by duplication, and it proposes a whole new mechanism for the generation of new genetic systems from pre existing ones. The observation that contrary to the predictions of the classical model, a large fraction of duplicated genes were maintained in the genome for long periods of evolutionary time, led to the investigation of alternative mechanisms of gene redundancy retention. In the DDC model, subfunctionalization is the fate of those duplicated genes that are retained in the genome and it explains how new functions can arise from the partition of pre existing functions among the new copies. The DDC model focuses on degenerative mutations affecting regulatory regions. It is proposed that following a gene duplication degenerative mutations can accumulate in different regulatory regions of the new duplicates. Since each copy will have a different deleterious mutation, the function they can now perform is partially affected and in order to recover the complete function of the original gene, the two copies must be expressed together. This sharing of functions would explain why nearly identical genes can be maintained in genomes for a long time and it is an explanation consistent with the theory of population genetics. This model differs radically from the classical one in that advantageous mutations are not a requisite to evolve new gene functions. Alternatively, it is expected that subfunctionalization will persist until one gene specializes in a different function or the two copies each diverge to perform two different and perhaps complementary functions. The DDC thus, opens the door for gene specialization and for a complex regulatory rearrangement of gene expression.

As I have mentioned already, the DDC is mainly concerned with complementary regulatory mutations among gene duplicates, however, it is possible to envisage a case where it is the original coding gene that encodes different functional or structural subunits in the protein, and the function of each domain is partitioned among the duplicate genes (Dermitzakis and Clark 2001). I have also suggested that a further form of subfunctionalization is possible (Aguileta et al. submitted; see Chapter II), one where the duplicated paralogs will perform a similar function as the original gene but they will be expressed at different developmental stages. Even though some divergence must occur following duplication, there need not be a burst of nonsynonymous substitutions because the paralogs can evolve under long-term differences in selective constraints. Here it is important to note that the DDC model does not exclude the possibility of neofunctionalization, that is, the DDC and the classical models are not mutually exclusive explanations of the same process. Actually, both neofunctionalization and subfunctionalization can occur at different stages in the evolution of gene families.

One of the critiques of the DDC model lies in the more or less explicit requirement that the original gene is itself partitioned in different modules or regulatory regions. Not only that but each "individual function" should contribute significantly to the general function, this means that it is important to determine the significance of multiple regions in order to test for subfunctionalization. Dermitzakis and Clark developed such a test called the paralog heterogeneity test (2001) and even though relatively effective and simple, it is based on the comparison of two orthologs to determine the pattern of substitution of each paralog. There is a limit to the power of their test method. Gu (1999, Wang and Gu 2001) approached this problem by incorporating phylogenetic information and the different substitution rates among

amino acids in order to detect functional divergence between paralogous genes. This allows the user to incorporate the information contained in multiple genes and is thus more powerful. It may be that adding information from the different codon positions, as would be possible with a codon-based model, would make it even more powerful to distinguish selective pressure between paralogs, yet again making the picture more detailed.

c) Other important contributions.- Since the classical model was proposed it has been challenged, criticized or expanded by different authors. Here I will talk about the major contributions to the discussion of gene family evolution by gene duplication by different researchers. Ohta made the first big challenge to the standing model when she added the concept of near neutrality to Ohno's model. That helped to build a more comprehensive model that incorporated selection in the dynamics of generation of new genes. It was Andrew Clark (1994) who made an investigation on the process of invasion and maintenance of a gene duplication. In this work he analysed the possibilities available for a newly duplicated gene following duplication and he concluded that even though fixation of a duplication could be reached entirely by drift, other forces must influence the fate of paralogs. He suggested that extra gene copies provided a buffer against deleterious mutations in the original gene and that recurrent duplication probably help to maintain chromosomal composition (i.e. prevents gene loss). His main conclusion was that a duplication can invade the population only if it provides an advantage to the organism. Later, Bruce Walsh (1995) published his work on the theoretical test of the probabilities that gene duplicates will become pseudogenes or will be fixed in the population. He assumed that null alleles were neutral and ignored the effects of linkage and gene conversion,

all of which made his results underestimates for the probability that an advantageous allele is fixed first. He found that for sufficiently large populations, gene fixation was the most likely fate for a newly duplicated gene. The contribution of Krakauer and Nowak (1999) was to propose different possible mechanisms by which redundant genes can be preserved. They cite asymmetric mutation, asymmetric efficacy, pleiotropy, developmental buffering, allelic competition and regulatory asymmetries. It is clear that they were simultaneously thinking on the same lines as Force et al. but although Krakauer and Nowak clearly stated that some form of asymmetry was necessary to maintain functional redundancy indefinitely, they did not formulated this as a single coherent model. Nevertheless their contribution to the debate was of great importance. Finally, I want to add the recent work by Kondrashov et al. (2002), who focused on the role of selection in the evolution of gene duplicates. These authors looked for evidence of increased evolution following duplication by analysing large datasets from different genomes and then measured the d_N/d_S rate ratio to establish selective pressures after gene duplication events. They found that paralogs evolve faster than orthologs with the same level of divergence and similar functions, but, they emphasize, these genes do not experience a phase of neutral evolution. So they claim that from their very time of origin, duplicate genes are advantageous and are thus retained. These genes can later develop new functions just as the classical model proposes, when a greater level of divergence has been reached.

1.5 The globin family of genes

Globins are perhaps the most thoroughly studied of all proteins and constitute the case study in this thesis. Since they were first identified, we have accumulated a huge

amount of information regarding their origin, evolution, function, regulation and structure. They have also been studied as models of gene family evolution by gene duplication, as globins have undergone all possible evolutionary routes available to newly duplicated genes. Here, I will provide a brief summary of the most important characteristics of globins, in general, and of β globins in particular.

1.5.1 The origin and taxonomic distribution of globins

Globins are part of a large group of proteins whose function is to bind oxygen non-covalently. Heme-binding proteins have been found in all kingdoms of living organisms, in prokaryotes, fungi, plants and animals (Hardison 1998). This widespread distribution indicates that the ancestral gene for globins must be ancient. In animals the most abundant globin protein is hemoglobin, which is found in erythrocytes in large concentrations and transports oxygen in the blood, from the lungs to tissues. In plants, hemoglobins were first found in the root nodules of legumes and were thus named leghemoglobins. In fungi, hemoglobins are also involved in oxygen transport and other functions. In *Saccharomyces* there has been a report of a fusion of a heme-binding domain and an FAD-binding domain. This fused flavohemoglobin shows no introns and is induced by high oxygen concentrations (Zhu and Riggs 1992). Intron positions in hemoglobins from different organisms offer clues as to the evolutionary history of globins that gave rise to the variety we know today. All plant hemoglobin genes are separated into four exons by three introns (Anderson et al. 1996). The first and third introns are in positions homologous to those of the two introns found in vertebrate hemoglobin and myoglobin genes. It was suggested that the ancestor to plant and animals had a hemoglobin gene with three introns, an arrangement that has been retained in plants and certain nematodes (Fig.

1.2). The central intron was lost before the divergence of annelids and arthropods, and is therefore absent in all vertebrate hemoglobin and myoglobin genes. Intron loss has varied in different organisms and there is an extreme example of total intron loss in the arthropod *Chironomus*. All models indicate that there was an ancestral hemoglobin gene around 1500 million years ago (MYA), before plants and animals diverged.

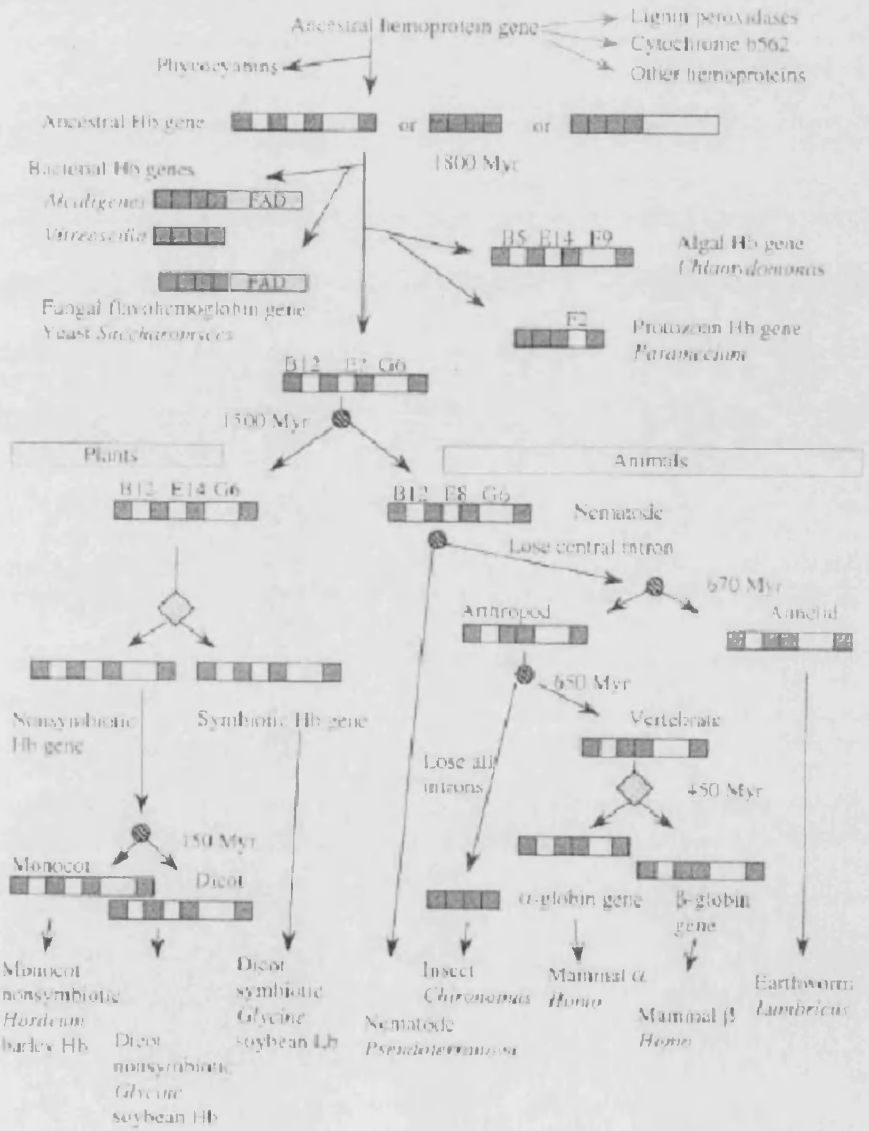


Fig.1.2 Hemoglobin gene evolution. Dark-filled boxes represent exons and genes, gray boxes represent a flavin-binding domain and introns are open boxes. Speciations are shown as circles and gene duplications as gray diamonds. Taken from Hardison 1998.

1.5.2 Major gene duplications in the globin superfamily of genes

In vertebrates, globins diverged from an ancestral gene and duplicated in successive occasions to give rise to the globin families we know today. Apart from hemoglobin (encoded by α and β globins), and myoglobin, there have been at least other two families reported in mammals, neuroglobin (Burmester 2000) and cytoglobin (Burmester et al. 2002). A different family, which conforms a separate clade in the globin superfamly tree, is that of truncated hemoglobins, with 20-40 fewer residues than vertebrate hemoglobins (Wittenberg et al. 2002). All these globin families are in most cases located in different chromosomes, as it occurs in humans with α hemoglobin in chromosome 16, β hemoglobin in chromosome 11 and myoglobin in chromosome 22 (Graur and Li 2000).

Myoglobin diverged from hemoglobin more than 800 MYA, preceding the emergence of annelid worms (Goodman 1976, for a review on myoglobin evolution see Romero-Herrera et al. 1978). In the great majority of vertebrates hemoglobin is composed of two types of protein chains whose encoding genes, α and β globin, diverged around 400 to 500 MYA probably by a tandem duplication (Goodman 1981). Initially, the tandem duplication generated two linked genes in the same chromosome, an arrangement that is retained in fish and amphibians. Chromosomal separation probably occurred around 300-350 MYA after the divergence of amphibians from amniotes but prior to the duplications that gave rise to the specific members of this family. The α globin cluster is composed of four functional genes: the embryonic gene ξ, two adult genes α_1 and α_2, θ_1. It also contains three unprocessed pseudogenes: ψε, ψα1, and ψα2. The β globin family has five functional genes: the embryonic ε globin; two fetal genes A_γ and G_γ; and two adult genes β and δ globins. There is also one unprocessed pseudogene, ψη (ψβ). Members of these gene

families vary in oxygen affinity. Fig. 1.3 shows the estimated dates of divergence among the members of these two hemoglobin gene families. In the case of α globin the most divergent is the ξ gene, which split more than 300 MYA, followed by the θ_1 gene, which branched off more than 260 MYA. Finally, the adult globin α genes and the ψα1 pseudogene diverged between 40-50 MYA. Among the β globins, proto ε diverged from proto β around 150-200 MYA (Efstratiadis et al. 1980, Czelusniak et al. 1982). Proto ε gave rise to ε and γ between 100 to 140 MYA, which in turn duplicated around 35 MYA in the simian primate lineage to give rise to A_γ and G_γ (Hayasaka et al. 1992). Also, proto β duplicated more than 80 MYA and originated β and δ, the adult globins (Hardison and Margot 1984, Goodman et al 1984).

1.5.3 The structure/function relationship in hemoglobin

Tetrameric vertebrate hemoglobin is composed of two identical α subunits and two identical β subunits. Each subunit contains a heme group. Oxygen binding is cooperative and is associated with a large shift in the quaternary structure of the heterotetramer: there are three known conformations, one is the relaxed or R conformation, which is adopted when hemoglobin is able to load oxygen; the opposite state is the tense or T conformation, which occurs when hemoglobin is ready to release oxygen to the tissues (Perutz 1970); and an intermediate conformation R2 has also been found and crystallized (Silva et al. 1992, Smith and Simmons 1994). Two types of interfaces participate in conformational transitions, namely α1β1 and α1β2. Perutz (1970) referred to the α1β1 interface as the packing surface and to the α1β2 interface as the sliding surface. During the transition form the T to R conformation, the α1β2 interface undergoes a large sliding movement, while the α1β1 interface is

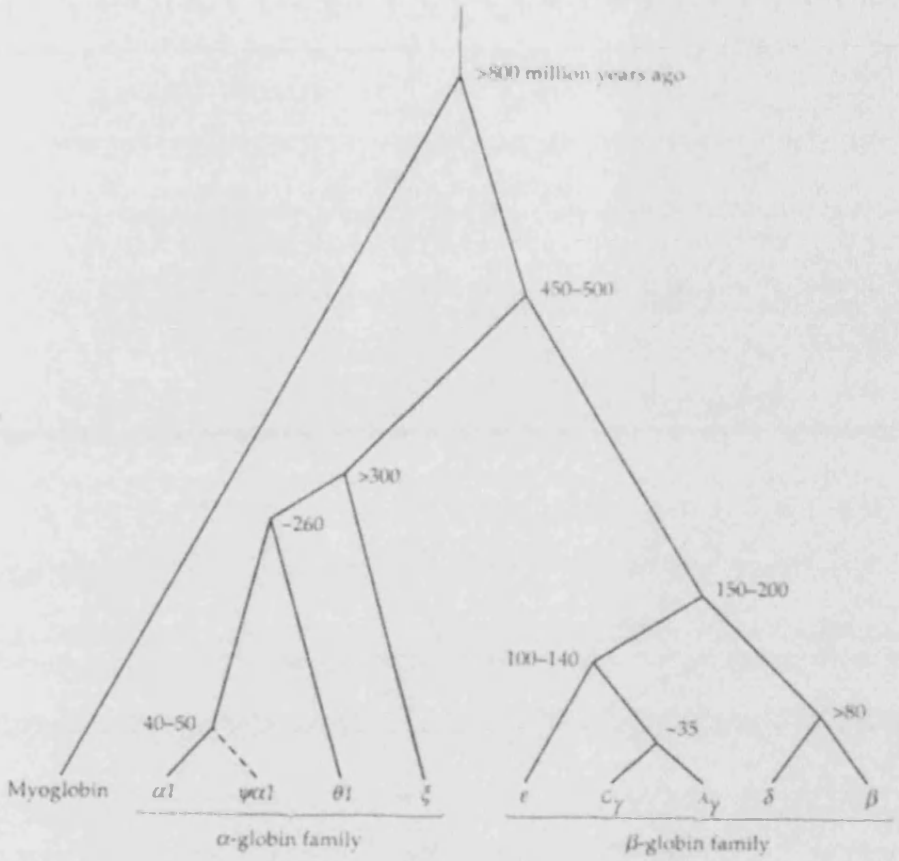


Fig. 1.3 Divergence date estimations for the major gene duplications in vertebrate globins. Taken from Li 1997.

practically unchanged. Hemoglobin properties such as oxygen affinity and cooperativity depend to a large extent on the intra-subunit and inter-subunit interfaces (Shionyu et al. 2001). There are monomeric and oligomeric hemoglobins; exactly what determines conformation into one or more monomeric units is not yet clear. From the physicochemical point of view, what determines the conformation of a protein into monomers or oligomers is the stability of the structure. It is known that the globin fold is a very stable and flexible structure, which has explored different structural and functional possibilities.

In terms of function, monomeric globins are not as sophisticated as oligomeric ones. Typically, monomers lack cooperative behavior whereas dimers and tetramers exhibit that property. Whereas myoglobin stores and delivers oxygen in tissue following a hyperbolic curve, hemoglobin oxygen intake and release is regulated in an allosteric way. The four hemes, one in each monomer of the hemoglobin, cooperate with one another and the affinity for oxygen is concentration dependent. At high oxygen concentrations, such as found in the lungs, hemoglobin has an increased oxygen affinity thus taking all the available oxygen. At low oxygen concentrations, such as found in metabolizing tissues, it has low oxygen affinity thus releasing all available oxygen to the tissues. Uptake and release of oxygen are also mediated by the conformation hemoglobin takes. The oxygen affinity is increased or lowered by as many as twenty six times between conformations, as part of the allosteric regulation of oxygen transport in hemoglobin. Chemical modulators such as bisphosphoglycerate, chloride ions, acids and carbon dioxide in turn regulate the transition rate between conformations. In metabolizing tissues, the acidic environment stabilizes the T conformation, whereas in the lungs, where blood is much less acidic it is the R conformation that is more stable.

1.5.4 Phylogenetic and physicochemical studies of amino acid sequence evolution in vertebrate globins

As Golding and Dean summarized it (1998), the field of molecular adaptation is divided between phylogenetics and physiological genetics or biochemistry. The former tend to look at pattern, while the latter see the process. The divide between history and mechanism is well exemplified here with this two opposing views on the evolution of the same molecule. Whereas Morris Goodman is more concerned with

the mode and tempo of evolution in globins and how molecular evolution is related to the evolution of species, Max Perutz is interested in how evolution adapts the globin fold to perform different functions in response to different environmental challenges.

In the tradition of phylogenetic studies, Goodman and his associates (1975) were the first to claim that early evolution of globins was much faster than later evolution (i.e. they rejected the molecular clock). They attributed the fast evolution to positive selection and related it to the improvement of function. Also, they proposed that the slow evolution was associated with stabilizing selection, which acted once improvements were fixed (Goodman et al. 1976). Based on the data provided by the crystallization of hemoglobin (achieved by Max Perutz), Goodman et al. showed that most of the mutations in globins occurred during the evolutionary transition from monomeric myoglobin to the allosteric tetramer at the sites involved in multimerization (1976). It made perfect sense to Goodman and his associates that our pre Cambrian ancestors took advantage of a monomeric hemoglobin with strong affinity for oxygen (i.e. showing a hyperbolic equilibrium curve), since they were worm sized creepy animals (Tiplady and Goodman 1977). As later organisms evolved, which were larger and more mobile, a new hemoglobin was needed, one that responded to the new demands like a better regulation of oxygen uptake and release (i.e. an allosteric hemoglobin with a sigmoid curve and Bohr effect), that could only be attained by a multimeric protein. Goodman thinks that such a change is more likely to be brought about by positive natural selection, which would also explain the acceleration of early evolution when there was considerable room for improvement. Opposite to this trend, once a stable structure had evolved, most mutations would be detrimental, thus slowing down the rate of amino acid changes.

In terms of the relationship between function and structure, Goodman showed there was a clear pattern in the amino acids that were more likely to be replaced (Goodman 1976). He and his colleagues found that amino acids forming part of the contact between subunits, either inter or intra subunits, were those that tended to have high rates of amino acid replacement in early evolution. Not by coincidence, these positions are related to the multimerization process. Also, the residues associated with heme contacts were stable and rarely changed, being established as far back as 600 to 700 MYA. However, the heme associated amino acids exhibit evolution seven times higher in the pre amniotes compared to the amniotes, which suggests that heme-heme cooperative interactions were perfected only around 300 MYA (Czelusniak et al. 1982). On the other hand, α and β hemoglobins evolved independently after their origin by gene duplication, and a similar pattern was observed as early evolution proved to be faster than at later stages (Goodman 1981).

From his studies on the stereochemical basis of variation in the allosteric properties of vertebrate hemoglobins, Max Perutz (1983, 1998) found that tertiary and quaternary structures of deoxy (i.e. unliganded hemoglobin) and oxyhemoglobin (i.e. ligand bound hemoglobin) have remained almost invariant among vertebrates and that most amino acid replacements are functionally neutral. He found evidence that adaptations responding to environmental stimuli have arisen by only a few (one to five) amino acid substitutions in key positions. His interest in the physicochemical basis of these adaptations led him to compare the amino acid changes that occurred among species and relate those with adaptive changes (Perutz 1983). One of the most spectacular examples he found was that of oxygen transport and storage regulation in crocodilians. These animals are able to stay underwater for as long as an hour without coming up to breathe. To do this, they reduce oxygen consumption by shutting off the

circulation to their muscles, so that oxygen supply is concentrated in the brain and viscera. They use as much as possible of the oxygen stored in their lungs and blood. This is accomplished because their hemoglobin has unusual allosteric properties. In most vertebrates, hemoglobin is regulated by chemical effectors that compete with oxygen for binding the heme in hemoglobin. These chemical effectors regulate the affinity of hemoglobin for oxygen and make possible interesting effects such as the Bohr effect (i.e. hemoglobin releases oxygen from the heme atoms more readily at reduced pH). Effector molecules typically include organic phosphates, H^+, Cl^-, and CO_2. Temperature also affects oxygen affinity of the heme. In crocodilians, hemoglobin does not respond to the typical effector molecules, in contrast it responds strongly only to bicarbonate ion. Bicarbonate ion forms when carbon dioxide dissolves in water, it accumulates in the crocodile's blood when it is underwater and binds to the regulatory site which lowers the hemoglobin affinity for oxygen. So, instead of being retained by the heme, oxygen is delivered to the brain and viscera. This remarkable property, Perutz emphasized, is brought about by only three amino acid replacements, Val NA1β to Ser, His NA2β to Pro, and Lys HC1β to Glu. These replacements involve only four nucleotide base changes. Furthermore, all other substitutions are neutral (Perutz et al. 1981). The fact that globins have been used as examples by both selectionists and neutralists attest to the complexity of their evolution.

Chapter II

-

Gene conversion and functional divergence in the β globin gene family

2.1 Motivation

The globin gene family is a textbook example of evolution by gene duplication, as the paralogs that arose by this process have undergone all the evolutionary routes available to newly duplicated genes (Li 1997), for example: α and β genes retained their original function (i.e., encode the adult hemoglobin chains) (Bunn 1981), ψ globins, η and δ became nonfunctional in some lineages (Lacy and Maniatis 1980; Cleary et al. 1981; Li, Gojobori and Nei 1981; Martin et al. 1983; Goodman et al. 1984) and yet others like γ and ε changed their function and time of expression (Farace et al. 1984; Hutchinson et al. 1984; Fitch et al. 1991; Meireles et al. 1995; Johnson et al. 1996). The eutherian mammal ß globin family comprises 5 functional genes (ß, δ, ε, G_γ and A_γ globin) and one pseudogene (ψß) typically arranged in a specific linkage order (Fig. 2.1). Even though globins are one of the best studied proteins, I believe it is necessary to update and complement the information generated mostly during the 1980s regarding the evolution of this gene family by gene duplication and to reanalyse data using new and more powerful methods.

In this chapter I want to emphasize the complex ways in which different evolutionary forces (indicated by selective pressure) have operated to give rise to the functional divergence observed in the β globin gene paralogs, which have experienced frequent unequal crossing-over (gene conversion), episodes of positive selection, purifying selection, long-term differences in selective pressure among genes, and recurrent birth and death of some members in the cluster.

Traditionally, this gene family was considered an example in support of Ohta's classical model of evolution by gene duplication (Goodman 1981, Czelusniak et al. 1982, Ohta 1990). I am very interested in testing this model using new methods,

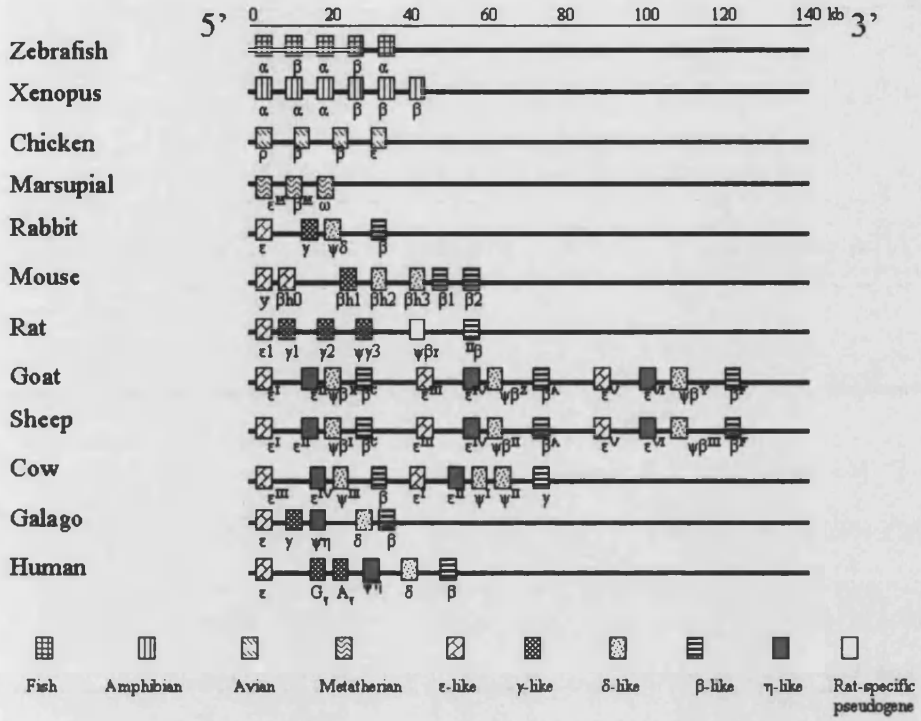


Fig. 2.1. ß globin gene linkage in different vertebrates (Cooper et al. 1996; Garner and Lingrel 1989; Konkel et al. 1979; Lacy et al. 1979; Kretschmer et al. 1981; Lingrel et al. 1983; Satoh et al. 1999; Schon et al. 1981; Shapiro et al. 1983; Townes et al. 1984; Schimenti and Duncan 1985b). Orientation is variable in fish and amphibian clusters (Gillemans et al. 2002; Hosbach et al. 1983).

and in particular, in comparing the predictions of the classical model with a very different alternative model, the DDC (Force et al. 1999). In a way, these two models represent extreme scenarios of evolution by gene duplication.

I assembled a dataset of 72 DNA sequences that include mammals, amphibians, fish, and birds. I inferred a phylogeny for the ß globin family and identified duplication events and gene conversions, some of which are reported for the first time. Specifically I tested for (*i*) a significant increase in the rate of nonsynonymous substitution following gene duplication events, a consequence of neofunctionalization predicted by Ohta (1988); and (*ii*) significant differences in selective constraints among paralogs. Even though the DDC model is concerned with the evolution of regulatory regions I hypothesize that, if subfunctionalization occurs in the protein coding sequences, as well as in the regulatory sequences, selective pressure should differ between paralogs. I measured selective pressure by using the nonsynonymous/synonymous substitution rate ratio (ω), as implemented in codon models of sequence evolution (Nielsen and Yang 1998, Yang et al 2000). An $\omega < 1$ indicates purifying selection, $\omega = 1$ is consistent with neutral evolution, and $\omega > 1$ indicates positive Darwinian selection (Yang and Bielawski 2000).

2.2 Theory and Methods

2.2.1 Phylogeny inference

A phylogenetic approach is assumed in this thesis to study the molecular evolution of the β globin gene family. One of the advantages of using this kind of approach is that questions can be put in a specific temporal and taxonomic context. For the phylogenetic study of the ß globin gene family, 72 sequences from various vertebrates including fish, amphibians, birds and mammals were obtained from GenBank. The

nomenclature of ß globin genes is rather chaotic. To avoid confusion, I have included species names and GenBank accession numbers next to each sequence in Figure 2.2. I used the bony fish clade as outgroup. The sampled sequences were aligned using Clustal X (Thompson et al. 1997), followed by manual adjustments. Alignment gaps were removed.

Using different methods I constructed a number of phylogenies for the ß globin genes to understand the relative order of duplication and speciation events, and to identify gene conversions. Trees were estimated from the nucleotide sequences using maximum parsimony, maximum likelihood and Bayesian analysis. Once phylogenies were estimated, I assessed their robustness by using different methods. Relative support for internal branches was measured by using Bootstrap analysis with PAUP* (Swofford 1998). I performed the SH (Shimodaira & Hasegawa 1999), KH (Kishino & Hasegawa 1989) and RELL (Kishino & Hasegawa 1989) tests to compare the inferred gene tree with an alternative topology derived from the expected species relationships. I compared two trees each time, the tree in Figure 2.2 and a tree modified by relocating the misplaced taxa according to the species phylogeny.

2.2.2 Tests of Gene Conversion

In general, gene conversion is inferred to have occurred when two genes, or regions in two genes, have synonymous sites that are more similar for these genes than for other genes (Drouin et al. 1999). I used the programs: (*i*) PLATO (Grassly and Holmes 1997) that detects anomalies that arise when phylogenies of different parts of the genome result in discordant topologies ; (*ii*) Pist (Worobey 2001) that examines whether the nucleotide substitutions observed between a set of genes are randomly distributed along these sequences or not ; (*iii*) GENECONV (Sawyer 1999) that looks

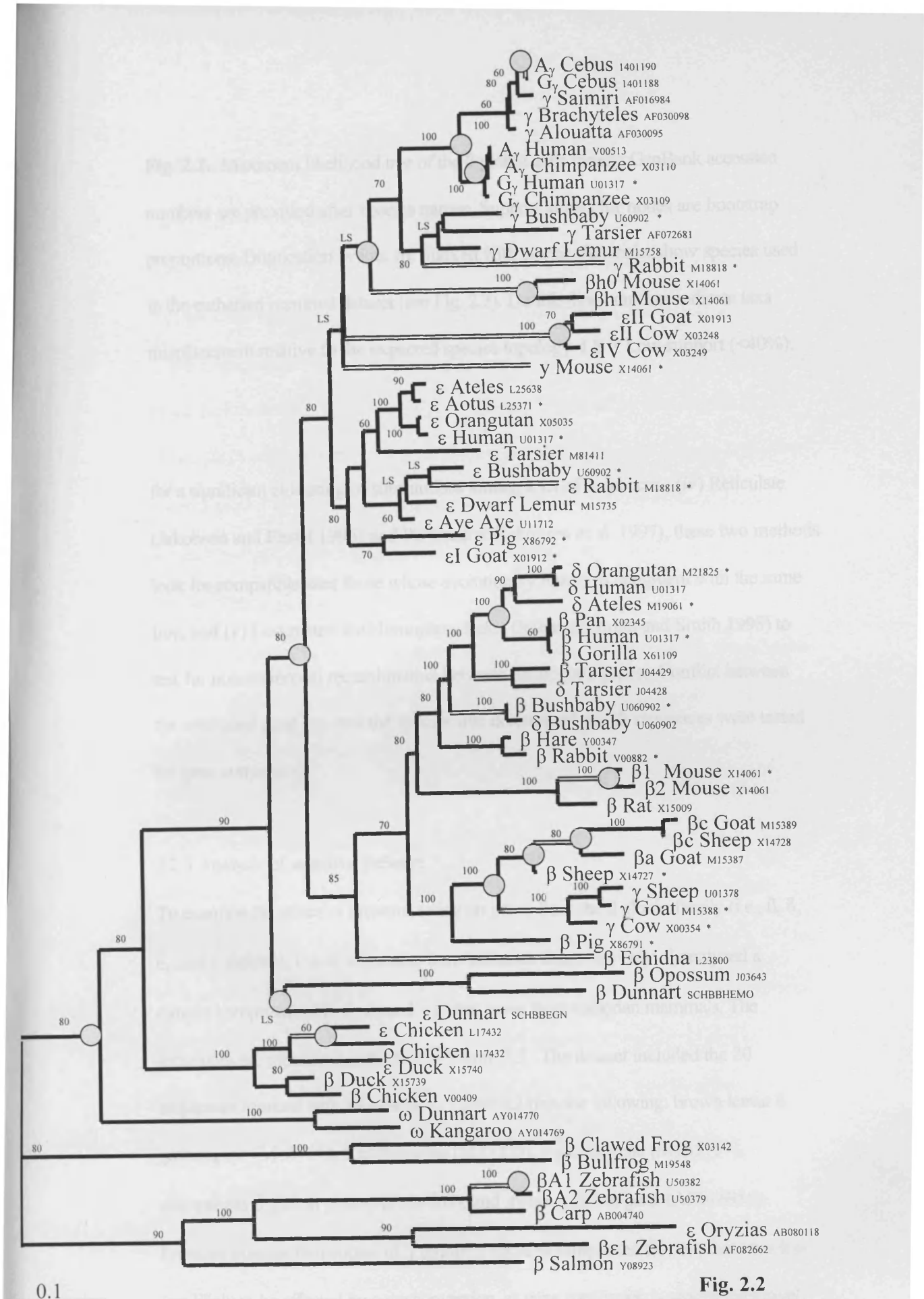

Aγ Cebus 1401190
Gγ Cebus 1401188
γ Saimiri AF016984
γ Brachyteles AF030098
γ Alouatta AF030095
Aγ Human V00513
Aγ Chimpanzee X03110
Gγ Human U01317 *
Gγ Chimpanzee X03109
γ Bushbaby U60902 *
γ Tarsier AF072681
γ Dwarf Lemur M15758
γ Rabbit M18818 *
βh0 Mouse X14061
βh1 Mouse X14061
εII Goat X01913
εII Cow X03248
εIV Cow X03249
γ Mouse X14061 *
ε Ateles L25638
ε Aotus L25371 *
ε Orangutan X05035
ε Human U01317 *
ε Tarsier M81411
ε Bushbaby U60902 *
ε Rabbit M18818 *
ε Dwarf Lemur M15735
ε Aye Aye U11712
ε Pig X86792 *
εI Goat X01912 *
δ Orangutan M21825 *
δ Human U01317
δ Ateles M19061 *
β Pan X02345
β Human U01317 *
β Gorilla X61109
β Tarsier J04429
δ Tarsier J04428
β Bushbaby U060902 *
δ Bushbaby U060902
β Hare Y00347
β Rabbit V00882 *
β1 Mouse X14061 *
β2 Mouse X14061
β Rat X15009
βc Goat M15389
βc Sheep X14728
βa Goat M15387
β Sheep X14727 *
γ Sheep U01378
γ Goat M15388 *
γ Cow X00354 *
β Pig X86791 *
β Echidna L23800
β Opossum J03643
β Dunnart SCHBBHEMO
ε Dunnart SCHBBEGN
ε Chicken L17432
ρ Chicken I17432
ε Duck X15740
β Duck X15739
β Chicken V00409
ω Dunnart AY014770
ω Kangaroo AY014769
β Clawed Frog X03142
β Bullfrog M19548
βA1 Zebrafish U50382
βA2 Zebrafish U50379
β Carp AB004740
ε Oryzias AB080118
βε1 Zebrafish AF082662
β Salmon Y08923
0.1

Fig. 2.2

Fig. 2.2. Maximum likelihood tree of the ß globin gene family. GenBank accession numbers are provided after species names. Support values for nodes are bootstrap proportions. Duplication events are marked with circles. Asterisks show species used in the eutherian mammal dataset (see Fig. 2.3). Double-line branches indicate taxa misplacement relative to the expected species topology. LS = Low support (<40%).

for a significant clustering of substitutions among a set of sequences ; (*iv*) Reticulate (Jakobsen and Eastel 1996) and Partimatrix (Jakobsen et al. 1997), these two methods look for compatible sites those whose evolutionary history is congruent with the same tree; and (*v*) I estimated the Homoplasy Index (Maynard Smith and Smith 1998) to test for non-reciprocal recombination between paralogous genes. Conflict between the estimated gene tree and the species tree determined which sequences were tested for gene conversion.

2.2.3 Analysis of selective pressure

To examine the selective pressure acting on genes from the ß globin family (i.e., ß, δ, ε, and γ globins), I used sequences from eutherian mammals only. I analyzed a dataset comprised of ß, δ, ε, and γ globin genes from eutherian mammals. The sequences are identified in the tree of Figure 2.3. The dataset included the 20 sequences marked with an asterisk in Figure 2.2 plus the following: brown lemur ß globin gene (M15734), ε globin gene (M15735), γ globin gene (M155757), chimpanzee δ globin gene (AF339363), and *Aotus* γ globin gene (AF016985). Primates possess two copies of γ globin; I chose to sample the G_γ copy because it is less likely to be affected by gene conversion, as gene conversion is almost exclusively

unidirectional with G_γ converting A_γ (Fitch et al. 1990). β globin genes converted δ globin genes in some lineages (Koop et al. 1989); therefore I excluded the converted δ copies. Also excluded were some of the internally duplicated genes in the ruminant ß globin cluster (goat εIII, εIV, εV, and εVI, and the cow εI and εIII) (Fig. 2.1). These sequences are very divergent due to inserted sequences (Saban and King 1994). From the mouse ß globin cluster (Fig. 2.1), I sampled one of the three copies of fetal globin (ßh1) and one of the two adult globin genes (ß1). Separate datasets also were constructed for ß, ε, and γ globin genes. There were too few sequences available for a separate analysis of the δ globin gene, where indicated these sequences were analyzed together with β globin genes.

Site-based analyses— A statistical approach was taken to study the selective pressure on the ß globin gene family in eutherian mammals. I used several codon models of molecular evolution that allow for heterogeneous d_N/d_S ratios at sites (Nielsen and Yang 1998; Yang et al. 2000). In the simplest model (M0 or one-ratio model), the ω ratio is an average over all the sites. The "neutral" model (M1) allows for conserved sites where $\omega = 0$ and completely neutral sites where $\omega = 1$. The "selection" model (M2) adds a third class to M1 at which ω can take values > 1. The discrete model (M3) uses an unconstrained discrete distribution with different ω ratios for K different classes of sites. Model M7 (beta) assumes a beta distribution of ω over sites. Model M8 (beta&ω) adds an extra class of sites to M7, thereby allowing ω values >1. Likelihood ratio tests (LRTs) were conducted to test M0 (one-ratio) against M3, M1 (neutral) against M2 (selection), and M7 (beta) against M8 (beta & ω). All analyses were based on the unrooted gene-tree topologies, and used the codeml program in the PAML package (Yang 1997).

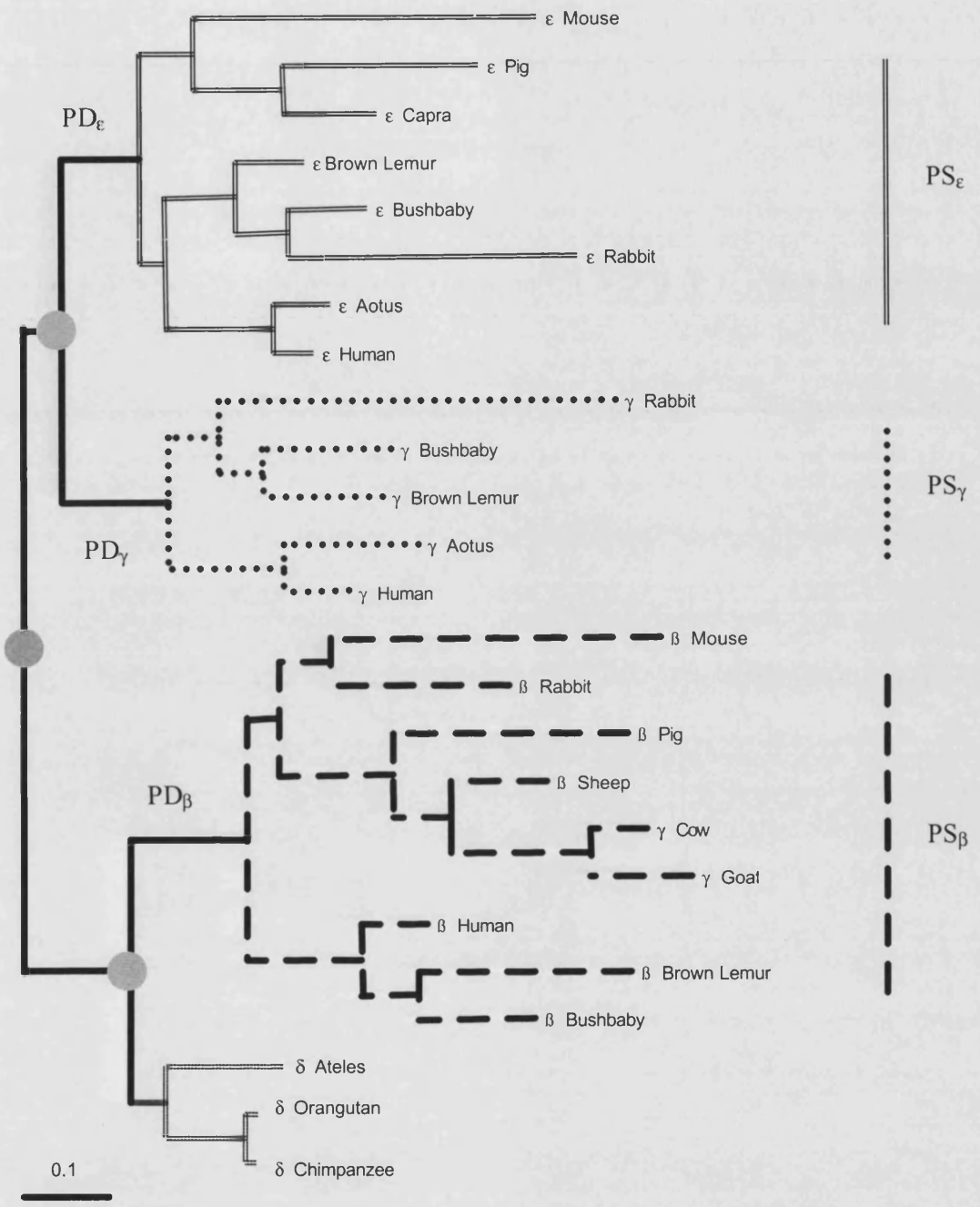


Fig. 2.3. Maximum likelihood tree of the ß, δ, ε, and γ globin genes from eutherian mammals. Grey circles indicate gene duplication events. Branches in the tree are partitioned into post-duplication (PD; those that immediately post-date the grey circles), and post-speciation (PS; those that postdate species-divergences) branches. The tree is rooted at the proto-epsilon and proto-beta globin duplication event to help interpret classification of branches. All analyses were conducted by using unrooted trees. Labels indicate the classification of PD (—) and PS branches (===; - - - ; = = =; and ····). This is a general representation, different tests focused on different PD and PS branches according to each case described in the text.

Branch-based analyses— To study changes in selective pressure in the context of gene duplication I implemented several models that allow for variable ω ratios among branches in the tree (Yang 1998; Bielawski and Yang 2003). The null model assumed the same ω for all lineages in the tree. The "PD-PS" model assigns different ω ratios for post-speciation and post-duplication branches in the tree (e.g. Fig. 2.3). This is based on the hypothesis that duplicated genes avoid nonfunctionalization because positive Darwinian selection promoted fixation of amino acid mutations that led to a new or modified gene function (Ohta 1988a). The hypothesis predicts a burst of amino acid replacements in the branches post-dating duplication events (Ohta 1983). After a new function evolves, however, amino acid evolution is expected to be dominated by purifying selection and the rate of nonsynonymous substitution should decrease (Ohta 1993). Hence there should be a higher rate of amino acid substitution along branches that immediately postdate duplication events (PD branches) as compared with those branches that immediately postdate speciation events (PS branches). An LRT can be conducted to compare the one-ratio model ($\omega_{PD} = \omega_{PS}$) with the two-ratio model PD-PS ($\omega_{PD} \neq \omega_{PS}$).

Another alternative model was based on the hypothesis that duplicated genes avoid nonfunctionalization because expression patterns and/or functions are partitioned among paralogs following gene duplication (Force et al. 1999). If subfunctionalization had indeed occurred in the protein-coding sequences, sites associated with such partitioning are expected to exhibit long-term differences in selection pressure. If the difference between paralogs is large, we might be able to detect paralog-specific differences in average selective constraint. I formalized this in a model called "Paralog", where an independent ω ratio is specified for each paralogous clade (e.g., $\omega_\beta \neq \omega_\gamma \neq \omega_\varepsilon$). To test for a significant difference in selective

pressure among paralogs I conducted an LRT comparing the one-ratio model (e.g., $\omega_\beta = \omega_\gamma = \omega_\varepsilon$) with the three-ratio Paralog model.

Branch-Site Analysis-The above approaches might not detect a short episode of positive Darwinian selection, such as immediately following a gene duplication event, if it occurs at just a fraction of amino acid sites. The "branch-site" models (models A and B) recently developed allow the ω ratio to vary both among lineages and among sites, and permits detection of lineage-specific changes in selective pressure at specific amino acid sites (Yang and Nielsen 2002). Branch-site models A and B have four ω site classes. The first two site classes, with ω_0 and ω_1, are uniform across the phylogeny, whereas the other two site classes are allowed to change from $\omega_0 \rightarrow \omega_2$ and from $\omega_1 \rightarrow \omega_2$ in a pre-specified branch of interest (the "foreground" branch). Note that ω_2 can take values > 1, thus allowing for positive selection. In branch-site model A, ω_0 is fixed to 0 and ω_1 is fixed to 1; hence positive selection is permitted at only the foreground branch. Model A is compared with model M1 (neutral) with degrees of freedom (d.f.) = 2. In model B, ω_0 and ω_1 are free parameters; therefore some sites can evolve under positive selection across all the branches in the phylogeny, whereas other sites are permitted to take ω values > 1 in the foreground branch. An LRT compares model B with model M3 (discrete) with K = 2 site classes and d.f. = 2. We used branch-site models A and B to test for possible adaptive evolution along lineages following gene duplications.

2.3 Results

2.3.1 Phylogenetic Analysis

The 72 ß globin family genes collected were used for phylogenetic reconstruction. The ML tree is shown in Figure 2.2. Both ML and Bayesian methods resulted in similar topologies, with support values for the internal nodes shown in Figure 2.2. The only case of disagreement between the two methods was in the placement of marsupial and monotreme sequences. In the Bayesian tree the echidna ß globin gene was sister to a marsupial clade (opossum and dunnart ß globins) and in turn this clade was placed sister to the eutherian ß globin clade. In the ML tree (Fig. 2.2), the echidna ß-globin gene was sister to the eutherian ß globin clade. Clearly, placement of the monotreme and marsupial ß globins is problematic and will probably require additional sampling to resolve. Interestingly, the marsupial ω globin genes were placed outside the mammalian ß globin clade, consistent with the earlier study of Wheeler et al. (2001).

Assuming no gene conversion, I expected (i) monophyly for each set of paralogs (i.e., ß, δ, ε, and γ globins), and (ii) to recover the expected species tree within each paralogous clade (Rowe 1999; O'Brien et al. 2001; Springer et al. 2003). However, I found some notable misplacements (double lines in Fig. 2.2): (*i*) the rabbit ε and γ sequences were sister to the primate ε and γ genes, respectively, rather than sister to rodent ε and γ globins; (*ii*) the cow ε II and ε IV genes and goat ε II comprised a monophyletic clade sister to the γ globins, instead of being within the ε clade; (*iii*) the mouse ε gene (a single-copy gene traditionally called y), did not appear within the ε clade but was sister to a clade including the cow ε II and ε IV and the goat ε II genes; (*iv*) tarsier and bushbaby δ globin genes were sister to tarsier and

bushbaby ß globin genes, respectively; (*v*) the genes traditionally labeled as γ globins in sheep, cow and goat were placed within the ß globin clade; (*vi*) chicken ε was sister to chicken ρ instead of being more closely related to duck ε globin; (*vii*) *Cebus* Gγ and Aγ were more closely related to each other than to their respective human and chimpanzee orthologs .

All misplacements were supported by high bootstrap proportions (> 70%) with the exception of the *Cebus* A_γ and G_γ, the rabbit ε, and the mouse y branches where there was low bootstrap support. I used the SH test to compare the expected placements with the estimated topology (Fig. 2.2). SH tests indicated significantly greater support for five misplacements (bushbaby δ: $P < 0.0001$; tarsier δ: $P = 0.002$; cow, sheep and goat γ: $P = 0.000$; echidna ß: $P = 0.053$ and *Cebus* A_γ and G_γ: $P = 0.000$). The remaining misplacements did not fit these data significantly better than the expected phylogenetic placements (mouse ε: $P = 0.095$; rabbit γ: $P = 0.217$; rabbit ß: $P = 0.59$; cow ε: $P = 0.193$; rabbit ε: $P = 0.289$ and mouse y, goat εII and cow εII and εIV: $P = 0.225$, chicken ρ: $P = 0.397$). Results under KH and RELL tests were the same as with SH tests (data not shown).

2.3.2 Extent of Gene Conversion

A potential source of conflict between the gene tree and species tree could be gene conversion (Ohta 1980, 1990). Hence, I used the misplacements to guide my tests of gene conversion. Tests were conducted on alignments of third codon positions only, by using different software programs (Grassly and Holmes 1997; Worobey 2001; Sawyer 1999; Jakobsen and Eastel 1996; Jakobsen et al. 1997; Maynard Smith and Smith 1998). I found evidence for two gene conversion events

that are not reported previously: (i) among the duplicates in the goat ß globin cluster between nucleotides 12 and 75 (site numbering refers to the human ß-globin gene, PDB file 2hhb) (PLATO z-score = 4.85); and (ii) among the mouse ß globin cluster genes between nucleotides 210 and 235 (PLATO z-score = 3.87). The analysis corroborated gene conversions previously suggested for tarsier and bushbaby δ globin genes (Koop et al. 1989; Grassly and Holmes 1997) between nucleotides 45-63 and 357-375, and in cow εII and cow εIV between nucleotides 12-30, in agreement with Schimenti and Duncan (1985a). However, I found no evidence for gene conversions between mouse ß genes ßh0 and ßh1 or between mouse ßh0 and mouse y, (see Figs. 2.1 and 2.2) reported by Hill et al. (1984), nor between *Cebus* Gγ and Aγ.

2.3.3 Analysis of selective pressure

Variable Selective Pressure Among Sites. - In order to minimize the effect of gene conversion, I excluded the converted genes. Given that gene conversion tends to have a direction in globins, it is known for instance that δ globins are generally converted by β globins and not vice versa. This prior knowledge allowed me to minimize gene conversion effects to some extent, although eliminating conversion altogether is impossible, as numerous events have characterized the evolution of β globin genes. I also compared tests of variable selective pressure using different datasets, both with and without misplaced sequences. Similar results were obtained for the different datasets, confirming that gene conversion, although probably present, did not greatly affect my results.

I expected selective pressure to vary among sites and among the genes of the ß globin family. I used codon models to detect among-site variability in selective pressure in the ß, ε and γ globin genes. From the one-ratio model (M0) I found that

the ω ratio averaged over all sites is 0.27, 0.26, and 0.17 for ß, γ and ε globin genes, respectively, when the three genes were analyzed as separate data sets (Table 2.1). The estimates suggested that, on average, the ε globin is more constrained than the γ and ß. However, an ω ratio averaged over sites is a crude measure of selective pressure. Therefore I used models that allow selective pressure to vary among sites. The discrete model (M3), with 3 site classes, revealed considerable variation in selective pressure among sites (Table 2.1). For example, ß globin had 65% of sites under strong purifying selection (ω = 0.02), 26% of sites were less constrained (ω = 0.57), and 9 % of sites were under positive selection (ω = 2.02) (Table 2.1). Interestingly, neither γ nor ε showed evidence of sites evolving under positive selection (Table 2.1). Evolution of the majority of sites in all three paralogs was dominated by strong purifying selection, with 65% of sites in ß, 52 % of sites in γ, and 66% of sites in ε evolving with $\omega < 0.05$.

I tested for variable selective pressure among sites by conducting an LRT comparing the one-ratio model (M0) with the discrete model (M3); results were highly significant for all three genes (Table 2.2). In general, ß globin was the most variable gene in the family, having an additional class of sites evolving under positive Darwinian selection.

I was interested in identifying regions that are conserved in all three genes in the cluster, which presumably indicate functionally important residues in the protein product. For ß-, γ- and ε-globin genes separately, I plotted the approximate posterior mean of the ω ratio at each site (Fig. 2.4). Four regions are highly conserved in all three genes: (i) residues 28 to 38, located in helices B and C; (ii) residues 57 to 63, located in helix E; (iii) residues 79 to 81, located in helix F; and (iv) residues 87 to

Table 2.1. Parameter estimates and likelihood scores in separate analyses of the ß, γ, and ε globin genes under site-specific models

Model	Parameter Estimates	ℓ
M0 (one-ratio)		
ß	$\omega = 0.27$	-1676.08
γ	$\omega = 0.26$	-1609.76
ε	$\omega = 0.17$	-2137.83
M1 (neutral)		
ß	$(\omega_0 = 0)$, $f_0 = 0.60$, $(\omega_1 = 1)$, $(f_1 = 0.40)$	-1621.00
γ	$(\omega_0 = 0)$, $f_0 = 0.57$, $(\omega_1 = 1)$, $(f_1 = 0.43)$	-1598.06
ε	$(\omega_0 = 0)$, $f_0 = 0.54$, $(\omega_1 = 1)$, $(f_1 = 0.46)$	-2145.53
M2 selection		
ß	$(\omega_0 = 0)$, $f_0 = 0.60$, $(\omega_1 = 1)$, $f_1 = 0.36$, $\omega_2 = 3.58$, $(f_2 = 0.04)$	-1617.42
γ	$(\omega_0 = 0)$, $f_0 = 0.52$, $(\omega_1 = 1)$, $f_1 = 0.006$, $\omega_2 = 0.57$, $(f_2 = 0.47)$	-1592.80
ε	$(\omega_0 = 0)$, $f_0 = 0.33$, $(\omega_1 = 1)$, $f_1 = 0.11$, $\omega_2 = 0.16$, $(f_2 = 0.56)$	-2100.22
M3 discrete		
ß	$\omega_0 = 0.02$, $f_0 = 0.65$ $\omega_1 = 0.57$, $f_1 = 0.26$, $\omega_2 = 2.02$, $(f_2 = 0.09)$	-1608.57
γ	$\omega_0 = 0.001$, $f_0 = 0.52$ $\omega_1 = 0.42$, $f_1 = 0.18$, $\omega_2 = 0.66$, $(f_2 = 0.31)$	-1592.78
ε	$\omega_0 = 0.04$, $f_0 = 0.66$ $\omega_1 = 0.27$, $f_1 = 0.24$, $\omega_2 = 0.89$, $(f_2 = 0.11)$	-2099.60
M7 beta		
ß	$p = 0.11$, $q = 0.29$	-1612.60
γ	$p = 0.23$, $q = 0.55$	-1593.31
ε	$p = 0.34$, $q = 1.50$	-2101.40
M8 beta&ω		
ß	$p = 0.16$, $q = 0.061$, $f_0 = 0.93$ $\omega_1 = 2.19$, $(f_1 = 0.07)$	-1608.76
γ	$p = 0.03$, $q = 0.64$, $f_0 = 0.57$ $\omega_1 = 0.60$, $(f_1 = 0.43)$	-1592.79
ε	$p = 0.95$, $q = 7.81$, $f_0 = 0.88$ $\omega_1 = 0.85$, $(f_1 = 0.12)$	-2099.67

Table 2.2. Likelihood ratio test statistics for comparing site-specific models for the ß, γ, and ε globin genes

Model	2δ	df	*P*-value
M0 (one-ratio) vs. M3 (discrete)			
ß	135.02	2	<0.0001
γ	33.97	2	<0.0001
ε	76.47	2	<0.0001
M7 (beta) vs. M8 (beta&ω)			
ß	7.68	2	0.020
γ	1.03	2	0.600
ε	3.46	2	0.177

101, located in helices F and G. When mapped onto the three-dimensional structure of the ß chain in hemoglobin (Fig 2.5), I found that sites within these four constrained regions were located mostly on the inner hydrophobic core of the subunit, the area around the heme pocket and the $\alpha_1\beta_1$ interface. In all cases the human ß globin chain structure (PDB: 2hhb) was used as reference to map sites into the three-dimensional structure. Residues 28 to 38, are distributed among the hydrophobic core, the $\alpha_1\beta_1$ interface between monomers, and part of the heme pocket.

The site-specific codon models were also used to identify positive selection at sites, indicated by $\omega > 1$. The selection model (M2), the discrete model (M3), and the beta&ω model (M8) allow $\omega > 1$ at a fraction of sites (Yang et al. 2000). All three models were generally consistent in suggesting a small fraction of sites (4 to 9 %) evolving under positive Darwinian selection (ω between 2.02 and 3.58) in the ß globin gene (Table 1). I tested significance of sites evolving under positive selection by an LRT comparing M7, which does not allow for such sites, with M8, which has an additional parameter that can accommodate sites with $\omega > 1$. The test is highly significant for the ß globin gene (Table 2.2).

Variable Selective Pressures Among Branches. - A burst of nonsynonymous evolution is often observed following gene duplication, and positive Darwinian selection is frequently invoked to explain this pattern. An LRT was used to test whether selective pressure is significantly different between postduplication (PD) and postspeciation (PS) branches in the ß globin gene phylogeny; i.e., $(\omega_{\beta(PD)} = \omega_{\varepsilon(PD)} = \omega_{\gamma(PD)}) \neq (\omega_{\beta(PS)} = \omega_{\varepsilon(PS)} = \omega_{\gamma(PS)})$. The LRT was not significant (Table 2.3), suggesting no difference

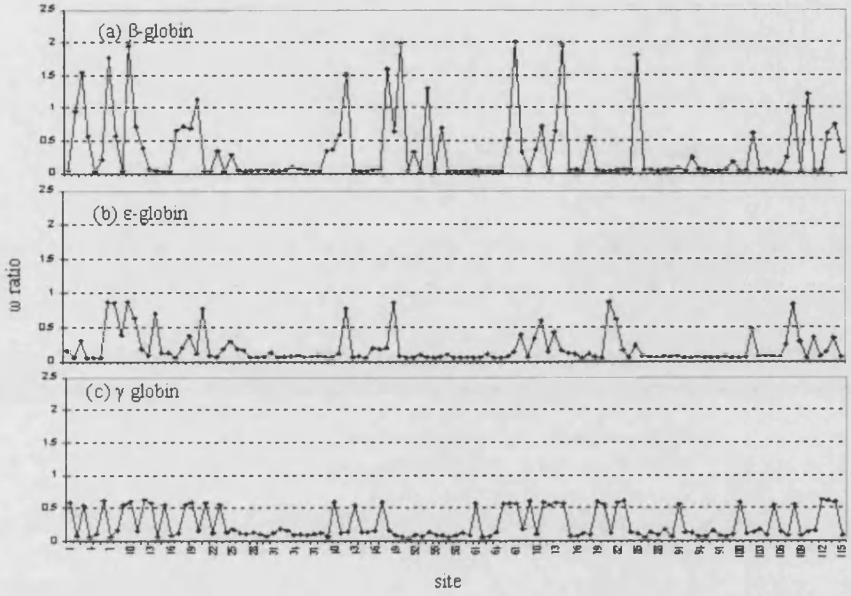


Fig. 2.4. Approximate posterior mean of the ω ratio for each site calculated under model M3 (discrete) for the a) ß globin; b) ε globin and c) γ globin genes.

between PD branches and PS branches. Furthermore, estimates of ω suggested strong purifying selection in both the PD and PS branches ($\omega_{(PD)} = 0.34$, $\omega_{(PS)} = 0.23$). I also fitted a more general four-ratios model in which the branches postdating the three duplication events in the phylogeny were assigned independent ω ratios ($\omega_{\beta(PD)}$, $\omega_{\varepsilon(PD)}$, $\omega_{\gamma(PD)}$, $\omega_{(PS)}$) and compared it with the one ratio model. Again, the LRT was not significant (Table 2.3), and none of the parameter estimates suggested positive Darwinian selection: $\omega_{\beta(PD)} = 0.41$, $\omega_{\varepsilon(PD)} = 0.22$, $\omega_{\gamma(PD)} = 0.08$, $\omega_{(PS)} = 0.24$. Note that in both PD-PS models tested, d_N values averaged 0.024 and d_S values averaged 0.101.

The above analysis averages rates over all sites in the gene and may lack power in detecting positive selection. Thus I also used branch-site models A and B (Yang and Nielsen 2002) to detect positive selection at a subset of sites along specific lineages. I tested each postduplication branch in the ß globin phylogeny as defined in Fig. 2.3. I found no evidence for positive selection at branches immediately following the duplication event that gave rise to proto-ß and proto-ε, nor after the duplication that created ε and γ (data not shown). The duplication event that resulted in A_γ and G_γ globins is hypothesized to have occurred along the branch leading to the simian primates (Slightom et al. 1985), but cannot be resolved on a gene tree because of frequent gene conversion events. However, when I used a specific dataset comprising ε and γ globins (Fig. 2.6), and tested the branch where the duplication is thought to have occurred I found an increase in nonsynonymous substitutions (M1 vs MA: $2\delta = 37.16$, df = 2, $P < 0.0001$; M3 vs MB: $2\delta = 18.66$, df = 2, $P < 0.0001$). The d_N value was 0.021 and the d_S value was 0.039, as measured as an average over all branches of the ε and γ globin tree. Parameter estimates under models A and B suggested

Table 2.3. Maximum likelihood estimates of ω ratios under branch-specific models and likelihood ratio test statistics when the model is compared with the null model M0 (one-ratio)

Alternative model	Parameter estimates	2δ	df	*P*-value
Post-duplication and post-speciation (PD-PS) models				
2-ratios	$\omega_{(PD)}$ =0.34, $\omega_{(PS)}$ = 0.23	1.10	1	0.29
4-ratios	$\omega_{\beta(PD)}$ = 0.41, $\omega_{\varepsilon(PD)}$ = 0.22, $\omega_{\gamma(PD))}$ =0.08, $\omega_{(PS)}$ = 0.24	1.10	3	0.78
Paralog models				
3-ratios	ω_{β} = 0.29, ω_{γ} = 0.23, ω_{ε} = 0.16	11.66	2	0.003
2-ratios	ω_{β} = 0.28, ω_{ε} = ω_{γ} = 0.19	7.89	1	0.005
2-ratios	ω_{ε} = 0.16, ω_{γ} = ω_{β} = 0.27	9.88	1	0.002
2-ratios	ω_{γ} = 0.23, ω_{β} = ω_{ε} = 0.23	0.045	1	0.832

positive selection at a few sites along the branch leading to simian primates ($\omega_{2(MA)}$ = 10.0, $\omega_{2(MB)}$ = 4.58 in Table 2.4). Interestingly, this branch is also thought to coincide with the recruitment of γ globins for fetal expression (double line in Fig. 2.6).

Globin genes are expressed at different developmental stages, so each gene might be subject to different selective pressures. To test for paralog-specific differences in selective pressure, I fitted the "Paralog" model, where ß, γ and ε globins have independent selective pressures (i.e., $\omega_ß \neq \omega_ε \neq \omega_γ$). This model fits the data significantly better than the one-ratio model, with parameter estimates $\omega_ß$ = 0.29, $\omega_ε$ =0.16, $\omega_γ$ =0.23 (Table 2.3). Average d_N value was 0.024 and average d_S value was 0.103. Those estimates are consistent with the ω estimates from the separate analysis of the paralogs, with ε globin to be more constrained than γ and ß globins (Table 2.1). Fitting additional models with two of the three ratios ($\omega_ß$, $\omega_ε$, $\omega_γ$) forced to be identical suggests that $\omega_γ$ is different from $\omega_ß$ and $\omega_ε$, while $\omega_ß$ and $\omega_γ$ are not significantly different (Table 2.3).

2.4 Discussion

2.4.1 Phylogeny and Gene Conversion

Gene conversion plays an important role in the evolution of multigene families, as it brings about the exchange of genetic material between related sequences (Schimenti 1994; Posada et al. 2002). It is a frequent mechanism of evolutionary change in globins and can act both to homogenize genes through concerted evolution (e.g., Aγ and Gγ in simian primates) or to introduce novelty among homologous genes

Table 2.4. Parameter estimates and log-likelihood scores for the γ globin gene under different sites and branch-site models.

Model	p	Parameter estimates	Positive Selection	ℓ
Site-specific models				
M1 (neutral)	1	$(\omega_0 = 0.00)$, $f_0 = 0.47$	No	-3170.38
		$(\omega_1 = 1.00)$, $(f_1 = 0.53)$		
M3 (discrete) ($K = 2$)	3	$\omega_0 = 0.06$, $f_0 = 0.74$	No	-3094.02
		$\omega_1 = 0.63$, $(f_1 = 0.26)$		
Branch-site models				
Model A	3	$(\omega_0 = 0)$, $f_0 = 0.45$	Yes	-3151.80
		$(\omega_1 = 1)$, $f_1 = 0.44$		
		$\omega_2 = 10$, $(f_{2+3} = 0.1)$		
Model B	5	$\omega_0 = 0.05$, $f_0 = 0.60$	Yes	-3084.69
		$\omega_1 = 0.63$, $f_1 = 0.22$		
		$\omega_2 = 4.58$, $(f_{2+3} = 0.18)$		

Note. p is the number of parameters in the ω ratio distribution. The foreground branch in the branch-site models is the branch leading to simian primates.

(e.g., cow εII and εIV). Gene conversion is known to affect gene phylogenies, as no single topology can relate the genes that have experienced conversion (Slatkin and Maddison 1989; Hudson et al. 1992; Maddison 2000). Given the general importance of the mechanism, its pervasiveness and its effects on phylogeny reconstruction, it is essential to test for gene conversion when topological discrepancies arise in a gene family tree (Drouin 2002). By using statistical methods, I found evidence of two unreported gene conversion events in ß globins, (*i*) among duplicates in the goat ß globin cluster, (*ii*) among duplicates in the mouse ß globin cluster, and confirmed many previously suggested cases. Furthermore, I suggest that the majority of misplacements in my β globin gene tree are the result of gene conversion events.

2.4.2 Different models of evolution by gene duplication

The traditional model of evolution by gene duplication predicts an increase in nonsynonymous substitution rate immediately after genes duplicate. It is a matter of debate whether this rate increase is due to a relaxation of selective pressure or to the action of positive selection for advantageous mutations (Massingham et al. 2001, Mazet and Shimeld 2002). Previous studies of the ß globin family supported the positive selection model, with this mode of evolution being suggested following the split of myoglobin and hemoglobin (Goodman 1981) and following the divergence of α and ß hemoglobins (Czelusniak et al. 1982). Accelerated amino acid evolution also occurred after the *en bloc* duplications within the ruminant artiodactyl lineage (Li and Gojobori 1983). In contrast to these examples, I found no significant evidence for a burst of nonsynonymous evolution in the branches postdating the initial duplications

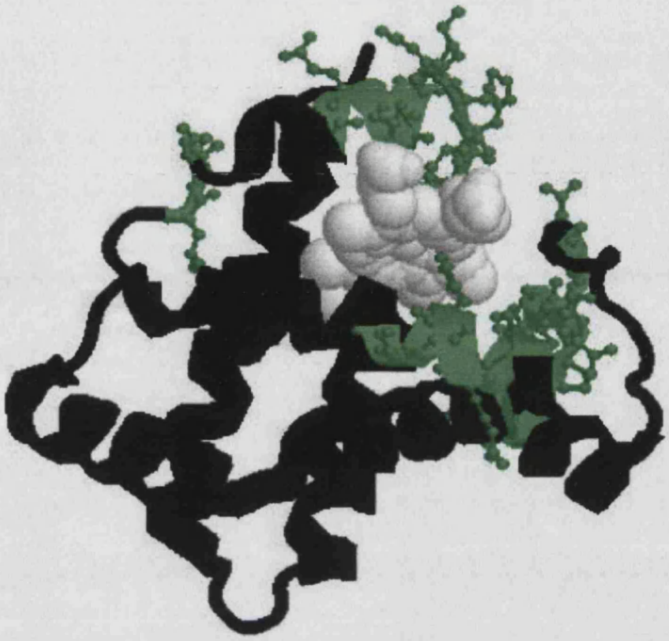


Fig 2.5. 3D structure of the ß globin chain of hemoglobin. The green part of the molecule corresponds to the residues conserved in all three genes that encode the subunit (ß, ε and γ globin genes). The ligand is depicted in white and using a different display mode for better appreciation. PDB 2hhb.

of the proto-ß and proto-ε genes, nor after the duplication giving rise to the ß and δ or to ε and γ clades, which correspond to the major duplication events within the gene family. I also tested for an increase in nonsynonymous substitutions at particular sites along the postduplication branches using branch-site models but failed to detect an evolutionary burst. These results appear consistent with a model suggesting a short-term advantage for preserving duplicate genes, where both paralogs initially evolve under equal constraints (Kondrashov et al. 2002).

There was one exception to the general pattern described above. In the lineage of stem-simians which represents the transition from embryonic to fetal expression of γ globins (Tagle et al. 1988, Fitch et al. 1991), I detected an acceleration in nonsynonymous substitution rates and identified positively selected sites. Whereas previously used methods employed raw counts of synonymous and nonsynonymous substitutions, and were thus unable to determine the source of amino acid evolution acceleration, the branch-site models indicated that nonsynonymous rate acceleration in the lineage of stem-simian γ globins was caused by positive Darwinian selection. It is possible that undetected gene conversion makes my tests for variable d_N/d_S rate ratios among branches more conservative, as sequences are more similar than they would be without its effects. It may also be that greater similarity among sequences reduced the power of the tests to detect an increase in d_N/d_S rate ratios following gene duplication. However, I make the observation that the tests were powerful enough in the case of the simian γ globin amino acid replacement acceleration. Furthermore, if adaptive evolution occurs by a single or a small number of substitutions it may not be detected by methods based on d_N/d_S ratios (Bielawski and Yang 2003). It is known that large phenotypic changes in globins can be achieved by only one or a few amino acid changes (Perutz 1983). A good example of the latter is provided by the deletion

of the NA1 valine residue from the protein chain encoded by γ-globin in some artiodactyls, which increases the oxygen affinity of the hemoglobin monomer (Poyart et al. 1992). Hence, in cases where I did not detect positive selection or even an increase in amino acid replacement rates, my findings do not exclude the possibility of neofunctionalization in ß globin genes by a few adaptive substitutions with large phenotypic effects.

The DDC model of gene copy preservation does not require a burst of nonsynonymous substitutions, and assumes purifying selection continues to act on both gene copies following duplication (Force et al. 1999; Zhang 2003). Nonetheless, if subfunctions are partitioned among the functional domains of the encoded protein, a potential outcome of the DDC model is heterogeneity in purifying selection among the gene copies. Dermitzakis and Clark (2001) proposed that identification of heterogeneity in patterns of amino acid substitution between different domains of the proteins encoded by paralogous genes could lead to the discovery of genes under subfunctionalization. While the DDC model has traditionally centred on regulatory sequences, I extend the possibility of finding subfunctionalization to protein-coding sequences by identifying heterogeneous selection pressure among paralogs. In the case of mammalian ß globins, genes are linked in a specific arrangement which, in most species, is known to be related to the order of expression of the genes (Hardison 1998). If the arrangement of ß globin genes in the cluster corresponds to a domain-like partition of function, each domain of expression could be subject to different selective pressures. Hence, my results are in agreement with a subfunctionalization model, as I found that each paralogous clade (i.e., domain of expression) is subject to significantly different selective constraints. My findings suggest a long-term process of divergence during which each paralog has been subject to different constraints by

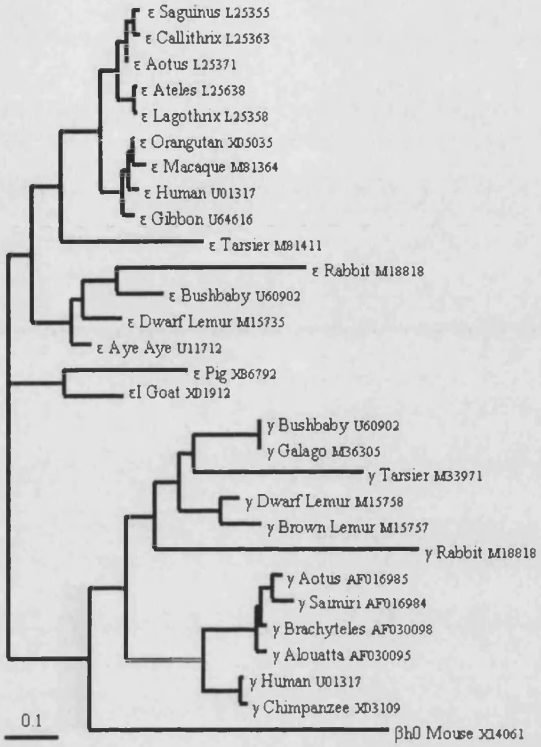


Fig. 2.6. Maximum likelihood tree of the ε, and γ globin genes from eutherian mammals. The double line corresponds to the branch where the G_γ and A_γ split is hypothesized to have occurred, in the ancestor of simian primates.

purifying selection, presumably related to differences in expression regulation. As described earlier, these findings do not exclude the possibility of brief episodes of increased amino acid replacement, in which case, other models (e.g., Ohta 1988b) may still be relevant to the evolution of β globins.

2.4.3 Selective pressure at sites in the vertebrate β globin genes

The ß globin gene is the only gene with sites predicted to be under positive selection in placental mammals. I identified twelve sites under Darwinian selection, consistent with the earlier study of Yang et al. (2000). These sites are located mostly at the exterior of the protein chain, with two sites located at the $\alpha_1ß_1$ interface between the α and ß subunits of hemoglobin (β116H, α111A and α115A). I tested for positive selection in the α globin genes currently available in GenBank, and found at least one positively selected site (115α) located at the $\alpha_1ß_1$ interface. My results raise the interesting possibility of long-term coevolution of some alpha and beta protein chain residues located in the $\alpha_1ß_1$ interface. I cover this topic more extensively in Chapter III.

Much is now known about what makes the globin fold a robust structure (Perutz et al. 1960; Bashford et al. 1987; Murzin and Finkelstein 1988; Brenner et al. 1997). Proteins whose secondary structures are mainly alpha helices, such as ß globin chains, are flexible and can easily accommodate many residues or prosthetic groups without disrupting tertiary or quaternary structural arrangements (Chothia et al. 1977; Efimov 1979). ß globins share the canonical features of the globin fold and have maintained a robust structure despite 200 million years of evolutionary divergence (Efstratiadis et al. 1980; Czelusniak et al. 1982). Arguably, the most important feature that explains the preservation of the globin fold is the clear conservation of

hydrophobic residues at buried positions in globin proteins (Lesk and Chothia 1980). In this study I identified regions conserved in all three ß globin genes, located in the interior or hydrophobic part of the subunit. Presumably, these conserved sites are involved in the maintenance of the secondary structures which in turn stabilize the tertiary and quaternary structures of hemoglobin. Furthermore, I found that some of the conserved sites are also part of empty concavities of the protein surface accessible to solvent (Liang et al. 1998). Concavities are particularly important as they are often associated with binding and catalytic activity (Liang and Dill 2001). For example, from the 23 sites which participate in interactions with the heme group, 15 correspond to the conserved sites in my study, with three involved in hydrogen bonding. With the exception of site 38Thr, all sites that participate in interactions with the heme ligand have hydrophobic-hydrophobic contacts, which stabilize the structure. Hence, during the long evolutionary history of the genes encoding the ß globin chain of hemoglobin, these functionally and structurally important sites have been preserved while at the same time a fraction of residues have been the targets of divergent fine tuning of the protein function.

2.4.4 The evolution of the β globin gene family

Gene family evolution reflects a balance between homogenization by unequal crossing over and gene conversion, and diversification by mutation (Ohta 2000). Both drift and selection play an important role in the evolutionary fate of duplicated genes, but only positive selection can account for the evolution of new functions (Ohta 1987). The dynamics of these forces are complicated (Ohta 2000), and my analysis of the ß globin family of genes illustrates this complexity. Gene conversion is clearly a frequent force for homogenization of some closely related members of this

family (e.g., Aγ and Gγ globins). As expected, gene conversion is less important to the evolution of the more divergent members, as it is prohibited when sequence divergence is too high (Ohta 2000). In addition to the partitioning of β globin paralogs into domains of expression, this gene family exhibits divergence both by positive Darwinian selection (β and γ globins), and by differential patterns of purifying selection pressure (γ and ε globins). While more tests are clearly necessary to fully discriminate between the DDC and Ohta models, I suggest that comparison between synonymous and nonsynonymous substitution rates provides a useful tool in studying relative roles of different evolutionary forces during the evolution of a gene family.

Chapter III

-

The evolution of the hemoglobin structure

3.1 Motivation

Tetrameric vertebrate hemoglobin is composed of two identical α subunits 141 residues long and two identical β subunits 146 residues long. Each subunit contains a heme group. Oxygen binding is cooperative and is associated with a large shift in the quaternary structure of the heterotetramer, the transition being from the deoxy (T) conformation to the oxy (R) form. There is at least another conformation available to hemoglobin known as R2, which is close to R. Two types of interfaces participate in conformational transitions, namely $\alpha 1\beta 1$ (or the equivalent $\alpha 2\beta 2$) and $\alpha 1\beta 2$ (or the equivalent $\alpha 2\beta 1$). Perutz (1970) referred to the $\alpha 1\beta 1$ interface as the packing surface and to the $\alpha 1\beta 2$ interface as the sliding surface. During the transition from the T to R conformation, the $\alpha 1\beta 2$ interface undergoes a large sliding movement, while the $\alpha 1\beta 1$ interface is practically unchanged. Hemoglobin properties such as oxygen affinity and cooperativity depend to a large extent on the intra-subunit and inter-subunit interfaces (Shionyu et al. 2001).

This chapter is about the evolution of hemoglobin structure and its relation to function. More specifically, it is about the way selective constraints have shaped the globin protein by preserving the functionally important residues while allowing changes in less compromising sites. Changes in structure may or may not be relevant to function. Are the changes we observe in globin proteins functionally relevant? I have tried to address this question in two important cases that involve not only β globins but other two related proteins, the monomeric myoglobin and the α chain of hemoglobin. This scope necessarily means going deeper in the globin phylogeny.

First, I analysed how the different sites in the α and β hemoglobin coding genes are constrained by varying selective pressure. My interest was to establish whether similar constraints act on both genes as they interact in hemoglobin to form

the two heterodimers in the molecule. Special emphasis was put on the sites, which in both genes, are part of the interfaces between the α and β hemoglobin chains. I hypothesize whether in order to maintain the crucial site-site interactions between the different chains in hemoglobin, sites at the αβ interfaces have coevolved.

Second, I was also interested in comparing hemoglobin genes with myoglobin because in this case there is a clear change in structural terms which has had a profound effect in function. The change here is the evolution of an oligomeric form in hemoglobin from a presumably monomeric ancestor. Myoglobin and hemoglobin also differ dramatically in their function. Whereas the latter is capable of allosteric regulation of oxygen trafficking, the former follows a hyperbolic equilibrium curve in response to oxygen concentration (Voet and Voet 1995). It is not clear whether oligomerization was brought about by selection or by chance but in either case we might be able to find traces of this evolutionary change by looking at the amino acid replacements between myoglobin and hemoglobin.

In order to investigate these questions I have aligned a number of sequences from myoglobin and α and β hemoglobins. I have estimated their gene trees and have measured selective pressure across their sites and among the different globin lineages. Also, I have inferred the ancestral states of those protein coding genes and looked at the direction and importance of the amino acid replacements in terms of structure and function.

3.2 Theory and Methods

4.2.1 Data and Phylogenetic Analysis

Three different datasets were constructed: (i) an alignment of myoglobin and β hemoglobin sequences that I will call the "myoglobin-hemoglobin" dataset; (ii) an

alignment of α and β globin sequences that I will refer to as the "α-β dataset"; and (iii) an alignment of α globin sequences for the individual analysis of this gene. Species names and GenBank accession numbers included in the three datasets are listed in Table 3.1. A total of 36 nucleotide sequences were retrieved from GenBank for the myoglobin-hemoglobin dataset, of which, 16 were β hemoglobin genes (column "β Globin" in Table 3.1) and 20 were myoglobin genes (column "Myoglobin" in Table 3.1); For the α-β hemoglobin dataset I used the 13 α globin sequences (column "α Globin" in Table 3.1) and 10 β globin sequences (marked with an asterisk in column "β Globin" in Table 3.1). The individual analysis of the α globin was done using the 13 nucleotide sequences listed in Table 3.1 (column "α Globin"). The individual analysis of the β globin gene was done previously in Chapter II (Table 2.1). All sequences were aligned using ClustalX (Thompson et al. 1997). Manual adjustments were done and gaps were removed.

Trees were estimated from the nucleotide sequences using maximum parsimony and maximum likelihood methods. The model of nucleotide substitution was HKY85 in all cases. All codon positions were analyzed together. In the case of ML trees I used a minimum evolution phylogeny as a starting topology from which model parameters were optimized. Both MP and ML methods resulted in similar topologies for each gene. Bootstrap analysis was performed using PAUP* (Swofford 1998) to assess relative support for internal branches.

3.2.2 Detection of positive selection across sites

I was interested in detecting positively selected sites, as these may be associated with important changes in globin structure and function. A statistical approach was taken

Table 3.1. List of the species comprised in the myoglobin-hemoglobin, the α–β globin, and the α globin datasets including their GenBank accession numbers

Myoglobin	GenBank Accession Number	β Globin	GenBank Accession Number	α Globin	GenBank Accession Number
Homo sapiens	NM_005368	β Human*	U01317	Mouse	V00714
Mus musculus	BC025172	βBushbaby*	U60902	Rat	U62315
Rattus norvegicus	AF197916	β Brown Lemur*	M15734	Cow	AJ242797
Sus scrofa	M14433	β Mouse*	J00413	Buffalo	AJ242731
Bos Taurus	NM_173881	β Sheep*	X14727	Sheep	X70213
Thunnus obesus	AB104433	β Rabbit*	V00882	Goat	J0044
Cannichthys rhinoceratus	AY341058	β Pig*	X86791	Horse	U70191
Makaira nigricans	AF291833	β Tarsier*	J04429	Rabbit	M11113
Thunnus albacares	AF291838	β Hare*	Y00347	Bushbaby	M29648
Euthynnus pelamis	AF291837	β Rat*	X15009	Rhesus monkey	J04495
Thunnus orientalis	AF291836	β Duck	X15739	Orangutan	M12158
Scomber japanicus	AF291835	β Chicken	V00409	Gibbon	M94634
Sarda chiliensis	AF291834	β Xenopus	X03142	Human	J00153
Thunnus alalunga	AF291832	β Xtropicalis	Y00501		
Thunnus thynnus	AF291831	β Tachyglossus	L23800		
Nothothenia coriceps	U71058	β Didelphis	J03643		
Gobionotothen gibberifrons	U71057				
Cryodraco antarticus	U71056				
Pseudochaenyctys georgianus	U71055				
Chionodraco ratrospinosus	U71059				

Note: The "myoglobin-hemoglobin" dataset comprises species in columns "Myoglobin" and "β globin"; the "α–β dataset" includes sequences in column "α Globin" and sequences marked with an asterisk in the column "β Globin". The "α Globin dataset" comprises the 13 sequences in column "α Globin".

to detect positive selection acting on sites in the myoglobin-hemoglobin, the α–β hemoglobin dataset and the α globin gene. I used the same method that was used in Chapter II to analyze variable selective pressure and detect positive selection in β globin cluster genes. Here too, I used several codon models that allow for heterogeneous d_N/d_S (ω) ratios at sites in the alignments, and in three models, for ω > 1, that is, the detection of positive selection (Nielsen and Yang 1998, Yang et al. 2000). In the simplest model (M0 or one-ratio model), the *ω* ratio is an average over all the sites. The "neutral" model (M1) allows for conserved sites where ω = 0 and completely neutral sites where *ω* = 1. The "selection" model (M2) adds a third class to M1 at which ω can take values > 1. The discrete model (M3) uses an unconstrained discrete distribution with different ω ratios for *K* different classes of sites. Model M7 (beta) assumes a beta distribution of ω over sites. Model M8 (beta&ω) adds an extra class of sites to M7, thereby allowing ω values >1. Likelihood ratio tests (LRTs) were conducted to test M0 (one-ratio) against M3, and M7 (beta) against M8 (beta & ω). All analyses were based on the unrooted gene-tree topologies, and used the codeml program in the PAML package (Yang 1997). Sites found to be under positive selection were studied by ancestral reconstruction analysis in order to evaluate their importance in terms of structure and function.

3.2.3 Analysis of selective pressure among lineages

I analysed the variable selective pressure acting on the different lineages of: (i) myoglobin and hemoglobin, and (ii) the α and β globin genes. I used the myoglobin-hemoglobin and the α-β globin datasets and constructed two likelihood ratio tests to determine whether positive selection was involved in the functional divergence between myoglobin and hemoglobin and between α and β globin following their split.

For each dataset I implemented two LRTs that allow for variable ω ratios among branches in the tree (Yang 1998; Bielawski and Yang 2003). The null model assumed the same ω for all lineages in the tree. In the first test, I assigned different ω ratios for the branches in each paralogous clade, so that branches in one clade had ω_0, and branches in the other clade had ω_1. This model (i.e. $\omega_0 \neq \omega_1$) was compared with the null model (i.e. $\omega_0 = \omega_1$) by means of an LRT. In the second test, using a rooted tree, I assigned an ω ratio for the ancestral branch of the two paralogous lineages (i.e. the ancestral branch to myoglobin and hemoglobin in one test, and the ancestral branch to α and β globin gene clades in the other); and different ω ratios were assigned for the branches in each paralogous clade, so that the ancestral branch had ω_0; branches in one clade had ω_1, and branches in the other clade had ω_2. Again, the alternative model (i.e. $\omega_0 \neq \omega_1 \neq \omega_2$) was compared with the null model (i.e. $\omega_0 = \omega_1 = \omega_2$) by an LRT.

3.2.4 Ancestral state reconstruction

It is important to investigate the substitutions that have occurred at sites in the protein coding genes with reference to the ancestral state of those sites because this kind of analysis provides us with an idea of the "direction" (change from ancestral to derived state) of the codon substitutions along evolutionary lineages. This is relevant when we are interested in finding the adaptive importance of substitutions. First, I was interested in inferring the ancestral states and the direction of the amino acid changes associated with the transition from a monomeric to an oligomeric globin structure. To do this I conducted an ancestral state reconstruction using the myoglobin-hemoglobin dataset in order to compare the amino acid states at the two ancestral nodes of myoglobin and hemoglobin. Specifically, I looked at the sites in the $\alpha\beta$ interfaces (Chien and Lukin 2001). Second, I reconstructed the ancestral states of the amino

acids that in both the α and the β globin chains correspond to αβ interfaces. This analysis allowed me to determine whether, in order to maintain the crucial site-site interactions at the interfaces, coordinated changes had occurred in the two hemoglobin chains. This required a similar sampling of the two datasets in order to compare the changes mapped in the two phylogenies. Since sites at the αβ interfaces are different in the α and β globin chains, I conducted the ancestral state reconstructions separately using the two individual datasets for each gene. Relevant replacements are mapped in Figures 3.3 and 3.4. Finally, it was also important to infer the ancestral states of the positively selected sites in α and β globins and in the myoglobin-hemoglobin dataset in order to establish whether the changes brought about by positive selection have a special relevance in terms of structure and function. Relevant replacements in α and β globins are mapped in Figures 3.3 and 3.4.

Ancestral state reconstructions were based on the maximum likelihood method described by Yang et al. (1995). Yang and collaborators developed a statistical method for reconstructing the nucleotide and amino acid sequences of extinct ancestors, given the phylogeny and sequences of extant species. The authors proposed a model-based likelihood approach to reconstructing ancestral sequences, which given the data at the site, the conditional probabilities of different reconstructions are compared and the reconstruction with the highest conditional probability is the best estimate at the site. Also, a measure of the accuracy of the reconstruction is possible since the method allows the calculation of the probability that the reconstruction is correct. Estimates of parameters in the models such as branch lengths of the tree are used to evaluate the possible reconstructions. In the approach of Yang et al. (1995), the ancestral amino acids are discrete random variables in the model and are estimated by maximizing the posterior probabilities. The assignment of a character state to a

node is obtained by summing the contribution to the probability of observing the data at a given site over all reconstructions at the site that assign the same amino acid to the node. Consequently, the best assignment at a node will be the amino acid that has the highest posterior probability. Once the ancestral nodes were inferred, I mapped the positively selected substitutions onto the respective rooted phylogeny in order to establish the direction of the changes.

3.2.5 Looking for correlated mutations in the evolution of α and β globins

In the case of α and β globin genes I was interested in identifying correlated substitutions in the two genes. Correlated changes are suspected when one identifies that a change at a given site occurred in the same branch in the two genes under comparison. It is thus necessary to have a similar sampling for both paralogs. In order to look for such correlated changes between α and β globin I used the program Plotcorr developed by Pazos et al. (1997) to investigate the pattern of correlated mutations at the interface sites. This program allows the user to compare amino acid replacements between two proteins and determine whether the changes occurred in a correlated fashion. According to the authors, correlated mutations are those that indicate a tendency of positions in proteins to mutate coordinately. Such coordinated mutations frequently occur between proteins that interact, such as monomers in a dimer. Correlated mutations are calculated according to Göbel et al. (1994). Each position in a multiple protein alignment is coded by a distance matrix. This position-specific matrix contains the distances between all pairs of sequences at that position. The distances are defined according to the scoring matrix of McLachlan (1971). Pazos et al. calculate a correlation coefficient for each pair of positions. They propose that it is possible to detect the signal for correlated mutations by studying compensatory

mutations between interacting proteins. Furthermore, the signal detected by their method corresponds mainly to networks of positions that have undergone compensating mutations during evolution.

3.2.6 Determining the structural and functional relevance of amino acid replacements

I was interested in testing the structural and functional relevance of residue substitutions between the proteins encoded by the α and β globin genes and between myoglobin and hemoglobin. Positively selected sites and sites at to the two αβ interfaces were mapped in the ancestral reconstruction analysis, and analyzed for the impact of the replacements they underwent across taxa. I classified replacements as conservative or adaptive according to the extent of the difference in terms of the physicochemical properties of ancestral and derived states. The physicochemical properties considered were volume, charge, and reactivity as described by Creighton (1993) and according to the classification by polarity and volume proposed by Zhang (2000).

3.3 Results

3.3.1 Phylogenetic Analysis

Myoglobin-hemoglobin.- The 35 myoglobin and hemoglobin sequences (Table 3.1) analyzed in a single dataset were used for phylogenetic reconstruction. Fig. 3.1 shows the rooted ML tree with bootstrap support proportions. Analysis using the correct species tree gave similar results. The myoglobin and hemoglobin clades are both monophyletic. Within the myoglobin clade, the fish, the rodent, and the artiodactyl clades are also monophyletic and agree well with their expected species tree. The human sequence appears more closely related with the

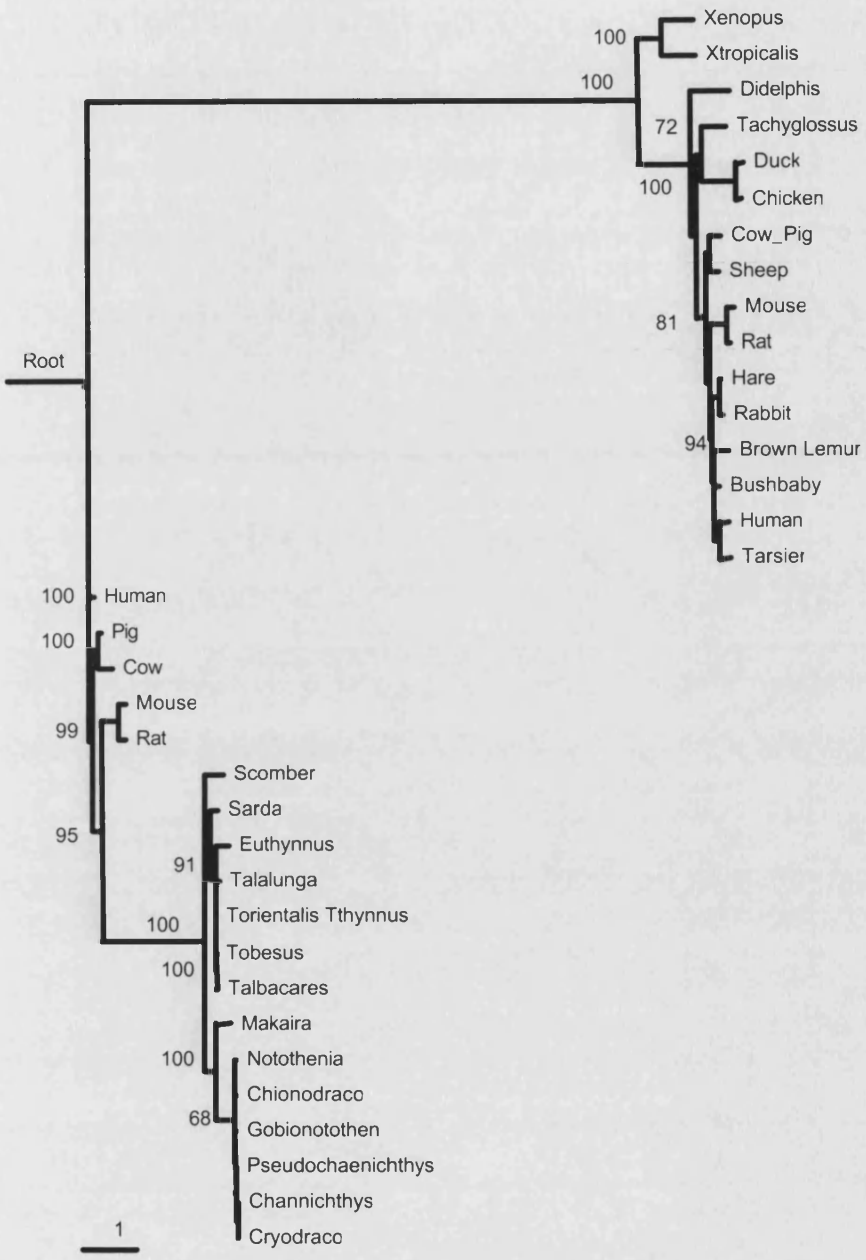


Figure 3.1. ML tree for the myoglobin-hemoglobin dataset, here rooted for clarity. Bootstrap support proportions are shown next to nodes. The clade on top right corresponds to hemoglobin sequences; the clade at bottom left corresponds to myoglobin sequences. The ancestral myoglobin node has 99 % bootstrap support and the ancestral hemoglobin node has 100 % bootstrap support.

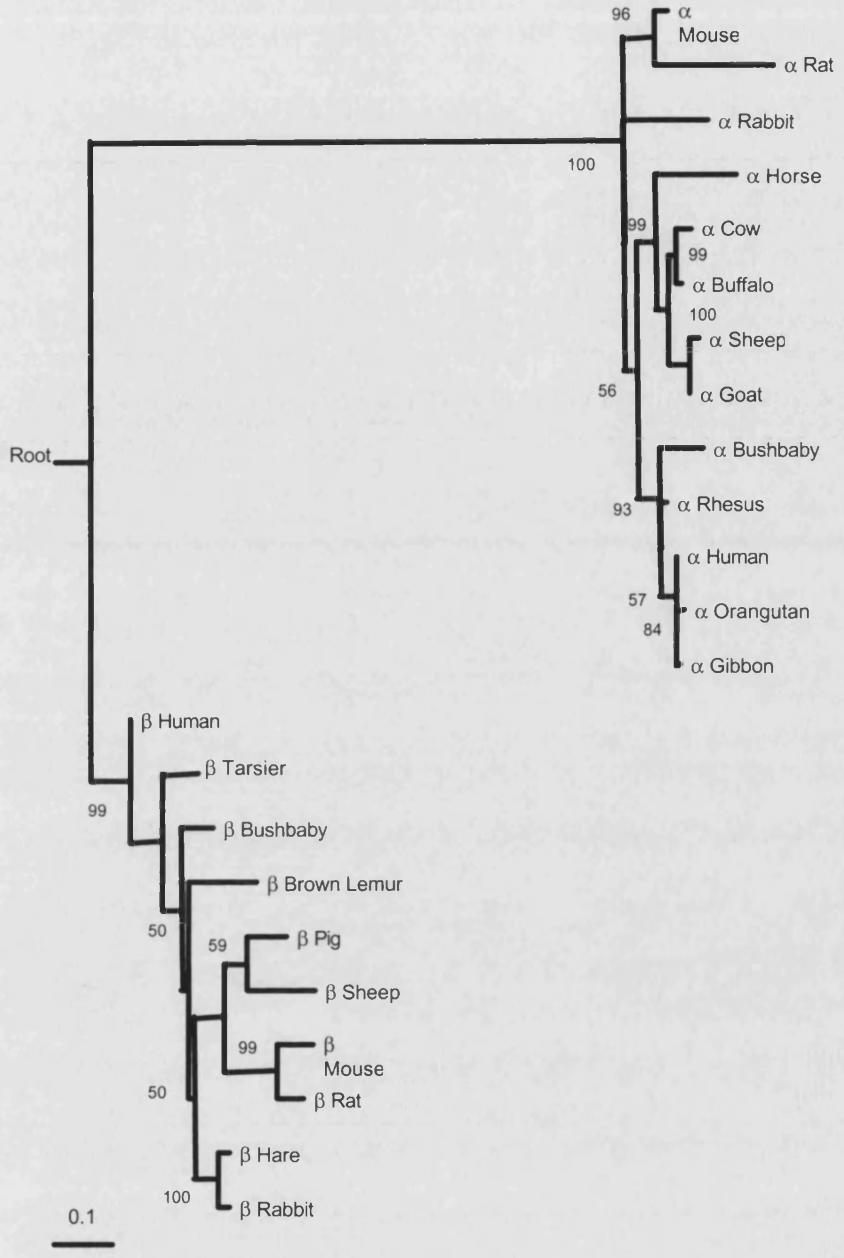


Fig. 3.2. ML tree for the α and β hemoglobin dataset, here rooted for clarity. Bootstrap support proportions are shown next to nodes. The clade on top right represents α globin sequences; the clade at bottom left corresponds to β globin sequences; the ancestral node to the α clade has 100% bootstrap support and the ancestral β clade has 99 % bootstrap support.

artiodactyls than the rodents, I speculate whether this unusual placement might be due to sampling bias as there is only one primate sequence in the dataset. Within the hemoglobin dataset, eutherian mammals are monophyletic and the expected species tree was recovered. Frogs are monophyletic and are at the base of the hemoglobin clade. The placement of the birds is unexpected, as it appears as a sister taxon with the monotreme gene. Sampling bias may have caused this result. The kangaroo (*Didelphis*) is placed as sister to the eutherians.

α–β globins.- I used the 13 α globin and the 10 β globin sequences listed in Table 3.1 to estimate the gene tree. I obtained a similar topology with both MP and ML phylogenetic reconstruction methods. Fig. 3.2 shows the rooted ML tree for the α–β gene tree. Both α and β globin clades are monophyletic. Within the α clade all the different vertebrate orders represented are also monophyletic. In the β clade I recovered monophyletic vertebrate orders. Some taxa have very short branch lengths (e.g., β tarsier). The monophyletic groups within the α clade were highly supported by bootstrap analysis with proportions > 90%. In contrast, lower bootstrap proportions (> 60%) were obtained for the monophyletic groups in the β clade.

α globin.- The 13 α globin genes in Table 3.1 were used to infer the phylogenetic tree. Both MP and ML methods produced similar topologies. Fig. 3.3 shows the rooted ML tree for α globin. In the inferred gene tree, the different mammalian orders are monophyletic. Also, I recovered the expected species tree within monophyletic clades except in the primate clade, where the orangutan and gibbon appear as sister taxa. The Rhesus monkey has a very short branch length. Bootstrap support of monophyletic clades have confidence levels > 60%.

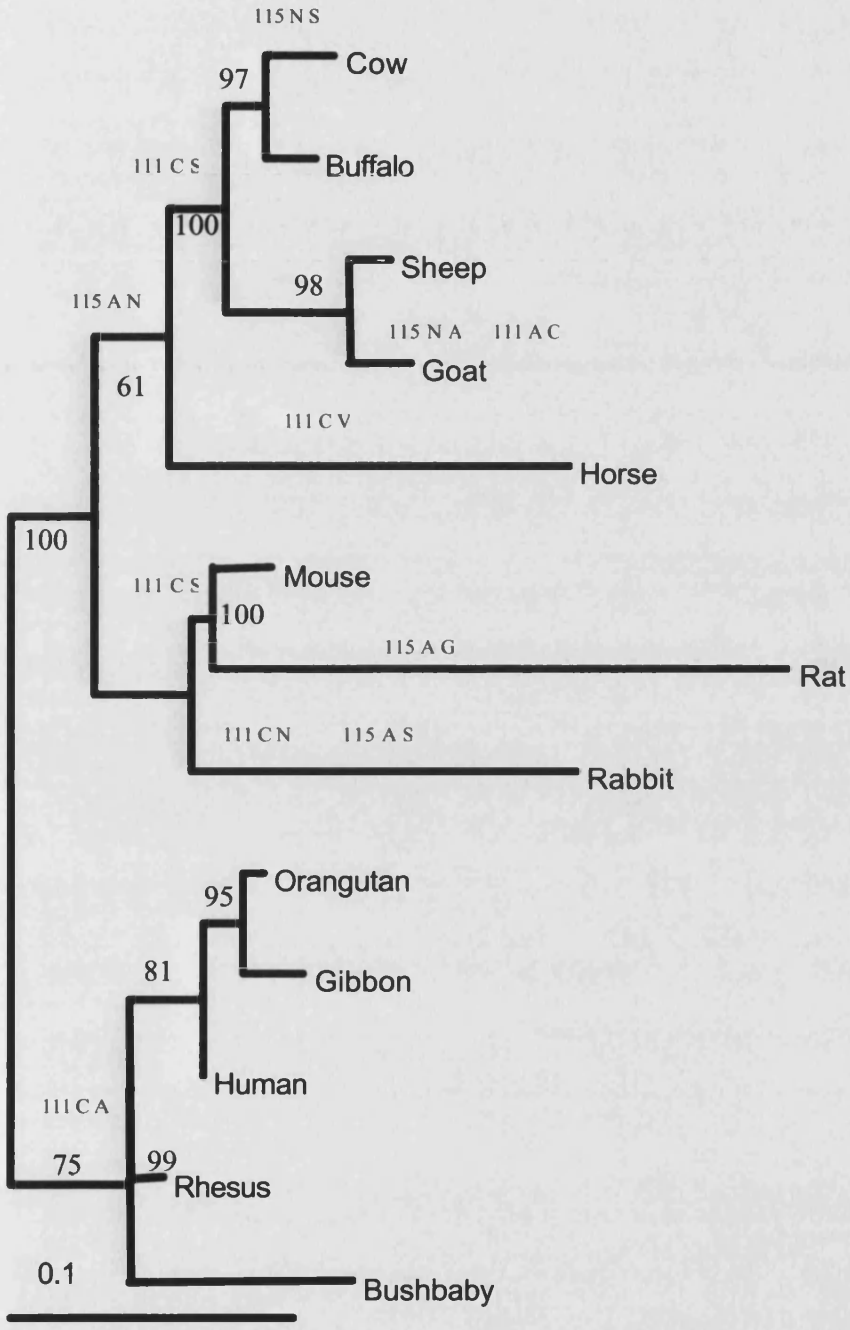


Fig. 3.3. ML tree for the α globin sequences. Numbers in black indicate bootstrap support. Relevant ancestral state reconstructions in blue are mapped onto branches indicating substitutions that occurred in those lineages. The number refers to the site in the protein involved; the letter at left is the original amino acid (at the root) for that site and the letter at right is the derived amino acid. Sites 111 and 115 are part of the α1β1 interface.

3.3.2 Analysis of selective pressure across sites

The analysis of selective pressure across sites in the studied genes (i.e. myoglobin, α and β globin) was conducted using the codeml program (Yang 1997). I implemented various codon models and performed LRTs to check for variability of selective pressure among sites and, more interestingly, positive selection acting at some sites.

Myoglobin-hemoglobin.- Table 3.2 presents the results of the analysis of selective pressure across sites in the myoglobin-hemoglobin alignment. The simplest model M0 gave an overall estimate of the ω ratio of 0.20. The selection model M2 indicated that 4% of sites are evolving under strong purifying selection (ω fixed at 0); that 92% of sites are evolving neutrally (ω fixed at 1.00); and that 4% of sites are positively selected (ω = 3.11). Positively selected sites include sites 8, 16, 17, 45, 62, 80, and 126, of which only sites 8 and 126 have a posterior probability > 95%. Also, model M3 resulted in 23% of sites to be evolving under strong purifying selection (ω = 0.03); 63% of sites are somewhat less conserved (ω = 0.17); and 12% of sites are evolving under relaxed pressure (ω = 0.74) although it did not detect positive selection. Model M8 indicated that 7% of sites are evolving under relaxed constraints (ω = 0.9) again without positive selection. The LRT M0 vs. M3 was significant (2δ = 177.39; df = 2; P-value < 0.0001) indicating there is variable selective pressure among sites in the genes; also the LRT M7 vs. M8 was significant (2δ = 9.21; df = 2; P-value = 0.01) although positive selection was not detected.

Table 3.2. Parameter estimates and likelihood scores for the myoglobin-hemoglobin dataset under site-specific models

Model	Parameter Estimates	ℓ
M0 (one-ratio)	$\omega = 0.20$	-5908.60
M1 (neutral)	$(\omega_0 = 0)$, $f_0 = 0.04$, $(\omega_1 = 1)$, $(f_1 = 0.96)$	-6126.87
M2 (selection)	$(\omega_0 = 0)$, $f_0 = 0.04$, $(\omega_1 = 1)$, $f_1 = 0.92$, $\omega_2 = 3.11$, $(f_2 = 0.04)$	-6117.18
M3 (discrete)	$\omega_0 = 0.03$, $f_0 = 0.23$ $\omega_1 = 0.17$, $f_1 = 0.63$, $\omega_2 = 0.74$, $(f_2 = 0.12)$	-5819.91
M7 (beta)	$p = 0.95$, $q = 3.23$	-5825.87
M8 beta&ω	$p = 1.39$, $q = 6.45$, $f_0 = 0.93$ $\omega_1 = 0.9$, $(f_1 = 0.07)$	-5821.27

Likelihood ratio test statistics for comparing site-specific models for the myoglobin-hemoglobin dataset

LRT	2δ	df	*P*-value
M0 (one-ratio) vs. M3 (discrete)	177.39	2	<0.0001
M7 (beta) vs. M8 (beta&ω)	9.21	2	0.010

Table 3.3. Parameter estimates and likelihood scores for the α-β dataset under site-specific models

Model	Parameter Estimates	ℓ
M0 (one-ratio)	$\omega = 0.23$	-3905.9769
M1 (neutral)	$(\omega_0 = 0)$, $f_0 = 0.17$, $(\omega_1 = 1)$, $(f_1 = 0.83)$	-3988.2276
M2 selection	$(\omega_0 = 0)$, $f_0 = 0.08$, $(\omega_1 = 1)$, $f_1 = 0.25$, $\omega_2 = 0.1$, $(f_2 = 0.66)$	-3797.1442
M3 discrete	$\omega_0 = 0.07$, $f_0 = 0.68$ $\omega_1 = 0.54$, $f_1 = 0.24$, $\omega_2 = 1.22$, $(f_2 = 0.08)$	-3792.7767
M7 beta	$p = 0.45$, $q = 1.29$	-3805.0356
M8 beta&ω	$p = 2.03$, $q = 21.03$, $f_0 = 0.75$ $\omega_1 = 0.81$, $(f_1 = 0.25)$	-3793.8039

Likelihood ratio test statistics for comparing site-specific models for the α-β dataset

LRT	2δ	df	*P*-value
M0 (one-ratio) vs. M3 (discrete)	226.40	2	<0.0001
M7 (beta) vs. M8 (beta&ω)	22.46	2	<0.0001

α–β globin dataset.- Table 3.3 presents the results for the detection of positive selection in the α–β globin dataset. Results show that substitutions at most sites are very constrained and only a small fraction of sites is subject to positive selection. Model M0 gave an overall estimate of the ω ratio of 0.23. The selection model M2 indicated that 8% of sites are evolving under strong purifying selection (ω fixed at 0); 25% of sites are evolving neutrally (ω fixed at 1.00); and 66% of sites are slightly less constrained with ω = 0.1. Model M3 resulted in 68% of sites to be evolving under strong purifying selection (ω = 0.07); 24% of sites are less constrained (ω = 0.54); and 8% of sites are under positive selection (ω = 1.22). Model M3 found 7 positively selected sites but only one with a high posterior probability of >99% (site 7). Model M8 identified 25% of sites evolving under weaker selective pressure (ω = 0.81). The LRT M0 vs. M3 was significant ($2\delta = 226.40$; df = 2; P-value < 0.0001); and the LRT M7 vs. M8 was also significant ($2\delta = 22.46$; df = 2; P-value < 0.0001). These results indicate that there is among-site variation in selective constraints and the evidence for positive Darwinian selection is weak.

α globin.- Results for the analysis of selective pressure across sites in the α globin gene are listed in Table 3.4. In the case of the α globin gene, I observed that the majority of sites are evolving under strong purifying selection. The simplest model M0, which assigns a single ω ratio over all sites in the gene, resulted in a ratio of 0.18. Model M2, which has a site class that allows ω to take values > 1, showed that 39% of sites are evolving neutrally (ω fixed at 1.00); 58% of sites are highly conserved (ω fixed at 0); and 3.4% of sites are positively selected (ω = 4.02). Sites 16, 69, 112, 116 and 132 were found to be under positive selection, however, none had a posterior

Table 3.4. Parameter estimates and likelihood scores for the α globin gene under different site-specific models.

Model	Estimates of Parameters	ℓ
One-ratio Model (M0)	ω_0=0.18	-2102.8461
Site specific		
Neutral (M1)	ω_0=0.00 , f_0=0.58(ω_1=1.00), (f_1=0.42)	-2045.4314
Selection (M2)	(ω_0=0.00) , f_0=0.58 ω_1=1.00, f_1=0.39 ω_2=4.02, (f_2=0.034)	-2043.1761
Discrete (M3), k=3	ω_0=0.001, f_0=0.48 ω_1=0.17, (f_1=0.35) ω_2=1.13, (f_2=0.17)	-2010.8425
Beta (M7)	p=0.16 , q=0.54	-2013.9719
Beta& ω (M8)	p=0.30 , q=2.96 , f_0=0.85 ω_1=1.19, (f_1=0.15)	-2011.5269

Note: sites in bold have posterior probabilities >95%.

Likelihood ratio test statistics for α globin gene.

LRT	2δ	df	P-value
M0 vs M3	184.007	2	<0.0001
M7 vs M8	4.89	2	0.087

probability higher than 95% so they may represent type I errors. Model M3 detected that 48% of sites are under strong purifying selection ($\omega = 0.001$), 35% of sites are fairly conserved ($\omega = 0.17$) and only 17% of sites are under positive selection ($\omega = 1.13$), albeit only roughly above 1. In this case, I recovered 24 sites under positive selection but only 6 have posterior probabilities higher than 95% (sites 22, 68, 78, 111, 115 and 131). Finally, model M8, which can also detect sites under positive selection, found that 15% of sites have ω values > 1 (21 sites were detected but only sites 68, 111 and 115 have high posterior probabilities). The LRT M0 vs. M3, which is a test of variable selective pressure among sites, was significant ($2\delta = 184.007$; df = 2; P-value < 0.0001). The LRT M7 vs. M8, which is a test of positive selection, was also significant ($2\delta = 4.89$; df = 2; P-value = 0.027).

3.3.3 Analysis of selective pressure among lineages

Myoglobin-hemoglobin.- The results of the LRTs constructed to test the action of positive selection following the duplication of myoglobin and hemoglobin are listed in Table 3.5. In the first case, I tested whether the average selective pressure on the myoglobin and hemoglobin lineages was significantly different to their ancestral branch. To do so, I constructed a two-ratio test by assigning a different ω ratio to the ancestral branch and one ω ratio to the rest of the branches in the tree. I compared that model with the null model where all branches in the tree have a single ω ratio. The LRT was not significant ($2\delta = 2.40$; df = 1; P-value = 0.12). The branches in the myoglobin and hemoglobin lineages had an average $\omega = 0.20$, indicating that purifying selection acted on these genes to maintain their respective function. On the other hand, the ancestral branch had $\omega = 2.44$, indicating that positive selection was driving the functional divergence of myoglobin and hemoglobin immediately

following the duplication event. This LRT was perhaps too stringent, as the average of all but one branch in the tree may have led to an underestimation of the likelihood of the alternative model. In order to alleviate this lack of power, I decided to add one parameter by adding a third class of sites to the alternative model just described. The second LRT constructed compared the null hypothesis (i.e., one ω ratio across all the branches in the tree) with an alternative hypothesis which assigned a different ratio to each paralogous clade and one ratio to the branch separating the myoglobin and hemoglobin clades in Fig. 3.1. Note that analysis was done using the unrooted tree. The LRT was also not significant although there was a slight improvement in the likelihood score ($2\delta = 3.03$; df = 2; P-value = 0.22). In this case, the d_N/d_S ratios obtained for the myoglobin and hemoglobin lineages were $\omega_{myo} = 0.22$ and $\omega_{hemo} = 0.19$, respectively; and the d_N/d_S ratio for the ancestral branch was $\omega_{ancestral} = 2.40$. The lack of significance in the two LRTs described above, suggests that immediately following the duplication which gave rise to myoglobin and hemoglobin there was not an episode of intense positive selection.

α-β globin dataset.- The results of the LRTs constructed to test the action of positive selection following the duplication of α and β globin are listed in Table 3.6. In the first case, I tested whether the selective pressure acting on the two paralogous lineages was significantly different from that of the ancestral branch (before their divergence). To do so, I constructed a two-ratio test by assigning each lineage a different ω ratio. I compared that model with the null model where all branches in the tree have a single ω ratio. The LRT was significant ($2\delta = 4.53$; df = 1; P-value = 0.03). The ancestral

Table 3.5. Maximum likelihood estimates for the myoglobin-hemoglobin dataset of ω ratios under branch-specific models and likelihood ratio test statistics when the model is compared with the null model M0 (one-ratio)

Alternative model	Parameter estimates	2δ	df	*P*-value
2-ratios	$\omega_{myo\text{-}hem}$ =0.20, $\omega_{ancestral}$ = 2.44	2.40	1	0.12
3-ratios	ω_{myo} = 0.22, ω_{hem} = 0.19, $\omega_{ancestral}$ = 2.40	3.03	2	0.22

Table 3.6. Maximum likelihood estimates for the α–β globin dataset of ω ratios under branch-specific models and likelihood ratio test statistics when the model is compared with the null model M0 (one-ratio)

Alternative model	Parameter estimates	2δ	df	*P*-value
2-ratios	$\omega_{\alpha\text{-}\beta}$ =0.22, $\omega_{ancestral}$ = 0.52	4.53	1	0.03
3-ratios	ω_{α} = 0.18, ω_{β} = 0.27, $\omega_{ancestral}$ = 0.53	9.59	2	0.008

branch had an $\omega = 0.52$, indicating that after duplication, a slightly increase in nonsynonymous substitutions occurred although purifying selection continued to act on the genes. On the other hand, the α and β lineages evolved with an average $\omega = 0.22$, indicating higher constraints once the two genes had diverged in function. The second LRT constructed compared the null hypothesis (i.e., one ω ratio across all the branches in the tree) with an alternative hypothesis which assigned a different ratio to each paralogous lineage and one ratio to the branch that occurs between the α and β clades in Fig. 3.2. Note that the analysis was done using the unrooted tree. The LRT was significant ($2\delta = 9.59$; df = 2; P-value = 0.008). The d_N/d_S ratios obtained for the α globin and β globin lineages were $\omega_\alpha = 0.18$ and $\omega_\beta = 0.27$, respectively; and the d_N/d_S ratio for the ancestral branch was $\omega_{ancestral} = 0.53$. The results suggest that immediately following the duplication which gave rise to α and β globin there was a slight increase in nonsynonymous substitutions followed by purifying selection in the two differentiated hemoglobin lineages.

3.3.4 Ancestral state reconstruction

I conducted an ancestral state reconstruction for myoglobin-hemoglobin and the individual α and β globin datasets in order to establish the direction of the amino acid replacements at the interfaces. This was done in order to determine whether changes at those sites had any relevance in terms of their dimerization potential (i.e., changes that create appropriate surfaces for interfaces such as exposed hydrophilic residues that promote H-bonding and salt bridges) and to investigate the possibility of correlated changes at $\alpha\beta$ interfaces between the two hemoglobin subunits. I also reconstructed the ancestral states at the positively selected sites in each chain, as these replacements could be related with important changes in terms of structure and

function in the evolution of the different globin chains. The method developed by Yang et al. (1995) reconstructs the pathways of amino acid substitutions for each site in the sequences. The branch lengths of the species trees were estimated by using the JTT model of amino acid substitutions (Jones et al. 1992). All the generated reconstructions were evaluated by assigning all observed amino acids at a given site to each interior node. Only the reconstructions with posterior probabilities within the range of 0.05 to 1.0 are shown in tables.

Myoglobin-hemoglobin.- I wanted to infer the amino acid state at each of the two ancestral nodes of myoglobin and hemoglobin in order to compare the physicochemical characteristics of the amino acids present at the $\alpha\beta$ interfaces. There are two interfaces in hemoglobin, the$\alpha 1\beta 1$ interface occurs between intradimeric monomers, and the $\alpha 1\beta 2$ interface occurs between interdimeric monomers (Chien and Lukin 2001). Table 3.7 shows the amino acid state at the ancestral node of myoglobin and hemoglobin (human sequence used as reference) and the changes that have occurred at the $\alpha 1\beta 1$ and $\alpha 1\beta 2$ interfaces. In the case of the former interface, of the 24 sites that are part of that interface, 16 sites differ between the ancestral node of hemoglobin and the ancestral node of myoglobin. Some of those replacements occurred only once in the phylogeny and that is exactly along the branch that separates the two ancestral nodes. Roughly 50% of the changes are radical (i.e. imply an important change in physicochemical properties). At the $\alpha 1\beta 2$ interface, out of the 17 sites that conform it, 13 sites had changed between the myoglobin and the hemoglobin ancestral nodes. Again most of the replacements were radical as is shown in the next section.

Table 3.7. Replacements at Interface Sites in the Myoglobin-Hemoglobin Dataset

Interface	Site	Myoglobin	Hemoglobin	Replacements across the phylogeny
α1β1				
	26	G	E	**GE** ED
	30	L	R	**LR**
	33	L	V	LV
	34	F	V	**FV**
	35	K	Y	**KT KY**
	51	E	P	AG AG AP **VS EA** AG AP AP **AV**
	55	K	M	SA **KG KM** ML **MS**
	101	Y	E	EQ **YN YE**
	108	V	N	**AC** VI **VA** VI VI **DG AV VN** ND ND
	109	I	V	VI IL IV **VM** VI
	111	Q	V	QE **QH** KH **QK QV** VI
	112	V	I	IV VI IV **VC** IV
	115	S	A	SA AG **SK SE** SA AS AG
	116	K	R	**HE RE** KR KR RH **RA RA**
	119	G	G	GS LM ST **GL** GS
	120	D	K	**KN** KH **KN** ST **DK KS**
	125	A	Q	**AQ**
	128	A	A	**AD AV** AG
	131	G	Q	GA **TK TQ** GT **GQ** QE
α1β2				
	34	F	V	**FV**
	35	K	Y	**KT KT TK KY**
	37	E	W	ED ED **EW**
	39	L	Q	**QT LQ QL**
	40	E	R	ED **EK ER**
	43	D	E	ED ED **EG DP** DE ED **EA ES**
	93	H	K	**KQ** DE HK **KD**
	98	P	V	PA **PV**
	99	V	D	VI **VD**
	100	K	P	**KN KI KP**
	101	Y	E	EQ **YN YE**
	102	L	N	**LF LN**
	105	I	L	**LR** LM IL **LR**

Note: the amino acid changes shown in the column "Replacements" were obtained from the ancestral state reconstruction (Yang et al. 1995). Each pair of letters represents a change at the different sites in the two αβ interfaces, from one amino acid to the other. Changes occurred in different lineages across the myoglobin-hemoglobin phylogeny. Columns "Myoglobin" and "Hemoglobin" show the amino acids present in their respective ancestral nodes using human sequences as reference. In bold are the replacements classified as radical according to Zhang (2000). Normal fonts correspond to conservative replacements.

α andβ globin genes.- I looked at the sites that are part of the αβ interfaces and mapped the changes that had occurred between the ancestral nodes of α and β globin genes in order to detect coordinated changes at those sites in the two chains. Table 3.8 shows the replacements that have occurred at the interface sites. There are remarkably different patterns of replacement between the two interfaces. Interface α1β1 has experienced many more substitutions than interface α1β2, which has changes at only two sites. Also, at the α1β1 interface replacements are mainly radical while the opposite is true for the α1β2 interface. This result would indicate that residues at the α1β2 interface are considerably more constrained than at the α1β1 interface, which is in agreement with the structural requirements in the α1β2 interface for sliding properly. Presumably, the role played by the sliding surface is more sensitive to amino acid changes than the packing surface α1β1.

The program Plotcorr was used to determine correlated mutations between sites at the αβ interfaces in the α and β globin genes. The program did not find correlated mutations between sites at the interfaces but it did find a few possible coevolving pairs in the exterior part of the protein (data not shown).

Positively selected sites in α globin.- The positively selected sites found in α globin with models M2, M3 and M8 are: 22, 68, 78, 111, 115 and 131. I obtained the marginal and joint likelihood reconstructions of ancestral amino acids. Table 3.9 shows the observed and inferred data for each of the positively selected sites. The results in the column labeled "Data" correspond to the observed amino acids at a given site in each taxon comprised in the dataset. The column labeled

Table 3.8. Replacements at αβ Interface Sites in the β and the α Globin Genes

Dataset	Site	State at root	Replacements across the phylogeny
ß-globin			α1β1
	26	E	EQ
	51	A	AP PA AP AS AP AS PA AP
	55	M	ML
	101	E	EQ
	109	V	VM VM **VE**
	111	V	VL
	112	I	VI **CV** VI **IT**
	115	A	AG AS
	116	H	**HE** TS TS IV **HT TK KI** HR HR
	120	K	**KN** KH **RG KN** SG KR **KN NS**
	123	T	TS **TN** TS
	124	P	PA
	125	E	**QP** EN **EV** ED DE **QA AE** ED EQ **AE**
	128	A	AS
α-globin			α1β1
	20	A	AG GA
	34	A	**AI AL**
	103	H	**HG**
	104	C	ST SA **CS**
	111	A	**SV SN AC SC** SA
	113	L	**LH LH LH HL**
	115	A	AG **NS NS** AS **AN**
ß-globin			α1β2
	43	E	ED ED
	101	E	EQ
α-globin			α1β2
	39	T	TS
	101	L	**LF** LV

Note: amino acid replacements were obtained from the ancestral state reconstruction (Yang et al. 1995) for the sites that are part of the αβ interfaces in the α and β globin genes. In relevant cases, replacements were mapped to the branches in the trees where changes occurred (see Figures 3.3 and 3.4). The columns "α" and "β" show the amino acids present at their respective ancestral node using human sequences as reference. In bold are the replacements classified as radical and normal fonts correspond to conservative replacements following the classification by Zhang (2000).

Table 3. 9. Reconstruction of the ancestral states of the positively selected amino acids in α and β globins by the likelihood method

Site	Data	Reconstructions and their posterior probabilities (in parenthesis)		# changes
	α globin			
22	EEEDAAEEDEDDE	EEEEEEEAEEED	(0.9950)	4
68	SNKKKKLKNLNNN	SSKKKKKKNNNN	(0.3986)	5
78	GGGGGGGGSHNNN	GGGGGGGGHHNN	(0.5310)	3
111	SSSSCCVNCAAAA	SSSSSSSCAAAA	(0.9847)	5
115	AGSNNSNSAAAAA	AAAANNNNAAAA	(0.9692)	5
131	SSNSNNTNSSSSS	SSSSSSSNSSSS	(0.4981)	4
	β globin			
5	PPAAADSPPPAAAAAAPAAAAEAAA	PPPPAAAAAAAAAAAAAAAAAPP	(0.9828)	5
9	SNAEAASATTASAASACTAAAAAAA	SSSSAAAAAAAAAAAAAAAAAST	(0.2616)	9
12	TCTLTSTANNMLTTTTSNTTTTTTL	TTTTTTTTTTTLMTTTTTTTTTN	(0.0907)	10
13	AASGGCAAAASSSSGGKGSSGSSSS	AAAASSSSSSSSSSGSSSSSSSAA	(0.5314)	8

52	DSSDDSNADDSSSSSSSSSSSSDD	DDSSSSSSSSSSSSSSSSDDDDD	(0.9846)	4
69	GSSQDTAGGGTTTTTTTSTTTTDD	GGSSSSTTTTTTTTTTTTDDDGG	(0.4885)	6
76	ANHKKNSAAAKKKKKKKKKKKKKK	AAKKKKKKKKKKKKKKKKKKKAA	(0.2604)	5
87	TKQKQSKQQQKKKKKKKKHHQQHSS	TTKKKKKKKKKKKKKHHHQKKSQQ	(0.1287)	8
116	HHHRHHHNNNHHHHHHHHHYLHYRR	HHHHHHHTTTTTTTTKKKIRRRRR	(0.0814)	9

Note: this table shows the results obtained for the joint ancestral reconstruction of the individual α and β globin clusters. The results in the column labeled "Data" correspond to the observed amino acids at a given site in each taxon comprised in the respective alignment. The column labeled "Reconstruction" presents the inferred ancestral amino acid at that site in each of the taxa. The posterior probability of the reconstructed data is in parenthesis. The number of changes refers to the substitutions that occurred along the different lineages in the tree that underwent an amino acid replacement.

"Reconstruction" presents the inferred ancestral amino acid at each internal node of the phylogeny, with the posterior probability of the reconstructed data in parenthesis. The number of changes refers to the substitutions that occurred along the different lineages in the tree that underwent an amino acid replacement. All the reconstructions have their posterior probability in parenthesis. The highest posterior probability was assigned to the reconstruction of site 22 and the lowest occurred for site 68. The number of changes indicates the total number of amino acid replacements that occurred along different lineages in the tree. Some sites underwent more changes than others. In the case of α globin the highest number of replacements was five at site 68 and the lowest number was three at site 78. I mapped the replacements in the rooted phylogenetic tree for the α globin gene in order to determine the direction of the changes. Fig 3.3 shows the relevant replacements mapped onto the tree.

Positively selected sites in β globin.- For this analysis I used the β globin dataset included in Chapter II (Fig. 2.3). The positively selected sites identified in the β globin dataset with models M2, M3 and M8 are: 5, 9, 12, 13, 52, 69, 76, 87 and 116. These sites were reconstructed by marginal and joint likelihood reconstructions. Table 3.9 shows the results obtained for the joint ancestral reconstruction of β globin. The number of replacements varies from 10 in site 12 to 4 in site 52. The direction of the changes was established by mapping the replacements in the rooted phylogenetic tree for the β globin gene. Fig 3.4 shows the relevant replacements mapped onto the tree.

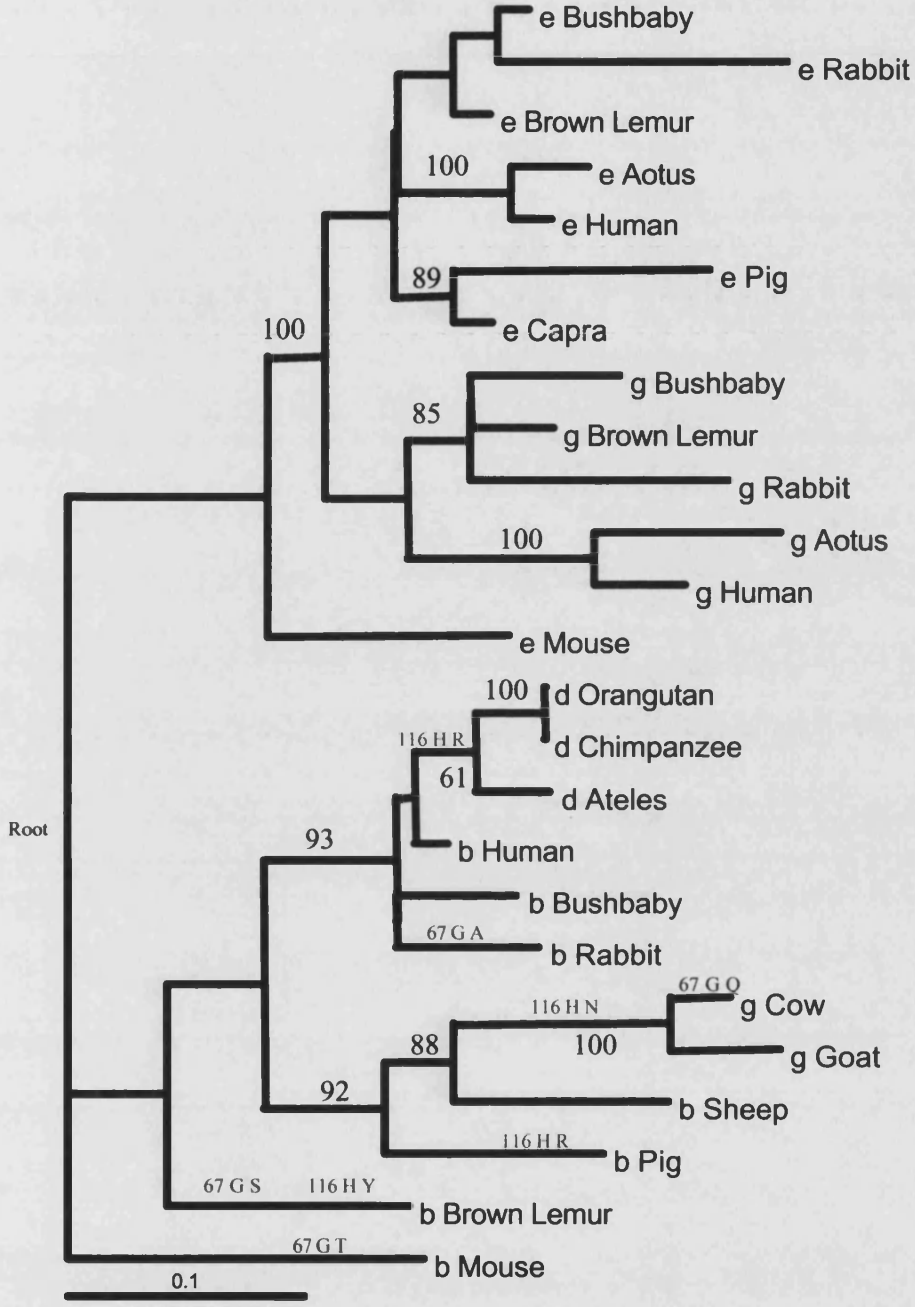


Fig. 3. 4. ML tree for the β globin cluster. Black numbers indicate bootstrap proportions. Relevant ancestral state reconstructions in blue are mapped onto branches indicating substitutions that occurred in those lineages. The number refers to the site in the protein involved; the letter at left is the ancestral amino acid at that site and the letter at right is the derived amino acid. Site 67 is located in the vicinity of the heme pocket and site 116 is part of the α1β1 interface.

In both the α and β globin genes one interesting case was noted, that of sites α111, α115 and β116. These sites appear in the respective reconstructions for α and β globins to have changed at the same branch in the artiodactyl lineage. Moreover, these sites all interact at the α1β1 interface in all three conformations of hemoglobin (Fermi et al. 1984, Shaanan 1983, Silva et al. 1992). This result may indicate a case of coevolution of interacting residues in different monomers.

Since all the positively selected sites in myoglobin-hemoglobin are located at the exterior of the protein chains and thus may not be very relevant in terms of the oligomerization potential or coordinated activity between monomers, I did not map the reconstructions in the phylogeny and only analyzed the physicochemical aspect of the replacements in the following section.

3.3.5 The structural and functional relevance of amino acid replacements

Myoglobin-hemoglobin.- In Table 3.10 we can see the comparison of the residues at the interface sites in the ancestral node of the myoglobin and hemoglobin clades. It can be seen that most amino acid replacements between the two ancestral nodes are radical. At the α1β1 interface, the replacements involved changes in physicochemical properties, such that in myoglobin most amino acids are either nonpolar or neutral and in hemoglobin they are polar, thus contributing to the formation of polar patches for building interfaces. Most replacements at the α1β1 interface imply a change in polarity. In the case of the α1β2 interface, most changes imply differences in the size of residues and to a lesser degree in their polarity or charge. From table 3.7 it is clear that sites at the different interfaces differ also in the number of radical replacements

they have undergone. At the $\alpha 1\beta 1$ interface, 40 replacements out of 87 are radical, as they change important physicochemical properties such as volume and polarity. On the other hand, at the $\alpha 1\beta 2$ interface, 28 out of 42 replacements are radical. Note that not all sites changed with a similar pattern. For instance, sites 35 and 100 experienced only radical replacements, whereas site 43 had more conservative changes. At both interfaces some sites experienced few radical changes relative to the total number of changes they underwent (Table 3.7). For example, site 109 had replacements that are mostly conservative, that is, it retained essentially unchanged physicochemical properties. On the other hand some sites have experienced replacements that are, in most cases if not always, radical as in the case of sites 120 or 35. The latter sites probably correspond to the 7% of sites identified by model M8 to be under relaxed selective pressure. A rough measure like the previous count fails to capture two important features of substitutions, which are when and where did changes occur. Mapping the replacements onto the estimated tree shows that in most cases replacements are evenly distributed across the phylogeny. This suggests no episodes of increased substitution, the only exception being at the ancestral branch of both myoglobin and hemoglobin. There, the number of changes is considerably elevated with respect to the rest of the branches in the tree. For the two $\alpha\beta$ interfaces, the ancestral branch registered the highest number of replacements anywhere in the tree. This indicates that following their divergence a large number of amino acid changing substitutions occurred at the sites that compose the interfaces.

Table 3.10. Comparison of amino acid changes at the αβ interfaces in Hemoglobin and Myoglobin

Site	Myoglobin	Hemoglobin	Relevance
α1β1			
26	G (1.00)	E (0.99)	radical
30	L (1.00	R (1.00)	radical
33	L (1.00)	V (1.00)	conservative
34	F (1.00)	V (1.00)	radical
35	K (1.00)	Y (1.00)	radical
51	E (1.00)	P (1.00)	radical
55	K (1.00)	M (0.99)	radical
101	Y (1.00)	E (1.00)	radical
108	C (0.48)	N (0.74)	radical
109	I (1.00)	V (0.98)	conservative
111	Q (1.00)	V (1.00)	radical
112	V (1.00)	I (0.97)	conservative
115	S (0.99)	A (1.00)	conservative
116	K (1.00)	R (0.78)	conservative
119	G (1.00)	G (1.00)	conservative
120	D (1.00)	K (1.00)	radical
125	A (1.00)	Q (1.00)	radical
128	A (1.00)	A (1.00)	conservative
131	G (1.00)	Q (0.99)	radical
α1β2			
34	F (1.00)	V (1.00)	radical
35	K (1.00)	Y (1.00)	radical
37	E (1.00)	W (1.00)	radical
39	L (1.00)	Q (0.98)	radical
40	E (1.00)	R (1.00)	radical
43	D (1.00)	E (0.84)	conservative
93	H (1.00)	K (0.99)	conservative
98	P (1.00)	V (1.00)	radical
99	V (1.00)	D (1.00)	radical
100	K (1.00)	P (1.00)	radical
101	Y (1.00)	E (1.00)	radical
102	L (1.00)	N (1.00)	radical
105	I (1.00)	L (0.70)	conservative

Note: this table shows the differences in the amino acids at sites in the αβ interfaces between myoglobin and hemoglobin. Columns "Myoglobin" and "Hemoglobin" show the different amino acids present in the ancestral nodes of myoglobin and hemoglobin clades using human sequences as reference. Posterior probabilities are shown in parenthesis. Classification of differences as radical or conservative was done following Zhang (2000).

Interface sites in α and β globins.- comparing the amino acids at the two αβ interfaces in both α and β globin genes, I found that there have been more replacements at the α1β1 interface than at the α1β2 interface, which presumably indicates that sites at the latter are more constrained than at the former. Also, replacements at the α1β2 interface were almost exclusively conservative, whereas both radical and conservative changes occurred at the α1β1 interface. Table 3.8 shows the comparison of replacements at α1β1 interface and at the α1β2 interface. Replacements of sites at the α1β1 in the β globin chain were conservative in the majority of cases (35/53). For example, site 51 underwent 8 changes, all of them conservative. Site 120 is exceptional in that most of the changes there were radical. On the contrary, replacements of sites at the α1β1 interface in the α chain were radical in the majority of cases (15/22). Sites 34, 103, and 113 experienced only radical changes. At the α1β2 interface, there were very few changes in both α and β chains and they were, with only one exception, conservative.

Positively selected sites in myoglobin and α and β hemoglobins.- Table 3.11 shows the results for the replacement analysis of the positively selected sites in all three globin genes. Replacements of the positively selected sites in β globin were in most cases conservative (38/70). In contrast, most replacements were radical in both α globin (24/30) and between myoglobin and hemoglobin (51/83) although the latter occurred in places that presumably do not greatly affect structure and function. As a final note, the α111, α115 and β116 residues that I hypothesize to have coevolved in the artiodactyl lineage evolved under positive selection and experienced radical changes, albeit in the case of β116, only 4 out of 9 changes were radical.

Table 3.11. Replacement Analysis of Positively Selected Sites in the β and α Globin Genes and in the Myoglobin-Hemoglobin Datasets

Dataset	Site	State at root	Replacements across the phylogeny
ß-globin			
	3	P	**PD** PS PA PA PA AP **AE AV**
	7	A	AS SA AT **AN** AT AS **AC** AS
	10	T	**TC TL** TS **TN**
	11	A	**AE AN** AS AS AS AS AS **AC** AT
	50	S	SA **SD SN AD**
	67	N	ND NQ **NT NS** SG **SN** SA
	74	N	**NK NH NA NS KN**
	85	K	**KQ KN KQ KT NS NH**
	112	T	TN TL LM MS TS TC TA TN TL LT
	116	H	**HT HE** HR TS HR TS **TK KI** IV
α-globin			
	22	E	ED ED ED **DA EA** DE
	68	N	**NK NK KL NS NL**
	78	G	**GN** GS **NH NG**
	111	C	**CS CS CN CV CA AC**
	115	A	**AN AN** AG **NS NS**
	131	S	**SN SN NS NT**
myoglobin-hemoglobin			
	8	V	VI **VA CH** AS **AN AE** AT **AN VC CF FC VA** AS **AQ** mostly radical
	16	Y	VI YF **YH EK** ED ED **QH** VL **VY YH YQ** QE
	17	A	AP **TN TD** AT VL **VT** AT **AV VC VD**
	45	D	**DA SD** SA **SN DG** SA DE **DS DT DS**
	62	A	SG **SN TQ TD TN AK** AT TS SA TS SG
	80	L	LM LM LM **KT KQ KQ HS GQ LK** KH **KQ KG**
	126	L	**VT VA** VI **NT NT** ST **AD** LV VI VI **LH HN NS HA**

Note: amino acid replacements were obtained from the ancestral state reconstruction for the positively selected sites in each dataset. In bold are the replacements classified as radical according to Zhang (2000). Normal fonts correspond to conservative replacements.

4.4 Discussion

Amino acid changes drive the evolution of proteins. Replacements in amino acid chains can be brought about by drift or by the action of positive selection; therefore, testing the evolutionary forces behind protein evolution is essential in order to distinguish between plausible scenarios. I investigated the role of positive selection in the evolution of myoglobin and the α and β globin genes and asked whether positively selected sites were associated with important changes in terms of structure and function. Results show that changes between myoglobin and hemoglobin are mostly neutral, with a small fraction of sites being slightly less constrained than the average. The positively selected sites detected by model M2 are mostly located at the exterior of the protein and even though the changes are radical in terms of their physicochemical characteristics, these are not expected to introduce important differences in terms of structure and function, as they are not associated with crucial sites at interfaces or heme pocket. Furthermore, the evidence for positive selection was not strong, as only model M2 found a few positively selected sites. One possibility is that the signal for positive selection driving the changes between myoglobin and hemoglobin is diluted when averaging across all the branches in the phylogeny. If this were the case one would expect to find greater ω ratios at a few lineages rather than across the whole tree. In the case of the α and β globin genes, results show that following their divergence, each gene was subject to different selective constraints, α globin being in general more constrained than β globin. Also, while most of the changes brought about by positive selection in β globin are conservative in terms of structure and function, the opposite is true for α globin,

where changes in sites at the interfaces may have contributed to the evolution of the interaction between hemoglobin chains.

To find out whether positive selection acted at a specific time during the divergence of myoglobin from hemoglobin and of α globin from β globin, I tested for positive selection among branches in the phylogeny. This analysis was necessary as results of positive selection across sites indicated that averaging across branches could result in the lack of power of the tests. It was interesting to find different results for the myoglobin-hemoglobin and the α–β globin datasets. In the first case, I found an increase in nonsynonymous substitutions along the ancestral branch of both myoglobin and hemoglobin although the LRTs performed were not significant. Also, the myoglobin lineage was on average evolving under less constraint than the hemoglobin lineage. Positive selection appears to have driven some of the changes between myoglobin and hemoglobin at the early stages of their divergence. Once the two genes diverged, though, the role of positive selection seems to have been less important in the further evolution of these two lineages, as we know from the analysis of the evolution of the vertebrate β globin gene cluster (Chapter II). In contrast, the analysis of the α-β globin dataset yielded no evidence for positive selection acting in the ancestral branch of α and β globin genes. Prior to their divergence, the ancestor of both α and β globins had only a slightly higher rate of nonsynonymous substitutions which may account for the functional differences between the two genes and once the α and β globin genes had diverged, purifying selection acted to maintain their respective function. The contrast with the divergence of myoglobin and hemoglobin is interesting. It is not surprising to find evidence for positive selection in the divergence of myoglobin and hemoglobin, as the changes involved are more dramatic, both in terms of structure and function, than those involved between α and β globin. In the

latter case, fewer changes mediate the differences in structure and function and these could appear without the episodic action of positive selection. I do not rule out the possibility that positive selection was involved in the divergence of α and β globin, what I am suggesting is that it could have acted at only a few sites at particular time points during this process. Analysis with site-branch models may help to find out whether this was indeed the case. On the other hand, for myoglobin and hemoglobin, it appears that prior to their divergence important changes occurred in a relatively short period of time that promoted the differences in structure and function between these duplicated genes and may explain the oligomerization potential realized in hemoglobin. This is consistent with previous findings by Goodman (1981).

The possibility that oligomeric proteins, such as hemoglobin, arose from a monomeric ancestor is intriguing and deserves more attention than it has been given. To investigate this hypothesis, it is necessary to compare sets of sequences that represent at least one monomer and one oligomer that are closely related. A difficulty immediately arises because we can compare and analyze only present-day sequences, whereas the ancestral monomer and the first oligomers very likely evolved long before any of the data available came to being. Since there are few or no alternatives at all, we have to take the risk of comparing extant sequences and try to infer past events. I thus decided to investigate the oligomerization process in globins by comparing monomeric myoglobin and oligomeric hemoglobin. Sites involved in the creation of interfaces are of special interest when testing the oligomerization of proteins (D'Alessio 1999). Intersubunit interfaces are characterized by a significant presence of polar and charged residues, which are fundamental in establishing hydrogen bonds and salt bridges among protein subunits (Creighton 1993). In order to determine whether amino acid replacements between myoglobin and hemoglobin

evidence a tendency to change from non-polar or neutral residues at sites corresponding to interfaces, I reconstructed the ancestral states of such residues and established the direction and relevance of changes. Results show that most of the amino acid changes between myoglobin and hemoglobin at the interfaces are radical, according to Zhang's classification by polarity and volume, which he found to be the most informative criteria to classify replacements (2000). Moreover, when comparing the states of ancestral myoglobin and hemoglobin nodes, it is evident that residues in myoglobin at the sites corresponding to would-be interfaces are more often non-polar and neutral, whereas interface sites in hemoglobin show a marked tendency to be polar or charged. These changes likely contributed to the creation of the necessary patches in the globin monomer that later served as interfaces. Interestingly, oligomerization was a requisite for the sophistication in function that we observe in hemoglobin relative to myoglobin.

Once interfaces evolved in oligomeric globins it is expected that in order to maintain the crucial site-site interactions occurring between subunits, amino acid replacements in one monomer would be followed by compensatory mutations in the other. This coevolution hypothesis can be tested by analyzing amino acid replacements at sites in the $\alpha\beta$ interfaces. In this case also I reconstructed the ancestral states in the two protein chains and determined the direction and relevance of replacements. Results indicate that most of the replacements at the interfaces are conservative, as expected given the importance of those residues in establishing the contact between subunits. There are, however, a few sites at the interfaces that were detected to be under positive selection. One remarkable example is given by sites 111 and 115 in α globin, and site 116 in β globin which are positively selected and interact in all three hemoglobin conformations (R, T, and T2). In order to establish

whether replacements at those sites have compensated each other I used the program Plotcorr, which detects compensatory mutations between pairs of residues in an alignment. Plotcorr did not find evidence for such correlated mutations between interface residues in α and β globins; however, in a previous paper, Pazos et al. (1997) did find support for correlated mutations at the α1β2 interface so it might be that the signal for correlated replacements for these sites is not strong enough to be detected or because even though sites 111α, 115α, and site 116β are part of the α1β1 interface they do not make actual physical contact and therefore no compensatory mutations would be needed. Crystallographic data is needed to clarify this possibility. At any rate, amino acid replacements between the α and β subunits seem to be restricted to conservative changes in physicochemical properties, particularly at interface α1β2.

The problems of investigating protein structure evolution from sequence data have to do with the fact that while three dimensional structure is highly conserved, sequences can diverge widely. This is true for globins, specially in the comparison of long diverged molecules, such as myoglobin and hemoglobin. Although the level of divergence observed among sequences was relatively high (around 40%), it was still possible to analyze their sequences using codon models. Alignments were relatively easy to obtain although several gaps had to be added in order to have a good match. The level of sequence divergence among the different globin genes studied in this chapter probably affected the results by obscuring the evidence for positive selection, which may have acted for short intervals of time at particular lineages. Also, given the long divergence times between α and β globin genes, I expect it will be rather difficult to find evidence for correlated mutations at a particular set of sites. Interface sites differ in α and β globins, therefore, ancestral state reconstruction had to be

obtained on their individual datasets, thus complicating the analysis of convergent evolution in artiodactyls, as sampling was not strictly identical. In the case of myoglobin, the reconstruction of ancestral sites and replacement analysis were done under the assumption that interface sites corresponded to those in hemoglobin, an assumption that may be unwarranted. Furthermore, the quality of the alignment greatly influences the results. In this case I consider the alignment is not a problem (Fig. 3.5). However contentious, I think the methodological approach exposed in this chapter is useful in detecting general patterns of selective pressure variation among sites and lineages and how these are related to relevant changes in the function and structure of proteins.

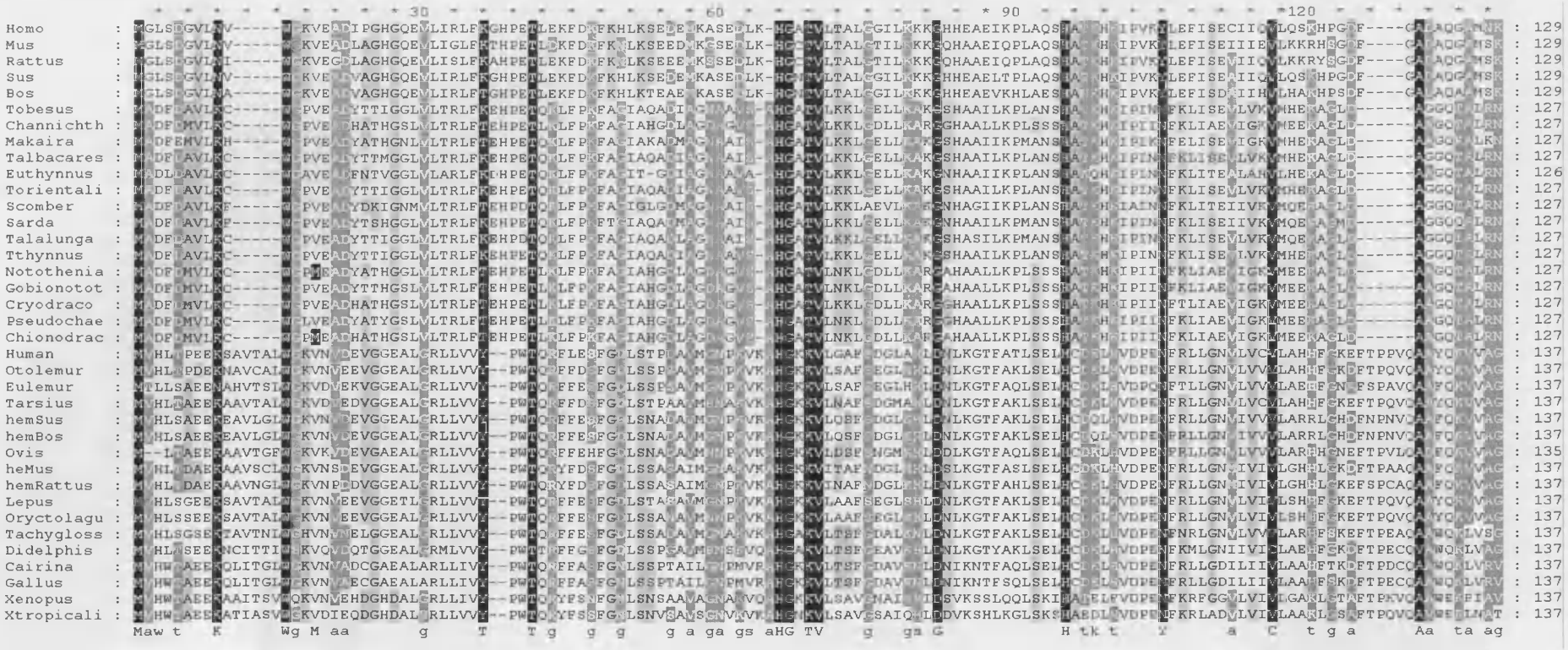
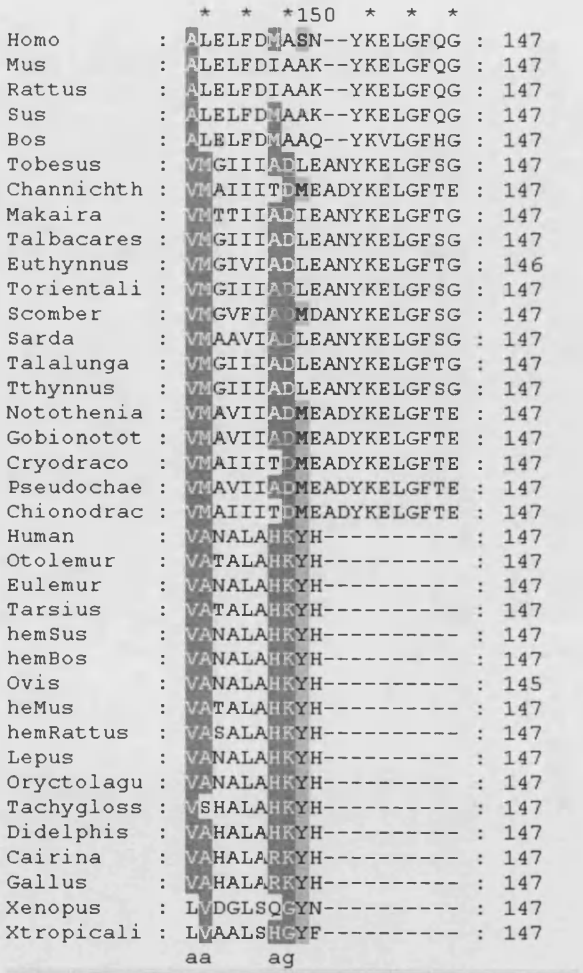


Fig.3.5. Amino acid alignment of the myoglobin and β globin sequences in the "myoglobin-hemoglobin" dataset. Darker shades correspond to greater similarity among sequences. Here gaps are shown but sites with gaps were excluded from the analysis.


* * *150 * * *
Homo : ALELFDMASN--YKELGFQG : 147
Mus : ALELFDIAAK--YKELGFQG : 147
Rattus : ALELFDIAAK--YKELGFQG : 147
Sus : ALELFDMAAK--YKELGFQG : 147
Bos : ALELFDMAAQ--YKVLGFHG : 147
Tobesus : VMGIIIADLEANYKELGFSG : 147
Channichth : VMAIIITDMEADYKELGFTE : 147
Makaira : VMTTIIADIEANYKELGFTG : 147
Talbacares : VMGIIIADLEANYKELGFSG : 147
Euthynnus : VMGIVIADLEANYKELGFTG : 146
Torientali : VMGIIIADLEANYKELGFSG : 147
Scomber : VMGVFIADMDANYKELGFSG : 147
Sarda : VMAAVIADLEANYKELGFSG : 147
Talalunga : VMGIIIADLEANYKELGFTG : 147
Tthynnus : VMGIIIADLEANYKELGFSG : 147
Notothenia : VMAVIIADMEADYKELGFTE : 147
Gobionotot : VMAVIIADMEADYKELGFTE : 147
Cryodraco : VMAIIITDMEADYKELGFTE : 147
Pseudochae : VMAVIIADMEADYKELGFTE : 147
Chionodrac : VMAIIITDMEADYKELGFTE : 147
Human : VANALAHKYH---------- : 147
Otolemur : VATALAHKYH---------- : 147
Eulemur : VANALAHKYH---------- : 147
Tarsius : VATALAHKYH---------- : 147
hemSus : VANALAHKYH---------- : 147
hemBos : VANALAHKYH---------- : 147
Ovis : VANALAHKYH---------- : 145
heMus : VATALAHKYH---------- : 147
hemRattus : VASALAHKYH---------- : 147
Lepus : VANALAHKYH---------- : 147
Oryctolagu : VANALAHKYH---------- : 147
Tachygloss : VSHALAHKYH---------- : 147
Didelphis : VAHALAHKYH---------- : 147
Cairina : VAHALARKYH---------- : 147
Gallus : VAHALARKYH---------- : 147
Xenopus : LVDGLSQGYN---------- : 147
Xtropicali : LVAALSHGYF---------- : 147
aa ag

Chapter IV

-

Dating gene duplications and species divergences in the vertebrate β globin gene family

4.1 Motivation

Since Zuckerkandl and Pauling proposed the concept of a molecular clock (1962, 1965), globins have been cited as examples of genes evolving with an approximately constant rate through time (Li 1997). Estimations of gene divergence dates based on molecular data often assume the molecular clock. Under this hypothesis, it is expected that substitutions accumulate at uniform rates along the branches in a phylogeny (Kimura 1983). However, this expectation is often violated. Also, divergence time estimation is known to be highly sensitive to assumptions about the evolutionary rate (Takezaki, Rzhetsky and Nei 1995, Rambaut and Bromham 1998, Yoder and Yang 2000, Aris-Brosou and Yang 2003).

New approaches based on maximum likelihood and Bayesian methods have been developed which allow the evolutionary rate to change among lineages when divergence times are estimated (Sanderson 1997; Rambaut & Bromham 1998, Thorne, Kishino and Painter 1998; Huelsenbeck, Larget and Swofford 2000; Yoder and Yang 2000; Thorne and Kishino 2002). Other interesting features include the possibility of using multiple calibration constraints based on fossil data (Rambaut & Bromham 1998, Thorne, Kishino and Painter 1998; Thorne and Kishino 2002), and using multiple genes or gene partitions. These approaches make a more efficient use of the information contained in sequences and account for differences in evolutionary processes among lineages.

In this chapter, I used recently developed methods based on maximum likelihood (ML) and Bayesian analysis to estimate dates for gene duplication and gene conversion events in the vertebrate β globin gene family. I compared dates inferred

assuming the molecular clock with dates obtained without this constraint. Also, I tested different models that allow variable rates among phylogenetic branches to investigate evolutionary rate variation among genes and vertebrate species.

4.2 Theory and Methods

4.2.1 Sequence data and phylogenetic analysis

62 vertebrate β globin nucleotide sequences were retrieved from GenBank. Genes from amphibians, birds and mammals were included, with fish used as outgroups. The species names and GenBank accession numbers for the ingroup are identified in the tree of Figure 4.1. The outgroup fish sequences include: Zebrafish βA1 (U50382), βA2 (U50379) and βE1 (AF082662); carp β globin (AB004740); salmon β globin (Y08923) and *Oryzias* ε globin (AB080118). Alignment was conducted using Clustal X (Thompson et al. 1994), followed by minor manual adjustments. Gaps were removed.

Maximum likelihood and Bayesian methods were used to estimate a phylogeny for the vertebrate β globin sequences. PAUP* (Swofford 2000) was used to conduct the ML analysis, with the model of nucleotide substitution being HKY85. A minimum evolution tree was used as a starting topology from which model parameters were optimised. Support for the branches was assessed with a bootstrap analysis with 100 pseudoreplicates. For the Bayesian analysis I used MrBayes (Huelsenbeck and Ronquist 2001). Substitution rates were assumed to be gamma distributed and base frequencies were estimated from the data. The Markov Chain Mote Carlo (MCMC) algorithm was run with three chains for ten million generations and sampled every 100 generations. I repeated this process three times to check

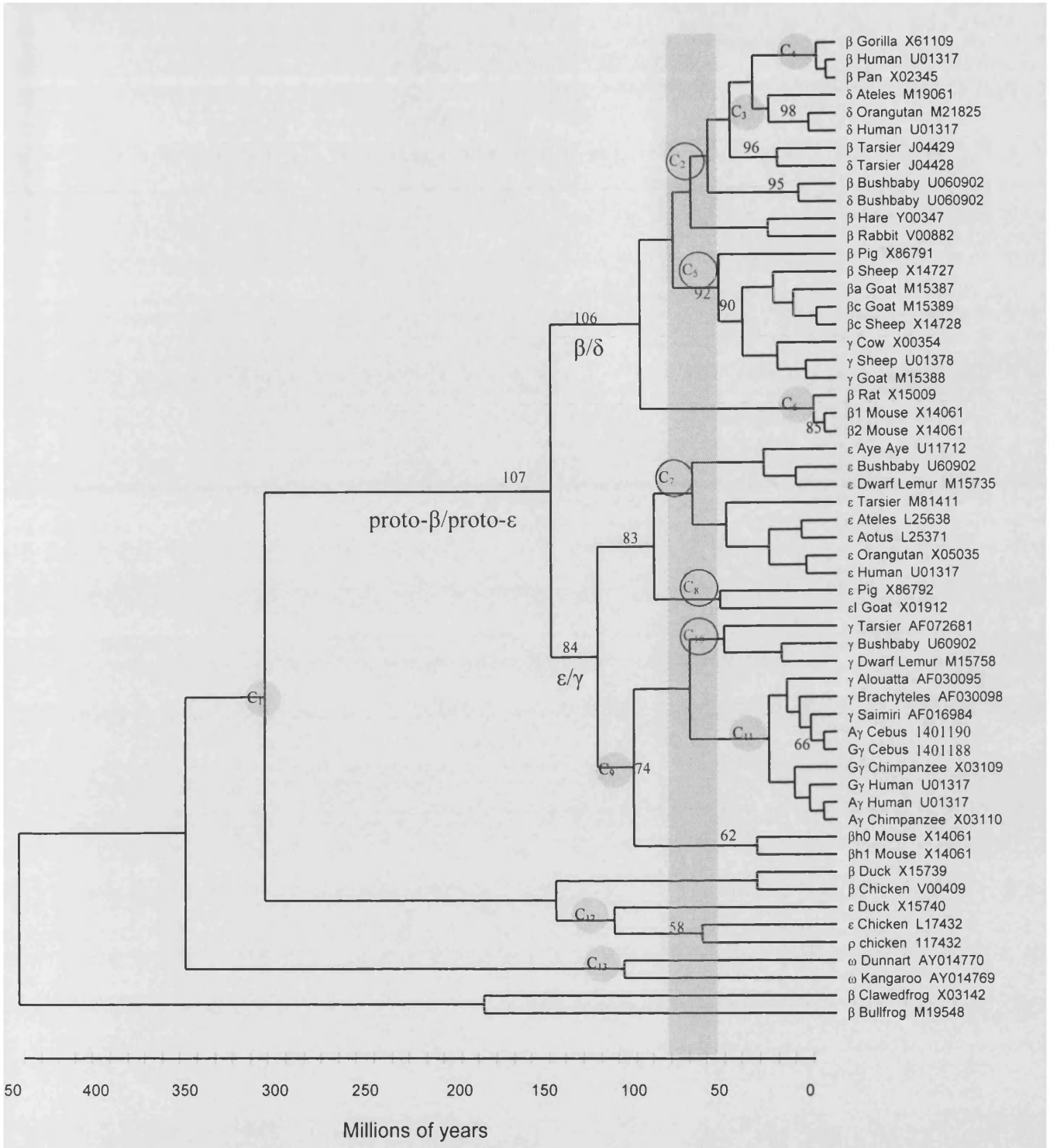


Fig. 4.1. Maximum likelihood phylogeny of the rooted ingroup tree for the β globin gene family. Species names and GenBank accession numbers are listed next to each sequence. Circles represent calibration nodes. Nodes of interest are numbered according to Jeff Thorne's multidivtime node assignment. The grey bar indicates the proposed time for the eutherian mammal radiation.

for convergence. Stationarity was reached after sampling 2000 trees; trees sampled before stationarity were excluded from the analysis.

4.2.2 Divergence date estimation

Divergence date estimation based on molecular analysis relies on fossil data used as calibration points. In this study, calibration points are available for 15 ancestral nodes distributed across the phylogeny (Fig. 4.1) and are listed in Table 4.1. The calibration dates used in this chapter come from fossil data of the relevant species divergences: human and gorilla (Shoshani et al. 1996), monkeys and apes (Shoshani et al. 1996), mice and rats (Jacobs and Downs 1994), birds and mammals (Benton 1990, 1999), chickens and ducks (Cracraft 2001), bony fishes and amphibians (Bromham et al. 1998); the oldest monotremes and marsupials (Luo et al. 2001; Cifelli 2000), the eutherian radiation (Goodman et al. 1987b), the basal radiation of primates (Gingerich and Uhren 1994; Martin 1993; Tavaré et al. 2002), and the Cetartiodactyla radiation (Bowen et al. 2002).

Likelihood Models of Global and Local Clocks- I analysed the three codon positions separately, and all codon positions simultaneously either with or without accounting for their differences in the substitution process (Yang and Yoder 2003). I used the ML tree for date estimation and estimated divergence dates under both the global-clock (i.e. enforcing a uniform rate across the phylogeny) and the local-clock (i.e. allowing for rate variation among sets of branches in the tree) models. Recently developed methods permit the use of multiple calibration points allowing the use of different fossil dates distributed across the phylogeny (Yang and Yoder 2003). The likelihood

implementation allows only fixed node ages, whereas the Bayesian approach allows a range to be used. In order to have comparable calibration points, in the likelihood analysis I used the mid-value of the range used with the Bayesian method. Parameters in the models include the substitution rate (μ) and the ages of nodes which are not calibration points. The nucleotide substitution models were JC (Jukes and Cantor 1969) and F84G (Felsenstein 2000; Yang 1994). The baseml program in the PAML package was used (Yang 1997; Yoder and Yang 2000; Yang and Yoder 2003).

Likelihood date estimation assuming different rate classes

The local-clock model allows rate variation among different sets of branches. I tested whether the paralogous genes in the β globin family were evolving under significantly different rates, to do so, four rates were specified for each gene clade in the phylogeny (i.e., β, γ, ε, and all other branches). I conducted a likelihood ratio test (LRT) to compare the local clock model (i.e., each gene clade was allowed to have a different rate) with the global clock model (i.e., all gene clades had the same rate) (Felsenstein 1981, Yoder and Yang 2000).

Bayes Method for Divergence Date Estimation- I used a Bayes MCMC approach as implemented in a new version of the program package written by Jeff Thorne (Thorne and Kishino 2002; see ftp://abacus.gene.ucl.ac.uk/pub/T3/). The program estbranches was used to produce the ML estimates of branch lengths for the rooted ingroup tree

Table 4.1. Calibration dates for ancestral nodes in Figure 4.1 (in millions of years)

Node	Range	Mid-Value
C_1 birds/mammals	288-310	299
C_2 primate basal radiation	63-90	77
C_3 monkeys/apes	32-38	35
C_4 gorilla/human	8-12	10
C_5 artiodactyl radiation	55-65	60
C_6 rat/mouse	11-13	12
C_7 primate basal radiation	63-90	77
C_8 artiodactyl radiation	55-65	60
C_9 eutherian radiation	51-120	85
C_{10} primate basal radiation	63-90	77
C_{11} monkey/ape	32-38	35
C_{12} duck/chicken	80-100	90
C_{13} stem marsupial	90-130	110

and the variance-covariance matrix. Fish sequences were used as outgroups, which are required by the program to locate the root in the ingroup tree. In this case, the model of nucleotide evolution assumed was F84G. The transition/transversion rate ratio and the shape parameter of the gamma distribution were obtained using PAML (Yang 1997). The output of estbranches was used in the program multidivtime in order to estimate divergence dates. In this analysis I used all codon positions accounting for the differences in rate among them. Also, this method does not require the use of fixed calibration points so I specified lower and upper bounds for the ancestral nodes. The MCMC chain was run at least twice for 100,000 generations after a burn-in of 10,000 generations. I sampled the chain every 10 generations. Multidivtime requires the specification of some priors. I assigned: (*i*) a gamma prior for the time of the root with mean 400 MY; and standard deviation 200 MY; (*ii*) a gamma prior for the rate at the root with mean 0.23; and standard deviation 0.12. Also specified were a gamma prior and the correspondent standard deviation for the parameter ν, which controls how variable rates are over time. The chosen value was 0.4 for both the ν parameter and the standard deviation.

4.2.3 Test of evolutionary rate variation

Local-clock models provide a good approach to test for evolutionary rate correlation between gene duplicates. I constructed several tests to examine whether, following duplication, different β globin gene duplicates evolved with similar rates. In this chapter I refer to sister genes that arose via a gene duplication event as a "pair". I compared the rate of a pair of genes that originated through gene duplication with the rate of the rest of the genes in the tree. I assigned one rate class to the tested pair of

duplicate genes and one rate class to the rest of the genes (two-ratios test). I conducted these tests for the following pairs of duplicate genes: (1) β and δ globin from tarsier; (2) β and δ globin from bushbaby; (3) βh0-βh1 from mouse; (4) β1-β2 from mouse; (5) G_γ-A_γ from *Cebus*; (6) G_γ-A_γ from chimpanzee; (7) G_γ-A_γ from human; and (8) βA1-βA2 from zebrafish. I compared each case with the global clock model using LRTs.

It is expected that through time, duplicated genes in a family will diverge in function and selective pressure. These changes would be reflected in the rate of evolution, with each gene evolving at a different rate. I used the same eight pairs of genes as described above and tested whether the two sequences in the duplicate pair evolved under a different rate from each other and from the average in the phylogeny. I assigned one rate class to each of the duplicate genes in the tested pair and one rate class to the rest of the genes in the tree (three-ratios model). I constructed the LRTs to compare the null (two-ratios) model with the alternative (three-ratios) model.

Lineage-specific effects, such as generation time, metabolic rate, population size, speciation rate, etc, are often invoked to account for the differences in rates among evolutionary lineages (Kimura 1983; Muse and Gaut 1997). I was interested in investigating the possible influence of lineage-specific effects in the variation of evolutionary rates among the sampled genes. I constructed several LRTs to compare the global clock model with a local clock alternative that assigns a particular rate class to each of the tested lineages. In each test I compared two lineages, assigning a specified rate class to each of the two and compared those models with the general clock model. If lineage-specific effects apply to my dataset the rejection of the null model is expected. I was particularly interested in testing lineage-specific effects on

the rates along the primate and rodent lineages. I used different sets of the sequences shown in Figure 4.1 to test each case. The sets of sequences used were: (1) primate and non-primate β globin genes; (2) primate and non-primate γ globin genes; and (3) rodent and non-rodent β globin genes.

4.3 Results

4.3.1 Phylogenetic Analysis

Maximum likelihood and Bayesian analyses resulted in similar topologies (Fig. 4.1). Most branches were supported by high bootstrap proportions (> 90%) and posterior probabilities (> 90%) (data not shown). In general, the gene tree is in agreement with the expected species tree except where gene conversion has affected the topology (e.g., tarsier β and δ globins appear as sister sequences) and where there has been gene duplication within some clusters (e.g., in the mouse and artiodactyl β globin clades). Besides dating gene duplications I was interested in dating gene conversion events, therefore, the conflicts in topology between gene and the expected species trees caused by gene conversion events are not a problem here. Most paralogous genes are grouped in different clades, the exceptions being genes affected by gene conversion and genes that are the result of within-cluster duplications. Gene conversion events and within-cluster duplications have been previously identified for the vertebrate β globin gene family (see Chapter II). Sequences affected by gene conversion include: tarsier β and δ globins; bushbaby β and δ globins; goat βa and βc globins; artiodactyl β and γ globins; mouse β1and β2 globins; human and chimpanzee A_γ and G_γ globins; and mouse βh0 and βh1 globins. All these sequences appear as

sister taxa instead of clustering with their true sister genes (i.e., δ tarsier appears sister to β tarsier instead of being located with the δ globins). In the case of sequences duplicated within clusters (e.g., γ cow, γ sheep, γ goat, and mouse βh0 and βh1), functionally different paralogous genes may still group with the gene from which they originated, as in the case of artiodactyl γ globins which originated from artiodactyl β globins and appear in the same clade. In the phylogeny, the most basal taxa are frog globins, followed by marsupial ω globins and bird globins. Marsupial ω globin is thought to be a relic gene that was lost in the eutherian lineage (Wheeler et al. 2001); thus, in accord with that hypothesis, it appears more closely related to bird sequences that to other mammalian globin genes. The eutherian mammal sequences are monophyletic, with the first duplication in that group being that of proto-β and proto-ε. Following this duplication, a duplication in the proto-β clade generated the β and δ lineages (node 98 in Fig. 4.1). In turn, the proto-ε lineage was divided by a duplication event, which yielded the ancestors of the ε and γ clades (node 84 in Fig. 4.1).

4.3.2 Estimation of Dates for Species Divergences and Gene Duplications

The dates obtained under the ML global clock model are listed in Table 4.2 in columns a)–d). The mid-values used for fixed calibration points are listed in Table 4.1. Dates for nodes other than those fixed as calibration points were estimated by ML, as were other parameters in the model. I either ignored the differences among the codon positions or accounted for them following Yang and Yoder (2003). When the three codon positions were analysed separately, date estimations did not differ greatly among codon positions (data not shown). I therefore decided to work exclusively on

the dataset with all codon positions with and without accounting for differences in the substitution process. The estimates listed in Table 4.2 under the label "all codon positions" are estimates averaged over the three codon positions, whereas the estimates under the label "combined" account for the differences in rate among them. Date estimates for the ancestral nodes of interest show the variation between the JC and F84G models, with the largest difference being about 17 MY at nodes 106 (99.5 - 117.7) and 107 (147.5 - 167.7) (Table 4.2). All dates are within the range of previously reported dates for gene duplication events although some speciation times seem somewhat young; for instance, my estimate for the G_γ/A_γ duplication is between 72 and 106 MY, when a previously proposed date places this duplication at around 35 MYA (Hardison and Margot 1984; Goodman et al. 1984).

The LRTs for the clock assumption under both the JC and the F84G models rejected uniform rates across the phylogeny. I estimated divergence dates assuming different rate classes for each of the three-lineages corresponding to the β, ε, and γ globin genes, and the rest of the phylogeny (i.e., r_1 for β, r_2 for γ, r_3 for ε globin, and r_4 for the remaining branches). ML estimates of duplication dates under the local clock model with four rate classes are listed in Table 4.2. Results in columns (e) and (f) correspond to the combined analysis of all codon positions accounting for their differences and assuming different rates for the specified rate classes. Lists in columns (g) and (h) in Table 4.2 correspond to the analysis of all codon positions assuming different rates for the indicated rate classes but without accounting for the differences in substitution among codon positions. Results are remarkably similar among the different local-clock models used, being within 3MY. Date estimates are more variable between the global- and local-clock models.

Table 4.2. ML and Bayes estimates of duplication dates with and without assuming the clock and comparing different models

	Maximum likelihood analysis								Bayesian analysis			
	global clock				local clock							
	all cod. pos.		combined		combined				no clock		clock	
Node	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)		(j)	
	JC	F84G	JC	F84G	JC	F84G	JC*	F84G*	F84G		F84G	
58 chicken ε/ ρ	50.5	49.4	50.8	49.4	56.4	55.1	56.4	55.1	59.3	(32.1, 87.8)	59.6	(33.5, 86.6)
62 mouse βh0/ βh1	26.7	25.0	26.4	23.9	23.2	21.5	23.2	22.2	29.0	(12.0, 50.3)	29.0	(11.5, 49.8)
66 Cebus A_γ/ G_γ	1.1	0.9	1.0	0.9	0.8	0.7	0.8	0.7	3.4	(0.1, 9.8)	3.0	(0.0, 9.4)
74 γ clade	85.0	85.0	85.0	85.0	85.0	85.0	85.0	85.0	105.6	(84.7, 119.2)	105.7	(85.3, 119.3)
83 ε clade	95.4	91.8	94.9	89.4	93.6	92.0	93.8	94.2	96.9	(78.4, 116.7)	96.7	(78.6, 116.0)
84 eutherian mammal ε/ γ	115.8	113.7	114.9	114.7	107.6	111.7	107.7	109.8	125.7	(103.9, 148.8)	125.9	(104.1, 149.0)
85 mouse β1/ β2	5.8	5.8	5.8	6.0	5.6	5.8	5.6	5.6	6.0	(0.6, 11.6)	6.0	(0.5, 11.6)
90 goat βA/ βC	32.9	31.6	32.6	31.3	28.9	27.8	28.9	28.2	20.6	(7.8, 35.2)	20.9	(7.8, 35.9)

92 artiodactyl β/ γ	51.4	51.7	51.3	51.8	48.2	48.8	48.1	48.9	41.4	(26.6, 57.0)	41.9	(26.4, 57.3)
95 bushbaby β-δ	8.7	7.5	8.4	7.4	6.8	6.0	6.8	6.1	14.8	(2.7, 29.1)	14.8	(2.6, 29.0)
96 tarsier β-δ	30.2	28.5	29.7	27.8	25.1	23.6	25.1	24.1	26.7	(12.5, 42.5)	26.81	(12.1, 42.6)
98 δ clade	35.0	35.0	35.0	35.0	35.0	35.0	35.0	35.0	34.8	(32.1, 37.8)	34.8	(32.1, 37.8)
106 β clade	117.7	112.5	115.8	113.5	100.5	100.4	100.4	99.5	102.4	(81.0, 129.6)	102.6	(81.4, 130.1)
107 eutherian β-globins	167.7	162.7	166.1	163.9	147.5	149.7	147.7	148.9	151.5	(123.9, 182.3)	151.7	(124.2, 182.9)

Note: The estimates in columns (a) and (b) are averaged over the three codon positions. The estimates in columns (c) to (f) are obtained from the three codon positions combined and accounting for the differences in rate among them. The estimates marked with * in columns (g) and (h) are obtained from the three codon positions combined but without accounting for their differences in rate. ML estimates are presented i columns (a) to (h) and Bayes estimates are in columns (i) and (j).

For instance, the estimate obtained for the ε and ρ divergence was 50.5 MYA under the global clock, whereas the local clock model resulted in 56.4 MYA (compare columns (a) and (e) in Table 4.2). The global clock estimate for the split between ε and γ was 115.8 MYA and the local clock model estimate was 107.6 (compare columns (a) and (e) in Table 4.2). However, all estimates are within the expected range according to previous reports and fossil data of the relevant species. The most important improvement in the likelihood scores is obtained by using the F84G substitution model with the combined analysis, column (f) in Table 4.2. The likelihood score for that column is -7431.1178 compared to -8073.5133 for column (a). The rates for each partition are listed in Table 4.3. Higher rates occur in the branches that follow gene duplication events. The β and γ globin genes evolve with a similar rate, with the β gene rate being slightly higher than that of γ globin in most cases. ε globin gene evolves with the slowest rate.

The Bayes method (Thorne, Kishino and Painter 1998) was applied to the β globin gene family data to estimate divergence dates. Posterior means of divergence times for ancestral nodes are listed in Table 4.2 columns (i) and (j). I analysed the data using all codon positions, accounting for differences in their substitution parameters, both with and without the clock assumption. The substitution model assumed was F84G, and the bounds listed in Table 4.1 are used for fossil calibration. First I estimated branch lengths and the variance-covariance matrices for the data at each codon position and treated these as gene partitions. The outputs for the three codon positions were used in multidivtime to estimate divergence dates. The prior mean rate at the root is the ML estimate under the global clock for all codon positions (Table 4.3). The dates obtained with and without the clock assumption do not vary

greatly (Table 4.2). Differences are higher between the ML and Bayes methods than between models using the same method. Apart from the inherent differences between the ML and Bayes approach, dates vary among methods as a result of using fixed and bounded calibration points. Bayesian date estimates for the major duplication nodes are largely in agreement with ML estimates (Table 4.2). The largest differences occur at four nodes: (i) node 66 is younger in the ML estimates by approximately 3 MY; (ii) node 95 is younger in ML dates by nearly 9 MY; (iii) node 90 is older in ML results by around 12 MY; and (iv) node 92 is older in ML estimates by around 10 MY (Table 4.2).

3.3.3Test for Rate Variation among Taxa

To test whether pairs of newly duplicated β globin genes are subject to different constraints, and therefore evolve with different rates from the rest of the phylogeny, I assigned a single rate to different pairs of newly duplicated genes (two-ratios model) and compared that model with the global clock model (one-ratio model). The duplicate pairs tested comprised genes which form part of a tandemly duplicated gene cluster or have been involved in gene conversion events. Bonferroni corrections were calculated to account for multiple tests. Of the eight duplicate gene pairs tested only two pairs of genes show a significant P-value: mouse βh0 and βh1 ($2\delta = 21.44$, df = 1, P-value < 0.0008); and mouse β1 and β2 ($2\delta = 42.25$, df = 1, P-value < 0.0008). Other results were: bushbaby β and δ globins ($2\delta = 7.39$, df = 1, P-value = 0.08); tarsier β and δ globin ($2\delta = 0.60$, df = 1, P-value = 3.52); *Cebus* G_γ-A_γ ($2\delta = 0.13$, df = 1, P-value = 5.76); chimpanzee

Table 4.3. ML Estimates of substitution rates for the four branch classes (x 10^{-8} substitutions per site per year)

Model	all codon positions		combined	
	JC	F84G	JC	F84G
Global clock	1.656	1.666	1.649	1.705 1st
			0.047	0.055 2nd
			0.171	0.186 3rd
Local clock	2.000	1.953	1.987	1.874 BG
	0.128	0.140	0.107	0.118 β
	0.083	0.086	0.057	0.059 ε
	0.137	0.144	0.093	0.098 γ
			0.034	0.039 BG
			0.071	0.077 β
			0.021	0.021 ε
			0.081	0.087 γ
			0.124	0.162 BG
			0.213	0.247 β
			0.176	0.189 ε
			0.242	0.254 γ

Table 4.3 [*] [a] Letters at the far right column indicate the rate for each corresponding codon position (*) or gene (BG, β, ε, γ). BG corresponds to background, used to designate the rate in the rest of the phylogeny. 1st, 2nd, and 3rd correspond to the three codon positions, respectively.

G_γ-A_γ ($2\delta = 0.01$, df = 1, P-value = 7.36); human G_γ-A_γ ($2\delta = 2.99$, df = 1, P-value = 0.64); and zebrafish βA1-βA2 ($2\delta = 0.54$, df = 1, P-value = 3.68). This suggests that rate differences among paralogs may not occur at all stages of the β globin gene family evolution.

To investigate whether the newly duplicated gene pairs in the β globin family differ in rate from one another as they begin to diverge in function, I assigned one rate class to each of the duplicate genes in the tested pair and one rate class to the rest of the genes in the tree (three-ratios model). I constructed the LRTs to compare the null (two-ratios) model with the alternative (three-ratios) model. Bonferroni corrections were calculated. In this case, only one LRT was significant: mouse β1 and β2 ($2\delta = 17.66$, df = 1, P-value = 0.008). Other results are: bushbaby β and δ globin ($2\delta = 6.63$, df = 1, P-value = 0.08); tarsier β and δ globin ($2\delta = 0.01$, df = 1, P-value = 7.36); mouse βh0 and βh1 ($2\delta = 0.00$, df = 1, P-value = 8); chimpanzee G_γ-A_γ ($2\delta = 4.80$, df = 1, P-value = 1.92); *Cebus* G_γ-A_γ ($2\delta = 0.61$, df = 1, P-value = 3.44); human G_γ-A_γ ($2\delta = 3.53$, df = 1, P-value = 0.48); and zebrafish βA1-βA2 ($2\delta = 0.29$, df = 1, P-value = 4.72). From these results it can be said that sequences in most duplicate gene pairs seem to evolve with similar rates between them.

To test whether lineage specific effects had influenced the rates of the genes in the phylogeny, I assigned each lineage of interest a different rate class from that of the rest (two-ratios model) and compared that with the global clock model (one-ratio model). The lineages tested included primate and rodent globin genes. Bonferroni

corrections were included. In all three tests genes in different lineages appeared to be evolving with a different rate from the rest of the branches in the tree and all LRTs are significant: primate β globins vs. non-primate β globins ($2\delta = 14.02$, df = 1, P-value = 0.0006); primate γ globins vs. non-primate γ globins ($2\delta = 21.44$, df = 1, P-value < 0.0003); rodent β globins vs. non-rodent β globins ($2\delta = 27.06$, df = 1, P-value < 0.0003). Rodents showed higher rates than mammals and other orders. There appears to be a lineage-specific effect on evolutionary rates acting on the primate and rodent lineages.

4.4 Discussion

Most dates estimated in this chapter for β globin genes are in fairly good agreement with previously reported dates, which were estimated assuming the clock: for the divergence between proto-β and proto-ε, I found dates around 150 MYA (node 107 in Table 4.2, see Efstratiadis et al. 1980; Czelusniak et al. 1982); the date I estimated for the β and δ globin divergence is between 100 and 118 MYA (node 106 in Table 4.2, see Goodman et al. 1984; Hardison and Margot 1984); for the ancestral node corresponding to the ε and γ globin divergence, I found values around 120 MYA (node 84 in Table 4.2, see Li 1997). The only discrepancy between my estimates and a previously published one occurred in the case of the duplication giving rise to A_γ and G_γ globins. Hayasaka et al. (1992), based on the fossil evidence from the divergence of the simian lineage (Anthropoidea) from the prosimians, proposed that the approximate date for the A_γ and G_γ globin divergence must be around 35 MYA. In contrast, my results give a much younger estimate for the *Cebus* A_γ and G_γ ancestral node of 3.4 MYA (node 66 in Table 4.2). I believe this result is likely affected by

gene conversion making my estimate younger than expected. As far as I know, this is the first time that dates for the ε and ρ globin split (49.4 – 59.6 MYA) are reported. This date is in agreement with bird speciation dates obtained with fossil and molecular data (Benton 1999). I dated two within-cluster duplications in the mouse: for the βh0-βh1gene divergence I found a range between 21 – 29 MYA; and for the mouse β1-β2 split a date around 6.0 MYA. These two dates would imply that the embryonic γ genes diverged well before the adult β globin genes, which is in agreement with studies of the evolution of the mammalian globin cluster (Hardies et al. 1984). I found that the goat βA-βC globin duplication occurred between 21 and 33 MYA; and that the artiodactyl β and γ globin divergence took place within 41- 52 MYA, these dates agree with fossil data for the artiodactyl origin and expansion (Bowen et al. 2002). Also, I provide an approximate date for the gene conversion events between β and δ globins in bushbaby, between 6 and 15 MYA, and tarsier, around 24 to 30 MYA. It is well known that these genes have experienced recurrent gene conversion (Drouin et al. 1999), even though this fact may have made the estimates younger, they agree with fossil and molecular speciation dates for tarsiers and bushbabies (Tavaré et al. 2002; Martin 2003). Another problem caused by gene conversion, and recombination in general, is that it may lead to the false rejection of the clock (Posada et al. 2002). To avoid falsely rejecting the clock I also used the clock test proposed by Muse and Weir (1992) that is independent of tree topology to check for rate constancy among lineages in the tree. The clock was not rejected when using this method, however, as it can only test three sequences at a time this approach lacks power (Bromham et al. 2000).

When estimating gene divergence dates, gene conversion obscures the results by making estimates younger, as sequences appear to be more similar than they actually are. In my date estimation analysis, I found that genes which were known, or suspected to be, affected by gene conversion usually gave younger estimates than expected from previous studies. Specifically, I believe that estimated dates for divergences of primate A_γ and G_γ globins and the mouse βh0 and βh1 and β1 and β2 globins are younger that the actual dates due to gene conversion. As methods are developed that allow estimation of dates from multiple datasets (Thorne and Kishino 2002), and also from non-clock like genes (Yoder and Yang 2000, Yang and Yoder 2003), gene families will become more attractive as data from which to estimate divergence dates. Gene families will be particularly useful because a single calibration, derived from a speciation event, can be used in multiple times in different paralogs. As such datasets will also have been prone to recombination, caution must be exercised when estimated dates are more shallow than the fossil record.

In most cases, dates estimated using ML were comparable to those obtained with the Bayes method. There are no apparent tendencies between the two methods because neither method necessarily produces younger or older estimates than the other. The dates for some ancestral nodes are particularly difficult to determine, like those that predate the duplication of two sequences with very short branches or those for very deep nodes. That is where I found the largest discrepancies between methods. One difference between the implementations of the Bayes and ML approaches is that the former accounts for uncertainties in the fossil record and the ML approach does not. From the results I can say that the way of assigning constraints on calibration nodes has a big impact in date estimation. Perhaps depending on the quality of the

fossil record a broader or narrower calibration range should be considered. Interestingly, other factors such as the priors required by the Bayes approach do not have as large an impact on date estimation as do calibration point bounds. This notion is in agreement with a previous report by Yang and Yoder (2003).

My investigation of evolutionary rate variation between pairs of newly duplicated gene sequences in the vertebrate β globin family produced two related results indicating a clocklike effect on rates. First, I found that following gene duplication most of the tested gene pairs were evolving with rates not significantly different from the rest of the phylogeny (one ratio-model vs two-ratios model). The two exceptions to this pattern were mouse βh0-βh1 and mouse β1-β2, which are affected by gene conversion. This result suggested that recently duplicated genes do not necessarily evolve with a different rate after duplication, or at least the change in rate does not occur immediately afterwards. Second, I found that following gene duplication, only in the mouse β1-β2 globin gene pair the two sequences were evolving at a significantly different rate from each other. The results indicate that most sequences in the dataset differ little in rate of evolution, whether they are the product of a recent duplication or not. In spite of observing local changes in rates in different parts of the β globin tree (e.g., changes in selection pressure, rejection of the global clock), I found very similar estimates for divergence dates when taken as an average over the entire tree. This would suggest that there are local changes in rates, but the average is rather clocklike.

Other factors that may influence rates of evolution in the β globin gene family are lineage-specific effects. I found that in all cases different vertebrate lineages were evolving with significantly different evolutionary rates. It would seem that lineage-

specific effects are very important in determining rates of β globin genes. This is in agreement with a previous observation regarding myoglobin suggesting that perhaps overall rates for separate lineages may be more informative than the average rate across the phylogeny (Romero-Herrera et al. 1973). Alternatively, this trend could reflect a large scale difference in selective pressure between β and γ globin rates, as suggested in a previous study on the evolution of vertebrate globin genes (Aguileta et al. submitted, see Chapter II).

Conclusions

The vertebrate β globin gene family is a perfect case study for the evolution of paralogous genes. Not only is this gene family well sampled in current gene and protein databases, but the wealth of knowledge about it is huge, providing the possibility of formulating and testing concrete hypothesis regarding the many aspects of its evolution with a solid background. In this thesis, I wanted to emphasize the complex ways in which different evolutionary forces have operated to give rise to the functional divergence observed in the β globin gene paralogs. What makes this gene family particularly interesting is that it has experienced frequent unequal crossing-over (gene conversion), episodes of positive selection, purifying selection, long-term differences in selective pressure among genes, and recurrent birth and death of some members in the cluster. In Chapter I it was shown how the evolution of the β globin gene family reflects a balance between homogenization by unequal crossing over and gene conversion, and diversification by mutation. I also showed how β globin paralogs are partitioned into domains of expression, thus facilitating their functional divergence, which is brought about both by positive Darwinian selection (as in the case of β and γ globins) and by differential patterns of purifying selection pressure (as seen in γ and ε globins). In order to fully discriminate between the competing DDC and Ohta models of gene family evolution, more tests are clearly necessary than those presented in Chapter I. However, I think that comparison between synonymous and nonsynonymous substitution rates provides a useful tool in studying relative roles of different evolutionary forces during the evolution of a gene family.

In Chapter III I investigated the evolution of globin structure and its relation to function, specifically, I was interested in the way selective constraints have shaped the globin protein. I compared hemoglobin and myoglobin looking for the evolutionary path through which an oligomeric form in hemoglobin evolved from a presumably monomeric ancestor, such as myoglobin. This evolutionary innovation in structure had a profound effect in function, allowing the allosteric regulation of oxygen transport in hemoglobin. As expected, the changes in hemoglobin interfaces are characterized by a significant presence of polar and charged residues, which are fundamental in establishing hydrogen bonds and salt bridges among protein subunits. Presumably, similar changes contributed to the creation of the necessary patches in the ancestral globin monomer that later served as interfaces. In contrast, the sites corresponding to interfaces in myoglobin show the inverse tendency and are mainly nonpolar. Once interfaces had evolved in oligomeric globins it became necessary to maintain the crucial site-site interactions between subunits. One hypothesis is that amino acid replacements in one monomer would be followed by compensatory mutations in the other. The results I obtained when I analysed α and β globins indicated that most of the replacements at the $\alpha\beta$ interfaces are conservative, as expected. However, the few sites at the interfaces detected to be under positive selection, did not appear to have undergone correlated mutations in α and β globins. This negative result might be caused by a weak signal for correlated replacements at these sites or because they do not make actual physical contact and therefore no compensatory mutations are needed. Crystallographic data will provide the last words on this matter.

In Chapter IV I put to the test the use of globin genes as examples of the molecular clock. I estimated dates for gene duplication and gene conversion events in the vertebrate β globin gene family and compared the dates inferred assuming the molecular clock with dates obtained without this constraint. Most dates estimated in this thesis for β globin genes are in fairly good agreement with previously reported dates, which were estimated assuming the clock. Furthermore, the dates obtained using new approaches based on maximum likelihood and Bayesian methods that allow the evolutionary rate to change among lineages when divergence times are estimated, provided similar results. One difference between the implementations of the Bayes and ML approaches is that the former accounts for uncertainties in the fossil record and the ML approach does not. One cautionary remark is that the way of assigning constraints on calibration nodes has a big impact in date estimation, especially when using the Bayes method. I also found that when estimating gene divergence dates, gene conversion obscures the results by making estimates younger, as sequences appear to be more similar than they would be without its effects. It is therefore advisable to test for recombination before estimating dates from molecular data. In general, dates estimated in this chapter for β globin genes produced results indicating a clocklike behaviour of rates. The evolutionary rates of the β globin genes I sampled might be influenced by lineage-specific effects, partly explaining their clocklike behaviour. Alternatively, this trend could reflect a large scale difference in selective pressure between β and γ globin rates.

Gene families have become very important objects of study in the field of molecular evolution, as they provide a natural framework for investigating the evolutionary dynamics of genes within the larger scope of the cell and genome. Future

work naturally points in that direction. Also, the work done using well known data, as in the case of globins, will help to encourage and guide investigation of new gene families. More genes will be sequenced for years to come and it is likely that most of them will be members of gene families, some perhaps not yet described. Understanding the dynamics of gene family evolution is of great importance to understand gene interactions at all levels.

References

Anderson CR, Jensen EO, LLewellyn DJ, Dennis ES, Peacock WJ (1996) A new hemoglobin gene from soybean: a role for hemoglobin in all plants. Proc Natl Acad Sci USA 93:5682-5687

Aris-Brosou S, Yang Z (2003) Bayesian Models of Episodic Evolution Support a Late Precambrian Explosive Diversification of the Metazoa. Mol Biol Evol 20:1947-1954

Bashford D, Chothia C, Lesk AM (1987) Determinants of a protein fold. Unique features of the globin amino acid sequences. J Mol Biol 196:199-216

Benton MJ (1990) Phylogeny of the major tetrapod groups: morphological data and divergence dates. J Mol Evol 30:409-424

Benton MJ (1999) Early origins of modern birds and mammals: molecules vs. morphology. Bioessays 21:1043-1051

Bielawski JP, Yang Z (2003) Maximum likelihood methods for detecting adaptive evolution after gene duplication. J Struct Funct Genomics 3:201-212

Bowen GJ, Clyde WC, Koch PL, Ting S, Alroy J, Tsubamoto T, Wang Y, Wang Y (2002) Mammalian dispersal at the Paleocene/Eocene boundary. Science 295:2062-2065

Brenner SE, Chothia C, Hubbard TJ (1997) Population statistics of protein structures: lessons from structural classifications. Curr Opin Struct Biol 7:369-376

Bridges CB (1936) The bar 'gene' a duplication. Science 83:210-211

Bromham L, Rambaut A, Fortey R, Cooper A, Penny D (1998) Testing the Cambrian explosion hypothesis by using a molecular dating technique. Proc Natl Acad Sci USA 95:12386-12389

Bromham L, Penny D, Rambaut A, Hendy MD (2000) The power of relative rate tests depends on the data. J Mol Evol 50:296-301

Brown TA (2002) Genomes 2nd ed. Oxford University Press, Oxford

Bunn HF (1981) Evolution of mammalian hemoglobin function. Blood 58:189-197

Burmester T, Weich B, Reinhardt S, Hankeln T (2000) A vertebrate globin expressed in the brain. Nature 407:520-523

Burmester T, Ebner B, Weich B, Hankeln T (2002) Cytoglobin: a novel globin type ubiquitously expressed in vertebrate tissues. Mol Biol Evol 19:416-421

Charlesworth D, Charlesworth B, McVean GA (2001) Genome sequences and evolutionary biology, a two-way interaction. Trends Ecol Evol 16:235-242

Chien H, Lukin JA (2001) Haemoglobin: cooperativity in protein-ligand interactions. In: Nature Encyclopedia of Life Sciences. London: Nature Publishing Group. http://www.els.net/ [doi:10.1038/npg.els.0001345]

Chothia C, Levitt M, Richardson D (1977) Structure of proteins: packing of alpha-helices and pleated sheets. Proc Natl Acad Sci USA 74:4130-4134

Chothia C (1992) Proteins. One thousand families for the molecular biologist. Nature 357:543-544

Cifelli RL (2000) Cretaceous mammals of Asia and North America. Paleontol. Soc. Korea Spec Publ 4:49-84

Clark AG (1994) Invasion and maintenance of a gene duplication. Proc Natl Acad Sci USA 91:2950-2954

Cleary ML, Schon EA, Lingrel JB (1981) Two related pseudogenes are the result of a gene duplication in the goat beta-globin locus. Cell 26:181-190

Cooper SJ, Murphy R, Dolman G, Hussey D, Hope RM (1996) A molecular and evolutionary study of the beta-globin gene family of the Australian marsupial Sminthopsis crassicaudata. Mol Biol Evol 13:1012-1022

Cracraft J (2001) Avian evolution, Gondwana biogeography and the Cretaceous-Tertiary mass extinction event. Proc R Soc Lond B Biol Sci 268:459-469

Crandall KA, Kelsey CR, Imamichi H, Lane HC, Salzman NP (1999) Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. Mol Biol Evol 16:372-382

Creighton TE (1993) Proteins: Structural and molecular properties. Freeman, New York

Czelusniak J, Goodman M, Hewett-Emmett D, Weiss ML, Venta PJ, Tashian RE (1982) Phylogenetic origins and adaptive evolution of avian and mammalian haemoglobin genes. Nature 298:297-300

D'Alessio G (1999) The evolutionary transition from monomeric to oligomeric proteins: tools, the environment, hypotheses. Prog Biophys Mol Biol 72:271-298

Dermitzakis ET, Clark AG (2001) Differential selection after duplication in mammalian developmental genes. Mol Biol Evol 18:557-562

Drouin G, Prat F, Ell M, Clarke GD (1999) Detecting and characterizing gene conversions between multigene family members. Mol Biol Evol 16:1369-1390

Drouin G (2002) Testing claims of gene conversion between multigene family members: examples from echinoderm actin genes. J Mol Evol 54:138-139

Efimov AV (1979) Packing of alpha-helices in globular proteins. Layer-structure of globin hydrophobic cores. J Mol Biol 134:23-40

Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA et al. (1980) The structure and evolution of the human beta-globin gene family. Cell 21:653-668

Farace MG, Brown BA, Raschella G, Alexander J, Gambari R, Fantoni A, Hardies SC, Hutchison CA, III, Edgell MH (1984) The mouse beta h1 gene codes for the z chain of embryonic hemoglobin. J Biol Chem 259:7123-7128

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368-376

Fermi G, Perutz MF, Shaanan B, Fourme R (1984) The crystal structure of human deoxyhaemoglobin at 1.74 A resolution. J Mol Biol 175:159-174

Fitch DH, Mainone C, Goodman M, Slightom JL (1990) Molecular history of gene conversions in the primate fetal gamma-globin genes. Nucleotide sequences from the common gibbon, Hylobates lar. J Biol Chem 265:781-793

Fitch DH, Bailey WJ, Tagle DA, Goodman M, Sieu L, Slightom JL (1991) Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. Proc Natl Acad Sci USA 88:7396-7400

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531-1545

Futuyma DJ (1998) Evolutionary Biology 3rd ed. Sinauer, Sunderland, Mass

Garner KJ, Lingrel JB (1989) A comparison of the beta A-and beta B-globin gene clusters of sheep. J Mol Evol 28:175-184

Gillemans N, McMorrow T, Tewari R, Wai AW, Burgtorf C, Drabek D et al. (2002) A functional and comparative analysis of globin loci in pufferfish and man. Blood 2842-2849

Gingerich PD and Uhen MD (1994) Time of origin of primates. J Hum Evol 27:443-445

Göbel U, Sander C, Schneider R Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins: Struct Funct Genet 18:309-317

Golding GB, Dean AM (1998) The structural basis of molecular adaptation. Mol Biol Evol 15:355-369

Goodman M, Moore GW, Matsuda G (1975) Darwinian evolution in the genealogy of haemoglobin. Nature 253:603-608

Goodman M, Ingwall RT, St Pierre S (1976a) Synthesis and conformation of sequential polypeptides of L-alanine and beta-aminobutyric acid. Macromolecules 9:1-6

Goodman M (1976) Protein sequences in phylogeny. In: Ayala A (ed) Molecular Evolution. Sinauer, Sunderland Mass, pp 141-159

Goodman M (1981) Globin evolution was apparently very rapid in early vertebrates: a reasonable case against the rate-constancy hypothesis. J Mol Evol 17:114-120

Goodman M, Koop BF, Czelusniak J, Weiss ML (1984) The eta-globin gene. Its long evolutionary history in the beta-globin gene family of mammals. J Mol Biol 180:803-823

Grassly NC, Holmes EC (1997) A likelihood method for the detection of selection and recombination using nucleotide sequences. Mol Biol Evol 14:239-247

Graur D, Li HW (1999) Fundamentals of molecular evolution. Sinauer, Sunderland Mass

Griffiths A et al. (1999) Introduction to genetic analysis 7th ed. Freeman, New York

Gu X (1999) Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol 16:1664-1674

Gu X, Vander VK (2002) DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. Bioinformatics 18:500-501

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421:63-66

Haldane JBS (1932) The causes of evolution. Longmans, New York

Hardies SC, Edgell MH, Hutchison CA, III (1984) Evolution of the mammalian beta-globin gene cluster. J Biol Chem 259:3748-3756

Hardison R (1998) Hemoglobins from bacteria to man: evolution of different patterns of gene expression. J Exp Biol 201 (Pt 8):1099-1117

Hardison RC, Margot JB (1984) Rabbit globin pseudogene psi beta 2 is a hybrid of delta- and beta-globin gene sequences. Mol Biol Evol 1:302-316

Harrison PM, Echols N, Gerstein MB (2001) Digging for dead genes: an analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome. Nucleic Acids Res 29:818-830

Hayasaka K, Fitch DH, Slightom JL, Goodman M (1992) Fetal recruitment of anthropoid gamma-globin genes. Findings from phylogenetic analyses involving the 5'-

flanking sequences of the psi gamma 1 globin gene of spider monkey Ateles geoffroyi. J Mol Biol 224:875-881

Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK, Hood L (1997) Gene families: the taxonomy of protein paralogs and chimeras. Science 278:609-614

Hill A, Hardies SC, Phillips SJ, Davis MG, Hutchison CA, III, Edgell MH (1984) Two mouse early embryonic beta-globin gene sequences. Evolution of the nonadult beta-globins. J Biol Chem 259:3739-3747

Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R (1996) Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. Nucleic Acids Res 24:4420-4449

Holm L, Sander C (1995) Dali: a network tool for protein structure comparison. Trends Biochem Sci 20:478-480

Hood L, Campbell JH, Elgin SC (1975) The organization, expression, and evolution of antibody genes and other multigene families. Annu Rev Genet 9:305-353

Hooper SD, Berg OG (2003) On the nature of gene innovation: duplication patterns in microbial genomes. Mol Biol Evol 20:945-954

Hosbach HA, Wyler T, Weber R (1983) The Xenopus laevis globin gene family: chromosomal arrangement and gene structure. Cell 32:45-53

Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from DNA sequence data. Genetics 132:583-589

Huelsenbeck JP, Larget B, Swofford D (2000) A compound poisson process for relaxing the molecular clock. Genetics 154:1879-1892

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

Hughes AL, Nei M (1989) Evolution of the major histocompatibility complex: independent origin of nonclassical class I genes in different groups of mammals. Mol Biol Evol 6:559-579

Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proc R Soc Lond B Biol Sci 256:119-124

Jacobs LL and Downs WR (1994) in Tomida Y, Li CK and Setoguchi T (eds) Rodent and Lagomorph Families of Asian Origins and Diversification. National Science Museum Monographs, Tokyo, pp 149-156

Jakobsen IB, Easteal S (1996) A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. Comput Appl Biosci 12:291-295

Jakobsen IB, Wilson SR, Easteal S (1997) The partition matrix: exploring variable phylogenetic signals along nucleotide sequence alignments. Mol Biol Evol 14:474-484

Jeffreys AJ (1979) DNA sequence variants in the G gamma-, A gamma-, delta- and beta-globin genes of man. Cell 18:1-10

Jensen RA (1976) Enzyme recruitment in evolution of new function. Annu Rev Microbiol 30:409-425

Johnson RM, Buck S, Chiu C, Schneider H, Sampaio I, Gage DA et al. (1996) Fetal globin expression in New World monkeys. J Biol Chem 271:14684-14691

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275-282

Jukes TH and Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp 21-123

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624-626

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

King JL, Jukes TH (1969) Non-Darwinian evolution. Science 164:788-798

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

Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA et al. (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus. Nature 390:364-370

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplications. Genome Biol 3:RESEARCH0008

Konkel DA, Maizel JV, Jr., Leder P (1979) The evolution and sequence comparison of two recently diverged mouse chromosomal beta--globin genes. Cell 18:865-873

Koop BF, Siemieniak D, Slightom JL, Goodman M, Dunbar J, Wright PC, Simons EL (1989) Tarsius delta- and beta-globin genes: conversions, evolution, and systematic implications. J Biol Chem 264:68-79

Krakauer DC, Nowak MA (1999) Evolutionary preservation of redundant duplicated genes. Semin Cell Dev Biol 10:555-559

Kretschmer PJ, Coon HC, Davis A, Harrison M, Nienhuis AW (1981) Hemoglobin switching in sheep. Isolation of the fetal gamma-globin gene and demonstration that the fetal gamma- and adult beta A-globin genes lie within eight kilobase segments of homologous DNA. J Biol Chem 256:1975-1982

Lacy E, Hardison RC, Quon D, Maniatis T (1979) The linkage arrangement of four rabbit beta-like globin genes. Cell 18:1273-1283

Lacy E, Maniatis T (1980) The nucleotide sequence of a rabbit beta-globin pseudogene. Cell 21:545-553

Lesk AM, Chothia C (1980) How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J Mol Biol 136:225-270

Lewin B (2000) Genes VII. Oxford University Press, Oxford

Li WH (1997) Molecular evolution, second edn. Sinauer Associates, Sunderland Mass

Li WH, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292:237-239

Li WH, Gojobori T (1983) Rapid evolution of goat and sheep globin genes following gene duplication. Mol Biol Evol 1:94-108

Liang J, Edelsbrunner H, Woodward C (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci 7:1884-1897

Liang J, Dill KA (2001) Are proteins well-packed? Biophys J 81:751-766

Lingrel JB, Townes TM, Shapiro SG, Spence SE, Liberator PA, Wernke SM (1983) Organization, structure, and expression of the goat globin genes. Prog Clin Biol Res 134:131-139

Luo ZX, Cifelli RL, Kielan-Jaworowska Z (2001) Dual origin of tribosphenic mammals. Nature 409:53-57

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151-1155

Maddison WP (2000) Testing character correlation using pairwise comparisons on a phylogeny. J Theor Biol 202:195-204

Martin RD (1993) Primate origins: plugging the gaps. Nature 363:223-234

Martin RD (2003) Palaeontology: Combing the primate record. Nature 422:388-9, 391

Martin SL, Vincent KA, Wilson AC (1983) Rise and fall of the delta globin gene. J Mol Biol 164:513-528

Massingham T, Davies LJ, Lio P (2001) Analysing gene function after duplication. Bioessays 23:873-876

Maynard SJ, Smith NH (1998) Detecting recombination from gene trees. Mol Biol Evol 15:590-599

Mazet F, Shimeld SM (2002) Gene duplication and divergence in the early evolution of vertebrates. Curr Opin Genet Dev 12:393-396

McLachlan AD (1971) Test for comparing related amino acid sequences. J Mol Biol 61:409-424

Meireles CM, Schneider MP, Sampaio MI, Schneider H, Slightom JL, Chiu CH et al. (1995) Fate of a redundant gamma-globin gene in the atelid clade of New World

monkeys: implications concerning fetal globin gene expression. Proc Natl Acad Sci USA 92:2607-2611

Muller HJ (1939) Reversibility in evolution considered from the standpoint of genetics. Biol Rev 14:261-280

Murzin AG, Finkelstein AV (1988) General architecture of the alpha-helical globule. J Mol Biol 204:749-769

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536-540

Muse SV, Weir BS (1992) Testing for equality of evolutionary rates. Genetics 132:269-276

Muse SV, Gaut BS (1997) Comparing patterns of nucleotide substitution rates among chloroplast loci using the relative ratio test. Genetics 146:393-399

Nei M (1969) Gene duplication and nucleotide substitution in evolution. Nature 221:40-42

Nei M, Rogozin IB, Piontkivska H (2000) Purifying selection and birth-and-death evolution in the ubiquitin gene family. Proc Natl Acad Sci USA 97:10866-10871

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-936

Nowak MA, Boerlijst MC, Cooke J, Smith JM (1997) Evolution of genetic redundancy. Nature 388:167-171

Nutall GHF (1904) Blood immunity and blood relationship. Cambridge University Press, Cambridge

O'Brien SJ, Eizirik E, Murphy WJ (2001) Genomics. On choosing mammalian genomes for sequencing. Science 292:2264-2266

Ohno S (1967) Sex chromosomes and sex-linked genes. Springer Verlag, Berlin

Ohno S (1970) Evolution by gene duplication. Springer Verlag, Berlin

Ohta T (1980a) Evolution and variation of multigene families. Lecture Notes in Biomathematics. Springer, New York

Ohta T (1980b) Amino acid diversity of immunoglobulins as a product of molecular evolution. J Mol Evol 15:29-35

Ohta T (1983) On the evolution of multigene families. Theor Popul Biol 23:216-240

Ohta T (1987) Simulating evolution by gene duplication. Genetics 115:207-213

Ohta T (1988a) Evolution by gene duplication and compensatory advantageous mutations. Genetics 120:841-847

Ohta T (1988b) Time for acquiring a new gene by duplication. Proc Natl Acad Sci USA 85:3509-3512

Ohta T (1990) How gene families evolve. Theor Popul Biol 37:213-219

Ohta T (1992) The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst 23:263-268

Ohta T (1993) Pattern of nucleotide substitutions in growth hormone-prolactin gene family: a paradigm for evolution by gene duplication. Genetics 134:1271-1276

Ohta T (1998) On the pattern of polymorphisms at major histocompatibility complex loci. J Mol Evol 46:633-638

Ohta T (2000) Evolution of gene families. Gene 259:45-52

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH--a hierarchic classification of protein domain structures. Structure 5:1093-1108

Orengo CA, Pearl FM, Bray JE, Todd AE, Martin AC, Lo CL, Thornton JM (1999) The CATH Database provides insights into protein structure/function relationships. Nucleic Acids Res 27:275-279

Orgel LE (1977) Gene-duplication and the origin of proteins with novel functions. J Theor Biol 67:773

Ota T, Nei M (1995) Evolution of immunoglobulin VH pseudogenes in chickens. Mol Biol Evol 12:94-102

Page R D M and Holmes E C (2002) Molecular Evolution a Phylogenetic Approach. Blackwell Science, Oxford

Patthy L (1999) Protein Evolution. Blackwell Science, Oxford

Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) Correlated mutations contain information about protein-protein interaction. J Mol Biol 271:511-523

Perutz MF, et al. (1960) Structure of haemoglobin. A three-dimensional Fourier synthesis at 5.5 A resolution, obtained by x-ray analysis. Nature 185:416-422

Perutz MF (1970) Stereochemistry of cooperative effects in haemoglobin. Nature 228:726-739

Perutz MF, Bauer C, Gros G, Leclercq F, Vandecasserie C, Schnek AG, Braunitzer G, Friday AE, Joysey KA (1981) Allosteric regulation of crocodilian haemoglobin. Nature 291:682-684

Perutz MF (1983) Species adaptation in a protein molecule. Mol Biol Evol 1:1-28

Perutz MF, Wilkinson AJ, Paoli M, Dodson GG (1998) The stereochemical mechanism of the cooperative effects in hemoglobin revisited. Annu Rev Biophys Biomol Struct 27:1-34

Piontkivska H, Rooney AP, Nei M (2002) Purifying selection and birth-and-death evolution in the histone H4 gene family. Mol Biol Evol 19:689-697

Posada D, Crandall KA, Holmes EC (2002) Recombination in evolutionary genomics. Annu Rev Genet 36:75-97

Poyart C, Wajcman H, Kister J (1992) Molecular adaptation of hemoglobin function in mammals. Respir Physiol 90:3-17

Rambaut A, Bromham L (1998) Estimating divergence dates from molecular sequences. Mol Biol Evol 15:442-448

Romero-Herrera AE, Lehman H, Joysey KA, Friday AE (1973) Molecular evolution of myoglobin and the fossil record: a phylogenetic synthesis. Nature 246:389-395

Romero-Herrera AE, Lehman H, Joysey KA, Friday AE (1978) On the evolution of myoglobin. Phil Trans Roy Soc B 283:61-163

Rowe T (1999) At the roots of the mammalian family tree. Nature 398:283-284

Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK et al. (2000) Comparative genomics of the eukaryotes. Science 287:2204-2215

Saban J, King D (1994) Sequence of the sheep fetal beta globin gene and flanking region. Biochim Biophys Acta 1218:87-90

Sanderson MJ (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol Biol Evol 14:1218-1232

Sarich VM, Wilson AC (1967) Rates of albumin evolution in primates. Proc Natl Acad Sci USA 58:142-148

Satoh H, Inokuchi N, Nagae Y, Okazaki T (1999) Organization, structure, and evolution of the nonadult rat beta-globin gene cluster. J Mol Evol 49:122-129

Sawyer SA (1999) GENECONV: a computer package for the statistical detection of gene conversion, distributed by the author, Department of Mathematics, Washington University in St. Louis, available at http://.math.wustl.edu/~sawyer

Schimenti JC, Duncan CH (1985a) Concerted evolution of the cow epsilon 2 and epsilon 4 beta-globin genes. Mol Biol Evol 2:505-513

Schimenti JC, Duncan CH (1985b) Structure and organization of the bovine beta-globin genes. Mol Biol Evol 2:514-525

Schimenti JC (1994) Gene conversion and the evolution of gene families in mammals. Soc Gen Physiol Ser 49:85-91

Schon EA, Cleary ML, Haynes JR, Lingrel JB (1981) Structure and evolution of goat gamma-, beta C- and beta A-globin genes: three developmentally regulated genes contain inserted elements. Cell 27:359-369

Shaanan B (1983) Structure of human oxyhaemoglobin at 2.1 A resolution. J Mol Biol 171:31-59

Shapiro SG, Schon EA, Townes TM, Lingrel JB (1983) Sequence and linkage of the goat epsilon I and epsilon II beta-globin genes. J Mol Biol 169:31-52

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

Shionyu M, Takahashi K, Go M (2001) Variable subunit contact and cooperativity of hemoglobins. J Mol Evol 53:416-429

Shoshani J, Groves CP, Simons EL, Gunnell GF (1996) Primate phylogeny: morphological vs. molecular results. Mol Phylogenet Evol 5:102-154

Silva MM, Rogers PH, Arnone A (1992) A third quaternary structure of human hemoglobin A at 1.7-A resolution. J Biol Chem 267:17248-17256

Slatkin M, Maddison WP (1989) A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics 123:603-613

Slightom JL, Blechl AE, Smithies O (1980) Human fetal G gamma- and A gamma-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell 21:627-638

Slightom JL, Chang LY, Koop BF, Goodman M (1985) Chimpanzee fetal G gamma and A gamma globin gene nucleotide sequences provide further evidence of gene conversions in hominine evolution. Mol Biol Evol 2:370-389

Smith FR, Simmons KC (1994) Cyanomet human hemoglobin crystallized under physiological conditions exhibits the Y quaternary structure. Proteins 18:295-300

Springer MS, Murphy WJ, Eizirik E, O'Brien SJ (2003) Placental mammal diversification and the Cretaceous-Tertiary boundary. Proc Natl Acad Sci USA 100:1056-1061

Stephens SG (1951) Possible significance of duplication in evolution. Ad Genet 4:247-265

Swofford, D. C. 1998. PAUP* 4.0-Phylogenetic analysis using parsimony (* and other methods). Version 4.0. Sinauer Assoc., Sunderland, MA.

Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol 203:439-455

Takezaki N, Rzhetsky A, Nei M (1995) Phylogenetic test of the molecular clock and linearized trees. Mol Biol Evol 12:823-833

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631-637

Tavare S, Marshall CR, Will O, Soligo C, Martin RD (2002) Using the fossil record to estimate the age of the last common ancestor of extant primates. Nature 416:726-729

Thompson TD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple alignment aided by quality analysis tools. Nucleic Acids Research 24:4876-4882

Thorne JL, Kishino H, Painter IS (1998) Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol 15:1647-1657

Thorne JL, Kishino H (2002) Divergence time and evolutionary rate estimation with multilocus data. Syst Biol 51:689-702

Tiplady B, Goodman M (1977) Primitive haemoglobin. J Mol Evol 9:343-347

Townes TM, Fitzgerald MC, Lingrel JB (1984) Triplication of a four-gene set during evolution of the goat beta-globin locus produced three genes now expressed differentially during development. Proc Natl Acad Sci USA 81:6589-6593

Trabesinger-Ruef N, Jermann T, Zankel T, Durrant B, Frank G, Benner SA (1996) Pseudogenes in ribonuclease evolution: a source of new biomacromolecular function? FEBS Lett 382:319-322

van Ooyen A, van den BJ, Mantei N, Weissmann C (1979) Comparison of total sequence of a cloned rabbit beta-globin gene and its flanking regions with a homologous mouse sequence. Science 206:337-344

Van Tuinen M, Hedges SB (2001) Calibration of avian molecular clocks. Mol Biol Evol 18:206-213

Voet D, Voet JG (1995) Biochemistry 2nd ed. Wiley, New York

Walsh JB (1995) How often do duplicated genes evolve new functions? Genetics 139:421-428

Wang Y, Gu X (2001) Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. Genetics 158:1311-1320

Wheeler D, Hope R, Cooper SB, Dolman G, Webb GC, Bottema CD, Gooley AA, Goodman M, Holland RA (2001) An orphaned mammalian beta-globin gene of ancient evolutionary origin. Proc Natl Acad Sci USA 98:1101-1106

Wittenberg JB, Bolognesi M, Wittenberg BA, Guertin M (2002) Truncated hemoglobins: a new family of hemoglobins widely distributed in bacteria, unicellular eukaryotes, and plants. J Biol Chem 277:871-874

Worobey M (2001) A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria. Mol Biol Evol 18:1425-1434

Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306-314

Yang Z, Kumar S, Nei M (1995) A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641-1650

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555-556

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568-573

Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431-449

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496-503

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19:908-917

Yang Z, Yoder AD (2003) Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. Syst Biol 52:705-716

Yoder AD, Yang Z (2000) Estimation of primate speciation dates using local molecular clocks. Mol Biol Evol 17:1081-1090

Zhang J (2000) Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. J Mol Evol 50:56-68

Zhang J (2003) Evolution by gene duplication: an update. Trends Ecol Evol 18:292-298

Zhu H, Riggs AF (1992) Yeast flavohemoglobin is an ancient protein related to globins and a reductase family. Proc Natl Acad Sci USA 89:5015-5019

Zuckerkandl E, Pauling L (1962) Molecular disease, evolution and genic heterogeneity. In: Kash and Pullman (eds) Horizons in Biochemistry. Academic Press, New York, pp 189-225

Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. J Theor Biol 8:357-366