VERTEBRATE PHYLOGENOMICS AND GENE FAMILY EVOLUTION

James A. Cotton

A thesis submitted for the degree of Doctor of Philosophy to the
Division of Environmental and Evolutionary Biology
Institute of Biomedical and Life Sciences
University of Glasgow

June 2003

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

Declaration

I declare that the work recorded in this thesis is entirely my own, unless otherwise stated, and that it is my own composition. No part of it has been submitted for any other degree to any institution.

James Cotton
University of Glasgow
February 2003

Abstract

This thesis is about 2 topics: the evolution of gene families by the birth-death process of gene duplication and gene loss, and phylogenetic inference. It is a central theme that these two processes are intimately associated – the phylogenies of gene families (of any gene) are shaped by the processes of gene duplication and gene loss, as much as by the processes of speciation and extinction occurring among the species the gene is evolving in. This has two results. Firstly, that we need to know, or assume, something about the processes of gene duplication and loss to correctly understand the pattern of speciation, or cladogenesis, in a group of organisms. Secondly, that we need to know, or assume, something about this pattern if we are to fully appreciate the effect of gene duplication and loss on a gene family phylogeny.

The main part of this thesis investigates the use of reconciled tree methods in unravelling species phylogeny and the evolution of gene families. Part of this investigation involves placing reconciled tree methods (and the use of these methods to infer species phylogeny, known as gene tree parsimony), in the context of some related methods: supertree methods and "simultaneous analysis" of combined data. Two empirical studies complete this part of the thesis – one attempting to infer the higher-level phylogeny of vertebrates using gene tree parsimony, and another focusing on a lower taxonomic level, on primate phylogeny. This chapter attempts an integrated study of gene duplication and species phylogeny, which uses information about gene duplication to help date evolutionary events.

Despite the close relationship between gene duplication and speciation on phylogenies, it is possible to study gene duplication independently. If we restrict ourselves to genes sampled from a single genome, gene family trees represent gene duplications and gene losses occurring during the history of a single species, so the complication of speciation and extinction is eliminated. By realising that the processes of gene duplication and loss in these trees are analogous to the processes of speciation and extinction in species phylogenies, we can harness a toolkit of methods developed for more traditional phylogenies to study these molecular processes. Two such methods are models of cladistic tree shape and birth-death models, which

allow the first estimates of the rate of gene loss.

Acknowledgements

This work was supported by a Natural Environment Research Council studentship and the Wolfson Foundation.

Particular thanks go to my supervisor, Rod Page. Rod has been all that a supervisor should be and has been involved in every part of this thesis – his contribution is rightfully acknowledged by the fact that most of the work is co-authored by Rod, and these are not simply 'honorary' authorships.

Martyn Kennedy and Trevor Cotton each read large chunks of this thesis, and the final product has been greatly improved by their insightful comments and extensive red pen. Mike Charleston also read several of the chapters, and laughed regularly at my mathematical ineptitude.

Aside from Rod, Martyn, Trev and Mike, a number of people have contributed in various ways to particular chapters. I'm grateful to Olaf Bininda-Emonds for inviting me to write chapter 2. Gavin Naylor, the late Joe Slowinski and several anonymous reviewers provided constructive comments on various versions of the manuscript of chapter 4. For chapter 6, thanks to Steve Heard, Arne Mooers and Ed Stam for generously providing their data on species tree imbalances, and particularly to Kate Harcourt-Brown for providing both published and unpublished data. The manuscript for this chapter was greatly improved following comments from Arne Mooers and particularly Andy Purvis, who (along with Rod) stopped me making a slight fool of myself by relying on midpoint rooting. Xun Gu and colleagues made available their data, making the work described in chapter 7 possible. Thanks to David Sankoff for the invitation to write Appendix A. Two anonymous reviewers provided many helpful comments on Appendix B, and Mike Charleston translated Rod's "pseudo-mathematical gibberish" into real maths.

Thanks to my family for all their moral support during the last three and a bit years – and also to mum and dad for a number of essential "loans" and to Neill and

Trev for numerous free dinners.

Thanks to everyone in DEEB for making it a pleasure to come to work in the "mornings", and particularly to the members of the taxonomy group, past and present, and most particularly of all thanks to Tom, for fostering the poor working environment of 3-14c.

Last of all, and most of all, thanks to Claire for making me want to leave at the end of each day.

Contents

List of Figures

Chapter 1

Introduction: Gene Family Evolution and Phylogeny

A common ancestry relates all living things, as each species has evolved from another species. Phylogenetic trees represent the pattern of this relatedness, and these trees play a major part in understanding the evolutionary history of life on earth. Construction of phylogenetic trees originally depended upon examining and comparing anatomical features of organisms – known as morphological characters. This was (and still is) a time-consuming and specialised endeavour. When Zuckerandl and Pauling (1965) pointed out that molecular data could also be used to build phylogenies, they inspired a major revolution in systematics.

Perhaps the first major impact of molecular systematics was that phylogenies could be constructed showing the relationships between organisms that shared almost no anatomical similarity (or indeed, have virtually no anatomy) – the first phylogenetic trees incorporating the full known diversity of living things were constructed (Woese, 2000; Woese and Fox, 1977), and the diversity of microorganisms such as bacteria became apparent for the first time (DeLong and Pace, 2001; Pace et al., 1986). A second repercussion of the new data took a little longer to be felt. The technology used to isolate and sequence DNA has improved greatly, with advances such as the polymerase chain reaction (PCR) and automated sequencing of DNA, producing an explosion in the amount of molecular data available. This increase has been in both the scope of organisms covered (ranging from viruses and even prion proteins to extinct birds and insects) and the amount of

data available for particular species, particularly with the increasing amount of genomic-level sequence available. This represents an enormous amount of potential phylogenetic information. The increasing width of phylogenetic data available has begun to make assembling the "tree of life" – a phylogeny relating all known living things – a realistic (albeit distant) possibility. This has sparked renewed interest in methods for combining phylogenetic trees, or supertree methods, and I present a particular view on this endeavour. There can only be one 'true' phylogeny showing the pattern of speciation between relatives, but independent molecular markers often disagree about relationships between organisms. The increasing depth of phylogenetic data bearing on particular phylogenetic problems has focused interest on why this disagreement might exist, an ongoing theme in this thesis.

The final revolution prompted by molecular phylogenetics is perhaps the most subtle. It hinges on the simple realisation that molecular phylogenies are not phylogenies showing the relationships between organisms, but instead show the relationship between genes themselves. The main impact of this has been to move phylogenetic methods into the mainstream of molecular evolutionary biology. Population genetics now employs powerful and flexible coalescent models that require phylogenetic trees relating alleles (Tavarè, 1984), and studies of molecular adaptation use phylogenetic trees to locate selected substitutions. Perhaps most significantly, molecular phylogenetic methods have become central to understanding the evolution of genomes. Genome evolution is probably the fastest-growing area in evolutionary biology today. This thesis is also about how we can use molecular phylogenies to study the evolution of genes and genomes. Gene duplication and gene loss are, as we will see, among the most important processes shaping the diversity of genes within a genome, as well as being important mediators of genome size.

1.1 Thesis outline

The main aim of this thesis is to demonstrate the potential utility of reconciled trees in understanding how gene duplication has affected phylogeny.

The chapters in this thesis have been written as self-contained papers, so there is some repetition of introductory material and discussion. The rest of this intro-

duction presents background information about phylogenies and the processes of gene duplication. It is necessarily a biased, personal, and very concise summary, but is intended to introduce all of the material needed to place the rest of the thesis in context. Far more complete references on phylogeny are available (Page and Holmes, 1998; Swofford et al., 1996). The literature on gene duplication is more fragmented – Ohno (1970) is the classic reference on the subject and there is some recent coverage in molecular evolution textbooks (e.g. Hughes, 1999a; Li, 1997).

Part I consists of two chapters that explore reconciled trees and gene tree parsimony in more depth. Chapter 2 explains that gene tree parsimony can be seen as a supertree method. It tries to persuade supertree workers that correctly resolving conflict between subtrees relies on understanding the causes of this conflict, and suggests that reconciled trees can help in understanding one source of conflict. This chapter is currently in review for a forthcoming book on supertree methods, edited by Olaf Bininda-Emonds and to be published by Kluwer. It was co-authored with Rod Page. Chapter 3 contrasts gene tree parsimony with 'simultaneous analysis' or 'combined matrix' methods for tree reconstruction, and was written as a reply to a paper criticising gene tree parsimony methods (Simmons and Freudenstein, 2002) from this viewpoint. This chapter is currently in press at *Molecular Phylogenetics and Evolution*, and is co-authored with Rod Page.

Part II consists of two chapters attempting to use gene tree parsimony to reconstruct phylogeny. Chapter 4 attempts to resolve a long-standing debate over high-level phylogeny of vertebrates – the largest molecular dataset for vertebrates comes from whole mitochondrial genome sequences, and disagrees significantly with morphological and palaeontological views. We use reconciled tree methods to show that nuclear genes support the traditional picture of vertebrate phylogeny. This chapter has been published as *Proc. Roy. Soc. Lond.* B (2002) **269**, 1555-1561. It too was co-authored with Rod Page. Chapter 5 attempts an integrated study of primate phylogeny and gene duplication in the evolution of this group. It largely supports the current picture of primate phylogeny, and presents a molecular timescale for both phylogenetic events and gene duplications in the primates.

Part III consists of two chapters studying the process of gene duplication outside the context of species phylogeny. One important consequence of seeing gene duplications in a phylogenetic context is that a number of phylogenetic methods that have been developed to study the processes of speciation and extinction can

also be used to study the analogous processes of gene duplication and gene loss. Chapter 6 looks at what the the cladistic shape of trees for gene families might be able to show us about the process of gene duplication, by comparing these trees with conventional phylogenetic trees relating species. Chapter 7 uses birth-death models of speciation and extinction to study the distribution of gene duplications through vertebrate evolution, using these models to estimate both the duplication rate and, for the first time, the rate of loss of genes in vertebrates.

The first two appendices are published papers co-authored by myself and Rod Page, but where Rod played a larger part than in the rest of the work included. Appendix A is an introduction to using reconciled tree methods to study gene family evolution, which was published in *Comparative Genomics : Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*, D. Sankoff and J. H. Nadeau, eds. Kluwer Academic Publishers. Appendix B presents the duplication clustering algorithm used a number of times in this thesis, and has been published in *Pacific Syposium on Biocomputing, 2002*, R. B. Altman, A.K. Dunker, L. Hunter, K. Lauderdale and T. E. Klein, Eds., World Scientific Press (See `http://psb.stanford.edu/`). Appendix C contains a reprint of chapter 4 in its final published form.

1.2 Tree terminology

Phylogenetic trees are mathematical graph structures, so much of the appropriate terminology is mathematical. Phylogenetic trees are also *trees* in the mathematical sense – that is they are connected, acyclic graphs (Wilson, 1996). Trees are composed of *vertices*, or *nodes*, and *edges*, known as *branches*. Some vertices have degree one and are known as *leaves*. The vertices with higher degree are known as *internal nodes*. The leaves are labelled, representing the organisms whose evolution the phylogenetic tree represents. Internal nodes are generally not labelled.

Internal nodes of degree three are known as *bifurcations*, while those of higher degree are termed *polytomies*. A tree is *fully resolved* or *bifurcating* if all its internal nodes are bifurcations, otherwise it is known as *partially resolved* – in the special case of a tree with a single internal node of high degree, which is *unresolved*. Polytomies can be thought of as either representing uncertainty about the pattern of evolution at a particular place in the tree, in which case they are known

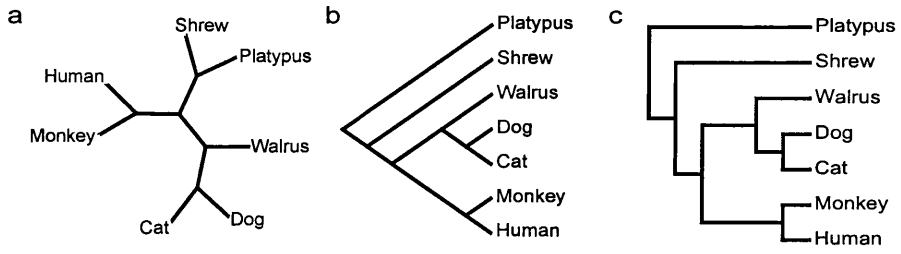

Figure 1.1: An unrooted phylogenetic tree (a) and two equivalent representations of a rooted tree (b and c).

as *soft polytomies,* or as correctly representing the fact that the pattern of evolution produced more than two lineages simultaneously, known as *hard polytomies* (Maddison, 1989). Polytomies are generally considered to be soft in most circumstances.

Trees as described so far represent the pattern of evolution, but include no information about the direction of evolution, and so cannot represent concepts such as 'more closely related' or 'ancestral'. To do this, one internal node of a tree is designated as being the oldest, as representing the earliest evolutionary event. This node is called the *root*, and such a tree is termed a *rooted tree* as opposed to an *unrooted tree* (figure 1.1), and is a directed graph – internal nodes have indegree one and outdegree two or more, leaves have indegree one and outdegree zero. These trees can show ancestors and descendants – node *a* is *ancestral* to node (or leaf) *b* precisely if the path from the root to *b* passes through *a*. Note this means that a node is its own ancestor – a node that is ancestral to but not identical to another node can be termed a *proper ancestor*. Node b is a *descendant (proper descendant)* of node *a* if and only if *a* is an ancestor (proper ancestor) of *b*. Note that in creating a rooted tree from an unrooted tree, the root is often placed along a branch of the tree, creating an additional internal node. This node, strictly speaking, has indegree zero and outdegree greater than zero, but is often represented with an additional incident edge.

In general, the lengths of edges on a tree have no meaning – serving only to create a pleasing representation of a tree (as in figure 1.1), but sometimes edge lengths do show information. In this case they show the evolutionary distance between the two vertices connected by the edge, or the inferred amount of change

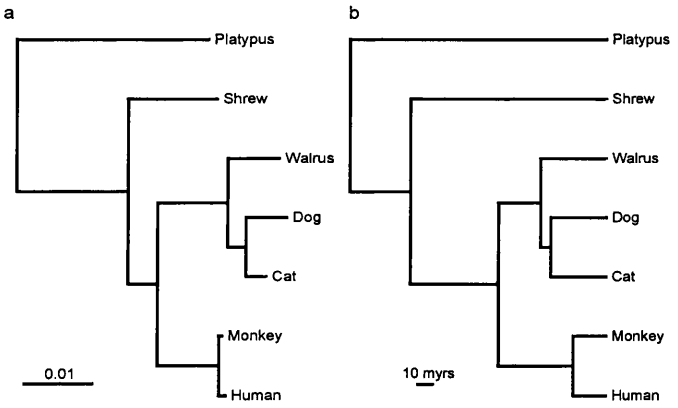

Figure 1.2: (a) A tree with branch length information, and (b) an ultrametric tree. Branch lengths on ultrametric trees can represent actual time.

in sequence (figure 1.2). This distance is taken to represent the amount of evolutionary change that occurred between the two speciation events connected by the edge. This change may be considered as representing the time between the speciation events. If this is the case, we would expect the tree to be *ultrametric* – if all the leaves are contemporaries, the lengths of all paths from the root to the leaves should be equal. An ultrametric and non-ultrametric tree are compared in figure 1.2.

1.3 How phylogenies are reconstructed

Molecular phylogenetics can be seen as the problem of reconstructing the branching tree diagram connecting a set of amino-acid or nucleotide sequences – the phylogenetic tree that has the sequences as its leaves. An almost bewildering variety of approaches have been taken to this problem over the years, which I will summarise only very briefly here.

Most methods split the process of inferring the phylogeny into two steps – separating *alignment* from inferring the tree itself. Alignment is the lining up of the sequences so that amino-acids and nucleotides that are thought to be evolutionarily related are compared side-by-side. This lining up is accomplished by inserting spaces, known as gaps, into one or more of the sequences. Alignment is probably the most challenging and least well-defined stage of phylogenetic reconstruction.

The methods used in this thesis are automated methods, but much alignment is still carried out subjectively, by eye. Automatic alignment relies on maximising the similarity between adjacent nucleotides across an alignment while minimising the number of gaps inserted into the sequences, which are combined into an alignment score. Finding the minimum score alignment for a pair of sequences is relatively easy, with exponential-time algorithms available to find this alignment by dynamic programming (Needleman and Wunsch, 1970). The problem becomes progressively harder for *multiple alignment* of more than two sequences. The dynamic programming algorithm can find optimal alignments in time $O(n^2)$ for two sequences of length n, but needs time $O(n^3)$ for three sequences, $O(n^4)$ sequences for four sequences, and so on to $O(n^m)$ for m sequences of length n. This is clearly impractical for more than a few sequences, so the most widely-used methods, such as implemented in ClustalW (Thompson et al., 1994) and used in this thesis, break the problem down into a number of pairwise alignments, following a rough phylogenetic tree known as a guide tree – more closely related sequences are aligned to each other first, then other sequences are aligned against this alignment, and so on, producing what is hopefully a reasonable approximation of the best alignment.

Once an alignment is available, the earliest methods used for reconstructing phylogenies involved calculating an evolutionary distance between each of the sequences in the alignment, and then using these distances to calculate a phylogenetic tree. The distances could be as simple as counting the number of amino-acids or nucleotides that were different between two sequences, but more complex distance measures are more generally used, that can correct for errors in this simple estimate – for example correcting for *multiple hits* (where a single nucleotide difference between two sequences has actually been caused by two sequential changes at the same position in the sequence). In theory, the distances between a set of sequences should precisely define a phylogenetic tree, which means they must be additive and satisfy the 4-point condition. This is rarely the case, however, due to estimation errors, so fitting a phylogenetic tree to a set of distances usually involves some distortion of these distances. There are a number of methods of combining distances into trees, some of which are very fast because they simply cluster sequences together rather than search for the tree best fitting the distances. Probably the most widely-used method, *neighbour-joining*, clusters most similar sequences together sequentially, and has been shown to be reasonably accurate in simulation

studies (Huelsenbeck and Hillis, 1993) and on real data (Hillis et al., 1992).

Distance methods have a number of drawbacks – converting an alignment into distances discards a great deal of information, and distance methods may be easily mislead by convergent evolution. These problems led to maximum parsimony methods becoming more popular. Maximum parsimony attempts to find the tree minimising the number of implied evolutionary changes between the sequences. A number of authors have claimed that parsimony has a philosophical justification in minimising the number of *ad hoc* hypotheses that similar characters have evolved in two species due to convergence rather than shared descent, and a fairly large number of biologists continue to prefer maximum parsimony to the exclusion of other phylogenetic methods. Indeed, until recently, it was the only practical alternative to distance methods for morphological data (Lewis, 2001). Another view of maximum parsimony methods is that they assume that evolutionary changes are rare (Felsenstein, 1973; Goldman, 1990) or that no common mechanism unites the evolution of a character across different parts of the tree (Tuffley and Steel, 1997). This assumption leads maximum-parsimony to make incorrect inferences of phylogeny when rates of evolution in different parts of a phylogeny are substantially different (long-branch attraction, Felsenstein, 1978b; see Sanderson and Shaffer, 2002, for an up-to-date discussion).

The next big development in phylogenetic methodology came about when people started considering phylogenetic inference as being analogous to other statistical inference problems. Just as inferring, say, the difference in weight between samples of tissue from two plants is properly seen as a statistical problem, so the problem of inferring a phylogenetic tree can be seen in the same light. This statistical view allows us to use statistical ideas like constructing confidence intervals for tree estimates (Holmes, 2003) and hypothesis testing (Huelsenbeck and Crandall, 1997). One significant problem is that the parameter to be estimated is not a simple scalar number but the rather more complex parameter of the phylogenetic branching diagram (Yang et al., 1995).

In fact, this viewpoint is rather old – parsimony methods were first proposed as a way of approximating the maximum likelihood estimate of a phylogenetic tree (Edwards, 1996). It was not, however, until Felsenstein (1978a, 1981) presented tractable probabilistic models of evolution, that allowed the likelihood of a phylogenetic tree to be computed, that statistical methods for inferring trees became

popular. The likelihood of a statistical model is the probability of the data given the model. The *principle of likelihood*, which dates back to R. A. Fisher (Edwards, 1972) states that the model which makes the observed data most probable (i.e. the maximum likelihood model) should be preferred. In the context of phylogenetics, the model includes the tree topology and a number of other parameters, such as branch lengths and the probabilities of changes from one nucleotide to another. Increasingly complex models are being formulated and used, for example incorporating substitution rates between codons (Yang and Nielsen, 2002) and incorporating information about protein structure (Thorne et al., 1996). The ability to model a wide variety of different evolutionary assumptions is a major advantage of probabilistic methods of phylogenetic inference. The main drawback of likelihood methods is that they can be extremely computationally intensive and so extremely slow, particularly as thorough searches of tree space appear to be needed to find multiple maximum-likelihood solutions (Chor et al., 2000).

Bayesian methods for phylogenetic inference have been introduced recently (Larget and Simon, 1999). They have much in common with likelihood methods, requiring the same probabilistic models, but these methods use Bayes' theorem to find the actual probability of a tree given some data. Bayes' theorem relates this probability, $p(model|data)$, to the likelihood, $p(data|model)$, by the equation: $p(model|data) = \frac{p(data|model)p(model)}{p(data)}$. Obviously, the preferred model is the most probable model, and having actual probabilities for models is very desirable, but comes at a cost – we need to assume prior probabilities for both the model and data, which can be difficult to do precisely (Huelsenbeck et al., 2002, pp. 684). One strength of these methods is the power to integrate across nuisance parameters such as branch lengths and substitution rates using Markov chain Monte Carlo (MCMC). The use of MCMC can also make Bayesian methods considerably faster than maximum-likelihood methods. Bayesian methods are becoming increasingly widely used.

Distance, parsimony and likelihood methods have been used at different points in this thesis.

1.4 Gene duplications

1.4.1 What are gene duplications?

As briefly mentioned above, there are a number of other evolutionary processes that can be represented and understood using phylogenetic trees besides speciation. For example, linguists have been using tree-like diagram for as long, if not longer, than evolutionary biologists (Craw, 1992), to represent the pattern of diversification of human languages. Any process generating this pattern of branching and divergence can reasonably represented by a tree diagram, and this is a common property of many evolutionary processes. One process showing particularly close parallels with the evolution of species is the evolution of genes themselves. Genes do not arise *de novo*, but rather are produced by the modification of other genes. This modification often follows the physical copying of the DNA that comprises the gene, allowing these modifications to occur without altering (and so, most likely, damaging) the function of the original DNA sequence. This multiplication of gene-carrying DNA is known as gene duplication.

Just as speciation splits two populations that subsequently follow their own evolutionary history, so two duplicated genes then have their own fates, accumulating mutations independently. In fact, the analogy goes even further than this, for just as two incipient species can be united by introgression, so the sequences of duplicated genes can be homogenised by the process of gene conversion (Archibald and Roger, 2002; Li, 1997, pp. 310-315). Molecular phylogenies are intended to represent the pattern of evolution of the species shown on the phylogeny, but gene duplication and speciation are, in fact, indistinguishable on a molecular phylogeny – both are splitting events that give rise to independent lineages of a gene. In one case, the lineages evolve independently because they are present in independent gene pools, in the other, the lineages are independent because they are present at two distinct loci in the same genome.

Although, as discussed later, much interest has focused on gene duplications in vertebrates, there is substantial evidence (e.g. Brenner et al., 1995; Wolfe and Shields, 1997) that gene duplications have also been important in other organisms, such as in the evolution of cell-to-cell communication pathways in the first multicellular animals (Ono et al., 1999; Suga et al., 1999). Gene duplication occurs

through a variety of mechanisms, at a variety of scales, and leads to the formation of gene families of related genes.

1.4.2 Mechanisms of gene duplication

Possible mechanisms for gene duplication are unequal crossing-over, replicative transposition, and replicative translocation (Nei, 1987; Ohno, 1970). Smaller duplications may be caused by slippage during DNA replication, but these probably only multiply small numbers of bases at a time, and can probably be ignored at the level of whole genes. Larger duplications of entire chromosomes can be caused by chromosomal non-disjunction, and the entire genome could be doubled by meiotic irregularities that produce gametes with unreduced chromosome number, leading to polyploidy. Some mechanisms of gene duplication are shown in figure 1.3, and an additional mechanism, replicative transposition, in figure 1.6.

Comparatively few studies have examined the causes of duplications, but we can speculate about the possible relative rates of the different mechanisms. Polysomy (duplication of an entire chromosome) may be relatively unlikely because large numbers of genes will be duplicated without their metabolic pathways, so gene dosage problems will frequently occur. The effects of these dosage problems lead to the multitude of symptoms of Down's syndrome, which is caused by trisomy of human chromosome 21, and trisomies of larger chromosomes are lethal in man. Similarly, the viability of partial polysomy mutations in *Drosophila* declines with increasing length of the duplicated segment (Li, 1997, p.270). It has also been argued that polyploidy would not be possible in organisms with chromosomal sex determination systems, such as birds and mammals, but a species of polyploid rat has been described (Gallardo et al., 1999), albeit one in which the sex chromosomes are the only chromosomes not duplicated – perhaps an exception that proves the rule.

1.4.3 Genome duplication

Genome duplication is a special case of gene duplication in which every gene in the genome is duplicated simultaneously – so that a diploid organism would become tetraploid. This creates a great amount of spare genetic material, and avoids a number of potential problems that could affect smaller-scale gene duplications.

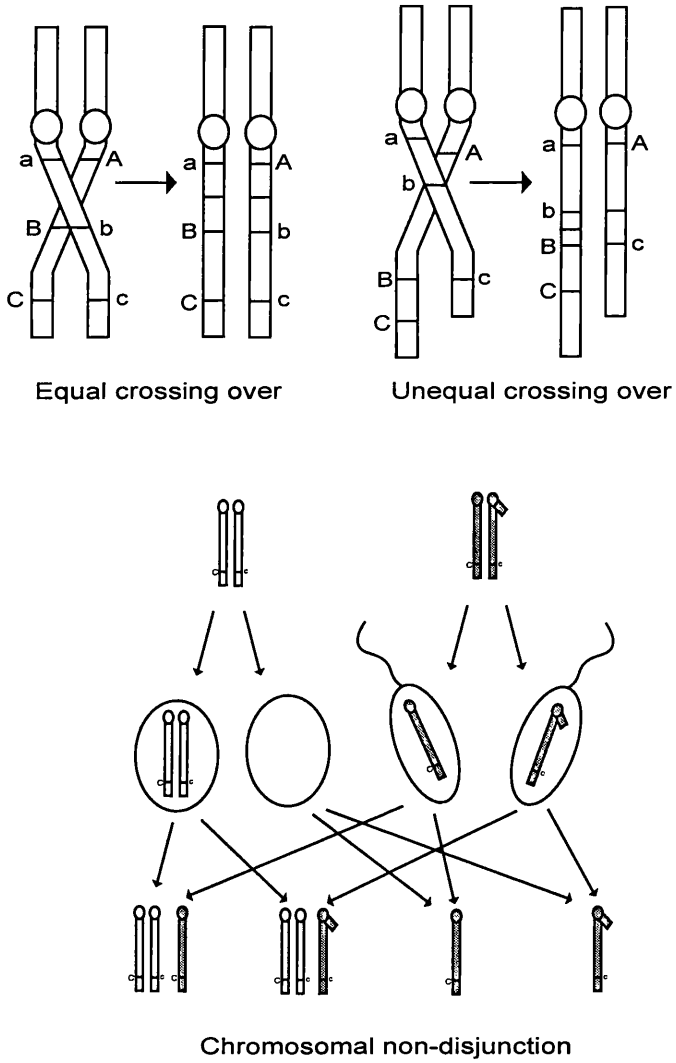

Figure 1.3: Unequal crossing-over and Chromosomal non-disjunction – two potential mechanisms of gene duplication. Unequal crossing-over occurs when homologous chromosomes pair incorrectly at meiosis or mitosis. Resolution of the chiasma will result in duplication of a locus on one chromosome and deletion of a locus from its homologue. Chromosomal non-disjunction occurs when a homologous pair of chromosomes fail to separate at metaphase, producing diploid gametes and trisomy in the zygote. Fusion of two diploid gametes can result in duplication of an entire chromosome pair.

Genome duplications would avoid the problem of dosage effects – as every gene in a metabolic pathway is duplicated, there will be no problems associated with the sudden doubling of transcription of some genes and not others. Secondly, genome duplication can also help populations avoid inbreeding problems and so survive population bottlenecks – in a tetraploid, the probability of a recessive phenotype is the fourth power of the frequency of the recessive allele, as opposed to the square of this frequency in diploids, so tetraploids may express far fewer deleterious recessive traits (Allendorf and Thorgaard, 1984; Li, 1980).

Genome duplication thus potentially supplies raw material for the development of new gene functionality on a great scale, which is probably a major reason why it was postulated that vertebrate genomes had undergone a doubling early in their evolution, associated with an increased complexity. This great increase in genetic material could also pose a problem for the organism, as Hughes (1999a, p.212) has suggested that the positive selection required for the evolution of new gene functions, if acting on a large number of newly duplicated genes, would impose so high a substitutional load that it would drive the population extinct.

Genome duplication can occur through two main mechanisms – *allotetraploidy* and *autotetraploidy*. In an autotetraploid, the genetic material of a genome is doubled, either by doubling of an individual genome or by crossing of two individuals of the same species producing a symmetrical genome with every gene now present in double the number of previous copies, and in exactly the same genetic context - with the same neighbouring genes, regulatory regions etc. Autotetraploidy makes a diploid genome into a tetraploid genome, so that every locus is segregating as four alleles. In most studied cases, it seems that a process of diploidisation has gradually taken place, as the tetrasomic loci have separated into two diploid loci that then diverge to produce two different copies of a gene. In an allotetraploid, two distinct diploid species have hybridised, so there are two similar but non-identical genomes are present. The loci will not show tetrasomic inheritance if the chromosome pairs are too dissimilar, so the genome will already be a duplicated diploid genome. An intermediate between these two situations, called segmental allotetraploidy, can occur if two very similar sister-species have hybridised. Theoretically, in such a situation, some loci could show disomic inheritance and some tetrasomic inheritance.

These three forms of inheritance have different consequences for the patterns

of relatedness of duplicated genes (Gaut and Doebley, 1997). For an autopolyploid, duplicate pairs should diverge from the onset of disomic inheritance, while duplicate pairs from an allotetraploid diverged at the divergence of the parental species' – potentially significantly older than the tetraploidy event itself. In a segmental allotetraploid, loci may become fixed for one of the parental alleles during the tetrasomic phase, before disomic inheritance begins and these identical alleles can diverge, or may preserve both parental forms. In the former case, divergence of the duplicated copies will date from the onset of disomic inheritance, otherwise it will date from the parental species' divergence. Under segmental allopolyploidy, paralogous loci duplicated in a single event could show quite different divergence dates.

These different mechanisms also have implications for the phylogenetic trees of genes multiplied in the genome duplication event – as Furlong and Holland (2002) have noted, if two allotetraploid-style genome duplications occurred in quick succession, before disomic inheritance was restored at every locus, a period of octosomic inheritance could occur. The resolution of these chromosome octuplets into tetrasomic quartets and disomic pairs would then take place, but under this mechanism, two rounds of genome duplication would not necessarily produce a symmetrical tree topology.

The most extensive literature on polyploidy is on plants, perhaps unsurprisingly, as one estimate suggests that as many as 70% of angiosperms may have experienced polyploidisation during their evolutionary history (Ramsey and Schemske, 1998; Soltis and Soltis, 1999). The importance of polyploidy in some plant populations' fitness is emphasised by the comparatively recent discovery that many polyploid plant species appear to have arisen multiple times from the same parent species, so polyploidy cannot be the evolutionary dead-end it was once considered (Ramsey and Schemske, 2002; Soltis and Soltis, 1995). Botanical data is also beginning to emphasise the rapid shifting of genetic material within polyploid genomes, both within and between the two diploid cohabitants. Traditionally, botanists have viewed allotetraploids as short-lived, with static genomes (Soltis and Soltis, 1999), but complete genomic maps for species like *Nicotiana*, *Avena*, *Brassica* and *Zea* show that the rate of genomic re-organisation in polyploids is significantly higher than in diploid genomes (Gale and Devos, 1998; Wendel, 2000). There is certainly a great deal of empirical evidence that genome duplication is also com-

mon across a wider range of organisms, but in the comparatively few cases where genomic-level sequence data are available to test for genome duplication a rather complicated pattern emerges.

The genome of at least one plant species – *Arabidopsis thaliana* – has been examined in detail to establish its polyploid nature. Despite this species being chosen for genome sequencing partly because of its supposedly compact genome (Meyerowitz, 2001), large internal repeats suggestive of a polyploid origin of the 125Mb genome were noticed by a number of authors (e.g. Lin, 1999; Terryn et al., 1999), which more detailed analyses have suggested arose from a series of major duplication events (Ku et al., 2000), which are probably whole-genome duplications (Vision et al., 2000), although the pattern is obscured by subsequent large-scale genome re-arrangements (Lin, 1999; McLysaght et al., 2000; Wolfe, 2001). There is probably less controversy over the suggestion that *Saccharomyces cerevisiae* has a polyploid past (Seoighe and Wolfe, 1999; Wolfe and Shields, 1997) but again, in this case, genomic re-arrangements have obscured the pattern (El-Mabrouk, 2000; Seoighe and Wolfe, 1998), and other interpretations are possible (Llorente et al., 2000).

1.4.4 Rates of gene duplication

Fairly little is known about the rate of gene duplication mutations. Part of the difficulty has, until recently, been the lack of comprehensive genome-scale information about genome structure, but there are more fundamental problems. Just as with point nucleotide mutations, only those duplications that are maintained will be observed, and there is a significant amount of data suggesting that duplicate genes are often selected against, and so will be rapidly lost. This makes inference about the actual rate of gene duplication (as opposed to the rate of maintained gene duplications) difficult, a point made explicitly by (Friedman and Hughes, 2003, pp.159-160). The rate at which gene duplications are maintained (and so can be observed) will be affected by selection – the number of genetic loci an organism can support is dependent on the mutational load – the fitness cost to the population of deleterious mutations. Mammals, with an average mutation rate of 10^{-5} per locus per generation (Kimura, 1983) can probably not support more than 100,000 genetic loci (Eyre-Walker and Keightley, 1999; Ohno, 1985).

One piece of evidence that points to a relatively high rate of gene duplication is that genome size is extremely labile – and especially in plants. Within a single genus, *Ranunculus*, estimated genome sizes vary from 5.8 pg[1] to 50.3 pg, and chromosome number varies from 16 to 108. Much of this variation is due to different ploidy levels, but within diploid species of *Vicia* the genome size varies from 3.4 pg to 27 pg. Variation within animal species is less marked, but still significant – the genome sizes of two different subspecies of the deer *Muntiacus muntjak* are 3,281 Mb and 2,521 Mb (Bennett and Leitch, 1995, Database of Genome Sizes - `http://www.cbs.dtu.dk/databases/DOGS/index.html`). There is also some data on the rates of polyploidy in plants, with the rate of autotetraploidy thought to be around 10^{-5}, and the rate of allotetraploidy significantly lower, but dependent on the frequency of interspecific hybridisation (Ramsey and Schemske, 1998).

Lynch and Conery (2000) presented the first study explicitly estimating duplication rates from genomic data, suggesting rates of around 0.0023 new duplicates per gene per million years in *Drosophila melanogaster*, 0.0083 for *Saccharomyces cerevisiae* and a substantially faster rate of 0.0208 *Caenorhabditis elegans*. Lynch and Conery (2000) suggest a high turnover of genes in eukaryotes, based on analysis of genes from the above species and human, mouse and *Arabidopsis* – they initially estimated that the half-life of duplicate gene copies ranged from 2.9 million years (for the two invertebrates) to 7.3 million years (for human and mouse). Lynch and Conery's paper was met with criticisms of both their data and methods (Long and Thornton, 2001; Zhang et al., 2001). This prompted them to slightly revise some of their estimates – for example, using a better-curated *Arabidopsis* genome sequence altered the estimated half-life of duplicates from this species from 3.2 million years to 32.4 million years, due to removal of allelic sequences and alternatively spliced forms of genes (Lynch and Conery, 2001). This is a striking demonstration of how bio-informatic analyses rely on well-assembled primary data.

It seems likely that gene duplication events vary in extent, and frequency: small tandem duplications may be quite common and larger sub-genomic to whole genome duplications appear to be rarer events. Evidence from the human genome

[1] pg is for picograms – genome size is usually estimated by fluorimetry, which estimates the mass of DNA per cell.

sequence suggests a rather different pattern, suggesting that most duplications are fairly large (>10 kb), but that there are duplicated blocks of almost every possible size (Lander, 2001). The same data also suggests a fairly continuous rate of duplicate formation – at least, duplicated blocks show a range of different sequence similarity. Other genome sequences from a range of organisms appear to show at least qualitatively similar patterns (The Arabidopsis Genome Initiative, 2000; Tomb et al., 1997; Venter, 2001). One recent analysis using human genome data has suggested a duplication rate of between 0.79 and 1.25 per million years across the whole genome (Gu et al., 2002).

1.4.5 Gene family evolution

The Darwinian paradigm of evolutionary change at the molecular level can be stereotyped as a model of gradual change due to nucleotide insertions, deletions and substitutions, but this model has difficulty in explaining the origin of new gene loci. The chance of a protein with a useful role in cellular processes evolving from an effectively random sequence seems vanishingly small. Duplication of existing genes, which already have functions, will produce new loci that are less constrained to perform a particular role, and might rapidly evolve to perform new functions, whether through gradual changes or through such processes as exon and domain shuffling. A group of such genes, related to one another, both in sequence and function, form a *gene family*. Given the gradual process of divergence over evolutionary time, the relatedness of genes within a family can vary. Dayhoff (1978) has suggested that genes with more than 50% similarity should be considered members of the same gene family, while related genes with less similarity than this should be grouped into a *superfamily*. This classification is convenient in this context – genes at these low levels of sequence similarity will be difficult to align and so difficult to investigate phylogenetically.

Gene families vary in both their pattern of phylogenetic relatedness and in their functional diversity. Figure 1.4 shows two different gene families, mammalian defensins and vertebrate lactate dehydrogenase. The gene duplications that have produced the diversity of mammalian defensins have occurred largely within the lineages leading to related species, leading to clades of defensins from each species. In contrast, ancient gene duplications produced the diversity of lactate dehydro-

genase (LDH), so that the two main clades on the LDH tree are different functional forms of the enzyme. The pattern shown by LDH, of different forms of the enzyme for different tissues, is common among enzyme gene families, and these different forms of the enzyme are known as *isozymes*. As is typical for isozymes, the two forms of LDH have different biochemical properties, differing in their affinity to NAD^+.

There is considerable interest in the existence of gene families from a number of different research areas. Scientists interested in protein structure use the diversity of gene families built around a common structural motif to understand how protein structure evolves, while molecular biologists are interested in identifying gene families to help identify potential functions for novel genes. It is notable that, despite the existence of a number of whole genome sequences for a variety of organisms, new families of proteins are still being discovered as rapidly as ever (Kunin et al., 2003). Early predictions (Chothia, 1992) that their would only be a limited number of gene families have proved wrong, underlining our still limited knowledge of genetic diversity. The very existence of families of paralogous genes provides powerful evidence for the importance of gene duplications. Figure 1.5 shows the size and number of gene families in vertebrates, confirming that gene duplications have indeed played a very powerful role in shaping genomes.

1.4.6 Fitness effects of gene duplication

One important function of gene duplication is the increase of the number of gene copies encoding a single function – with multiple copies of a gene, the transcription and translation machinery of the cell can produce a great abundance of a protein, and so increase the amount of the protein available. In fact, the best-known example of this is not a protein-coding gene at all, but the genes for cellular RNAs involved in protein synthesis. Genes for tRNAs and rRNAs can be present in many copies in the genome, enabling cells to very rapidly manufacture new ribosomes and new protein (Li, 1997, pp. 281). The multiple copies of RNA-encoding genes are, in general, kept identical by rapid gene conversion (Liao, 2000).

The most important consequence of gene duplication is the potential duplicate genes have for evolving new gene functions – genes do not arise de novo, but from modification of existing genes. It is difficult to see how an existing gene, care-

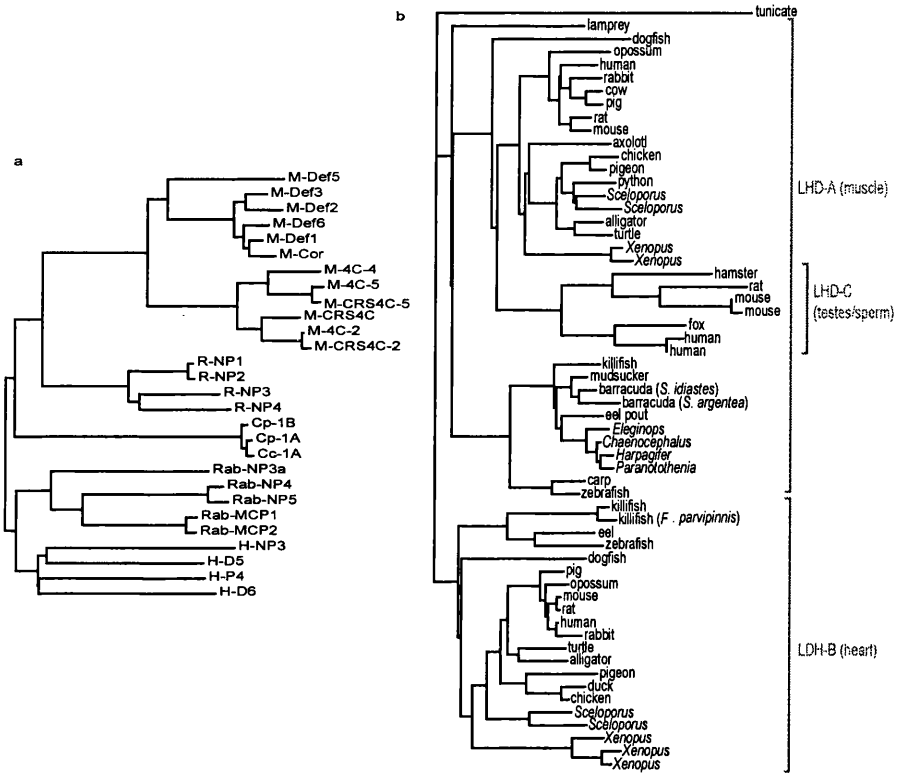

Figure 1.4: Two different gene families, differing in their pattern of diversification. (a) Mammalian defensins have diverged recently, leading to species-specific expansion of the family. Prefixes of the sequence names indicate species: Cc – *Cavia cutleri*; Cp – *Cavia porcellus*; H – human; M – mouse; R – rat; Rab – rabbit. (b) Lactate dehydrogenase diversified in early vertebrate evolution, producing cardiac and skeletal isozymes. A more recent duplication in tetrapods has produced the testis-specific isozyme of mammals. Part (a) redrawn from Hughes (1999a). Part (b) redrawn from Appendix A, figure A.3a.

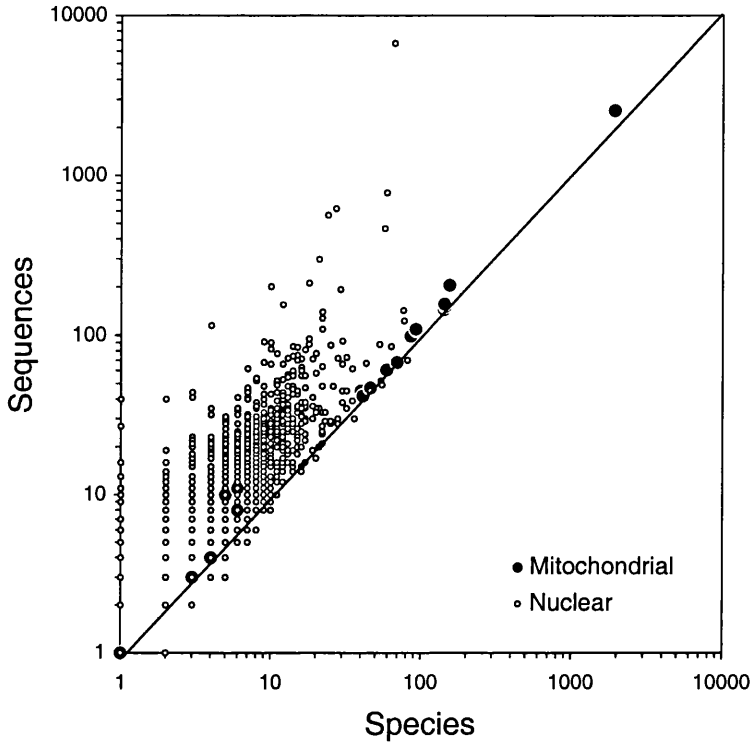

Figure 1.5: Number of sequences plotted against number of species for vertebrate gene families in release 29 (March 17, 1998) of the HOVERGEN (Duret et al., 1994) data base. Note that usually each species has a single mitochondrial sequence for a given gene (hence, the mitochondrial genes fall along the 1:1 line), whereas most nuclear genes are present in multiple copies. Due to redundancy in species names (for example, "human" and "*Homo sapiens*" being used to describe the source of different genes in the same family), some gene families appear to have fewer sequences than species. From Slowinksi and Page (1999, fig. 1).

fully adapted to a particular metabolic role, can be sufficiently free from selective constraints to evolve new function, but this is not the case with duplicate loci – one gene copy can continue performing an essential metabolic function while the second copy is free to vary, and to find a new function, whence positive selection will take over and finely tune the gene to its new place in metabolism. One copy of a duplicated gene is effectively free to wonder over the adaptive landscape of the genome until it finds a selective peak for some new function, which it will then climb and occupy.

This seems a convincing story, but there is a major problem – if the second copy of the gene is effectively 'spare' then the vast majority of non-silent substitutions in this gene will be neutral, and we might expect silencing mutations to be rather more common than the rare mutation moving the gene to the base of an adaptive "hill" – so the fate of the majority, and perhaps the vast majority, of duplicated loci will be oblivion – silencing followed by gradual expunging from the genome. A number of authors have visualised the early life of a duplicate locus as a race between fixing an advantageous mutation and fixing a null mutation, and there has been substantial interest in predicting how often these two fates occur, and so in establishing the true power of gene duplication to produce evolutionary novelty. This race is run particularly quickly in small populations – Watterson (1983) shows that the mean time until fixation of a null (nonfunctional) allele at one of two duplicated loci depends largely upon N_e, the effective population size, in large populations, and largely on the rate of mutation to the null state in small populations (see discussion in Li, 1997).

Walsh (1995) presents a population genetic model suggesting that, for large populations, 'new gene function, rather than pseudogene formation, is the expected fate of most duplicated genes', which would make gene duplication an impressively powerful mechanism for the evolution of novel biochemistry and novel developmental processes. Specifically, new functions are likely to evolve where $rS \gg 1$, where $S = 4N_e s$ and N_e is the effective population size, s is the mean selection coefficient of advantageous genes coefficient and r is the ratio of advantageous to other mutations. In fact, this model is likely to underestimate the rate of evolution of new gene functions, as it assumes that all non-advantageous mutations are neutral, where in reality many will be more or less deleterious. Ohta (1989) admits that 'gene duplication could well have been the primary mechanism

for the evolution of complexity in higher organisms', and presents models for the origin of 'gene families with diverse functions', concluding that natural selection should favour those genomes with more favourable mutations occurring in duplicated genes, so there should be selective pressure favouring mechanisms of gene duplication. Ohta has also presented a number of other simulation studies on the evolution of large gene families (Batson and Ohta, 1992; Ohta, 1987, 1988a,b) that support the likelihood of this model.

An alternative model has been suggested a number of times (Hughes, 1994; Li, 1980), which highlights a third possible fate for a pair of duplicated genes – rather than a locus either gaining a new function or being lost, the genes can each share part of the function of a pleiotropic parent gene, so that the gene functions become specialised. This model has been termed subfunctionalisation or 'duplication-degeneration-complementation' (Force et al., 1999), as it requires the two different loci to each mutate at least once, and for these mutations to be complementary. Gene conversion adds new complications for all of these models – while gene conversion may prevent, or slow, the divergence of two duplicated loci to form new functional genes, it may also prevent a gene becoming neutralised by a null mutation, or even resurrect a 'dead' gene copy (Li, 1997, p.333).

Theoretical studies have shown that gene duplications may be relatively likely to lead to new gene functions, and to increase the fitness of genomes in which they occur. Empirical studies (such as Nadeau and Sankoff, 1997) may suggest that the evolution of new functions is even more common than theoretical studies suggest, but there are a number of difficulties with the empirical work (Wagner, 1998). One interesting example of a duplicated gene acquiring a new function is *jingwei*. The high ratio of non-synonymous to synonymous substitutions (dN/dS ratio) after the duplication of *jingwei*, both before and after the divergence of the two *Drosophila* species (shown on figure 1.6), suggests that positive selection has acted on this gene to evolve a function distinct from that of its parent locus, *Adh*. This is reinforced by evidence that *jingwei* is expressed, but its function is currently unknown (Hughes, 1999a).

Another possible advantage of gene duplication is genetic redundancy – if multiple genes are capable of performing a particular metabolic role, an organism is robust against silencing mutations in one of these genes. A small-scale study on yeast has suggested that duplicate genes play little role in genetic robustness against

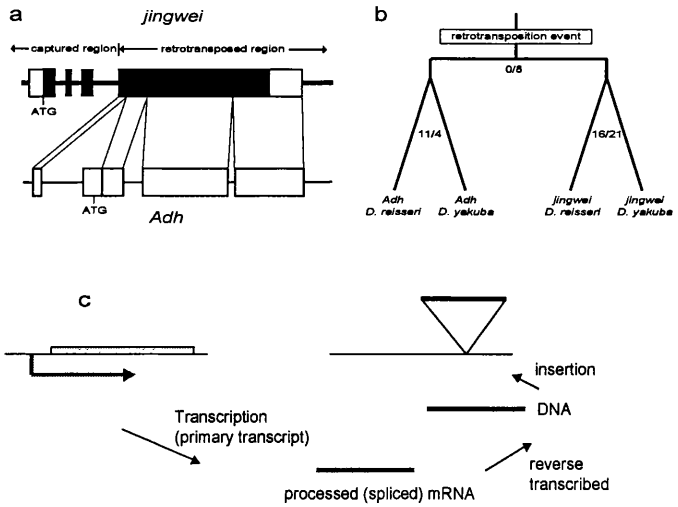

Figure 1.6: *Jingwei* has evolved from a retrotransposed copy of the *Adh* gene. The process of retrotransposition (c) produces gene copies identifiable by their lack of introns and regulatory regions removed during mRNA processing. *Jingwei* has also captured the 5' region of an unknown gene. (b) *Jingwei* duplicated sometime before the divergence of *Drosophila reisseri* and *Drosophila yakuba*. Figures on the branches are the numbers of synonymous and non-synonymous substitutions between the two species and between *adh* and *jingwei*. Parts (a) and (b) redrawn from Hughes (1999a), data from Long and Langley (1993).

gene silencing (Wagner, 2000). The availability of a nearly complete set of single-gene deletion mutants for *Saccharomyces cerevisiae* has enabled more recent work to estimate that at least a quarter of non-lethal deletions are compensated by duplicate genes (Gu et al., 2003). A rather different source of evidence suggests similar levels of redundancy from alternate biochemical pathways and duplicate genes (Kitami and Nadeau, 2002a,b). These results are particularly interesting in suggesting a possible fitness advantage maintaining duplicate genes in the absence of functional divergence.

1.5 The 2R hypothesis

There has also been considerable interest in genome evolution during the origin of vertebrates, particularly focusing on how gene duplications have produced the larger genomes of vertebrates. Ohno (1970) suggested that at least one whole-genome duplication occurred early in vertebrate evolution. This idea later became formalised as the '2R hypothesis', stating that two tetraploidisation events occurred sometime during the origin of higher vertebrates, so that the presence of multiple copies of many genes in vertebrates is due to duplication of the whole genome (Holland et al., 1994), prompted by the discovery of four *Hox* gene clusters in vertebrates compared to the single cluster of most invertebrates (Garcia-Fernandez and Holland, 1994).

If this seems unlikely, it is worth noting that both yeast (Wolfe and Shields, 1997) and maize (Ahn and Tanksley, 1993; Helentjaris et al., 1988) appear to be fairly recent degenerate tetraploids, and we would expect fairly little evidence of genome duplications so ancient to have survived. The fact that significant numbers of gene duplications have taken place during the evolution of higher vertebrates seems beyond doubt, but there is much debate over whether two rounds of whole-genome duplication best explains this. A number of empirical studies have attempted to unravel the picture (Hughes, 1999b; Martin, 1999a; Suga et al., 1999; Wang and Gu, 2000). Some reviews of this work have concluded either that there is still insufficient data to decide the question (Skrabanek and Wolfe, 1998), or even simply confirm that it's a very hard problem to tackle (Smith et al., 1999). One additional difficulty is that there has been much debate about when the two rounds of polyploidisation occurred (figure 1.7). Recent evidence suggests that one round

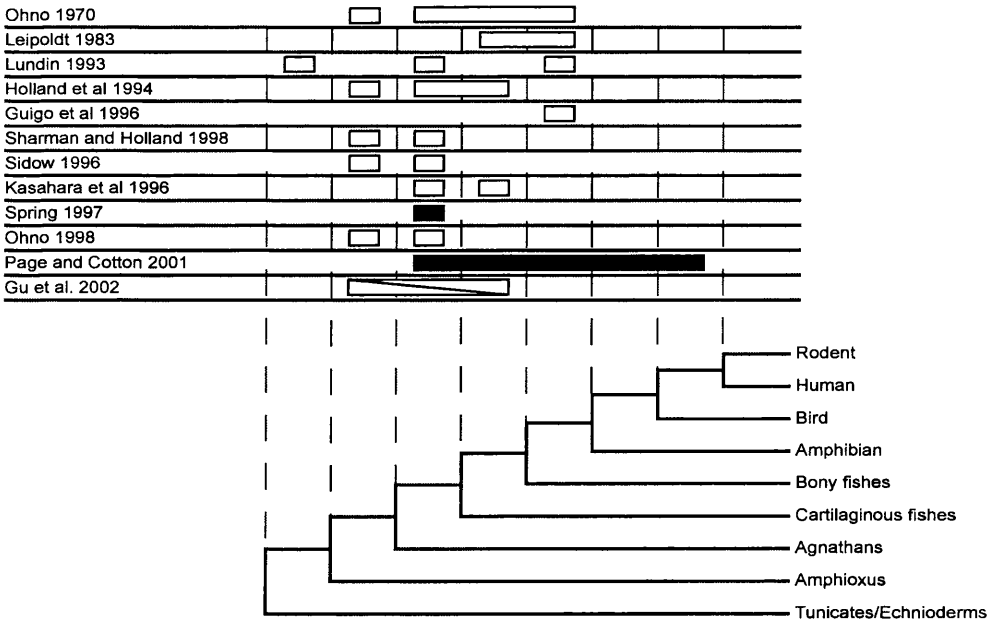

Figure 1.7: Suggested timings of genome duplications in vertebrate evolution. Different authors have disagreed about the timings of possible genome duplications in vertebrate evolution. Open boxes represent proposed timings of genome duplications, with extended boxes representing uncertainty in the timing. Shaded boxes represent suggested episodes of accelerated small-scale duplication. Gu et al. (2002) do not decide between one or two rounds of genome duplication. Modified from Martin (1999a); Skrabanek and Wolfe (1998).

occurred before the lamprey/gnathostome split and one after (Escriva et al., 2002).

1.5.1 Testing the 2R hypothesis

Genome sizes - physical and number of genes

Much of the evidence used by Ohno (1970) to support the original suggestion of something like the 2R hypothesis focused on differences in genome sizes and chromosome numbers. Today, we know that genome sizes are largely mediated by changes in the amount of non-coding DNA, and that they can be very fluid indeed, as discussed in section 1.4.4. It is hardly surprising that Ohno lacked much supporting evidence – very few gene sequences were known in 1970. Instead, Ohno

largely relied on arguments that small-scale tandem duplication was insufficient to produce the amount of additional genetic material observed in vertebrates. He argues that tandem duplications would be less effective than polyploidy because they would be more likely to cause deleterious gene dosage effects, would not duplicate regulatory elements and particularly, because tandemly duplicated regions would, in turn, encourage a higher rate of unequal crossing-over. Ohno envisioned a run-away process of more and more repetitive duplication (see Ohno, 1970, pp. 94-97), which seems remarkably prescient in the light of recent data that around 50% of the human genome comprises repetitive DNA, although the vast majority of this is transposon-derived (Lander, 2001).

One-to-Four rule

Spring (1997) proposed that it was possible to test the 2R hypothesis by checking its prediction that every gene present before two consecutive genome duplications would be present as four copies afterwards. Although subsequent gene loss will have lowered this figure, Spring claimed that the maximum ratio of human genes to their Drosophila orthologues was four. Furthermore, Spring found his 'tetralogues' on all 23 human chromosomes, supporting his idea that this pattern could only come from whole-genome events. Spring did find several gene families with other patterns than simple 1:4, but considered that more complete sequence data would split these families into simple tetralogue groups. More recent examinations of this idea, using the complete gene complement of *Caenorhabditis*, *Drosophila* and human have shown that there is, in fact, no excess of gene families showing Spring's 1:4 ratio than would be expected by the slightly larger genome of humans, and there are certainly a number of gene families where the ratio of human to invertebrate members exceeds the 1:4 maximum expected by Spring (Lander, 2001; Venter, 2001). This is, in fact hardly surprising – just as Spring admitted that gene loss following a genome duplication would affect this ratio, so would independent, smaller-scale gene duplication events. The interaction of subsequent gene loss and gene duplications could easily have erased any 1:4 signal dating from the ancient genome duplication events, if they occurred. The ratio of gene family members is simply not a sufficiently powerful statistic to test the 2R hypothesis.

Paralogous segments

Another major approach to testing the 2R hypothesis has been to look for the paralogous regions of vertebrate genomes that would be expected to be left by two successive doublings of the genome - some portions of some chromosomes should remain in four copies. The major difficulty of this approach is that genome rearrangements are very common – processes like inversions, transpositions and reciprocal translocations can shuffle genes around the genome, which will break up the quadruplicate pieces into smaller and smaller fragments. These small fragments could then as easily be the remnants of smaller-scale duplication events – as (McLysaght, 2001, p.30) admits 'finding as few as two genes in several linked clusters in a genome of over 30,000 is hardly overwhelming evidence for a genome duplication event'.

Initial evidence based on genetic maps was largely based on simply finding a number of related genes on 4 different chromosomes. Most famously of all, the four vertebrate Hox clusters are present on human chromosomes 2,7,12 and 17, with only a single similar cluster present in *Amphioxus* (Garcia-Fernandez and Holland, 1994), and a number of other genes co-occur on these chromosomes (Hughes et al., 2001). A number of genes around the MHC (major histocompatability) locus are found on human chromosomes 1,6,9 and 19 (Kasahara et al., 1996, 1997) and a single related cluster occurs in *Amphioxus* (Flajnik and Kasahara, 2001). More recent work has produced a great deal of additional evidence supporting the idea that the MCH region has duplicated twice in early vertebrates (Abi-Rached et al., 2002). Pebusque et al. (1998) have claimed that human chromosomes 4,5,8 and 10 form a similar set. Some rather dubious examples have been proposed – Gibson and Spring (2000) have claimed a relationship between human chromosomes X, 4, 5 and 11 based on evidence from only two gene families.

Most of these chromosome relationships have come in for criticisms. Phylogeneticists in particular have objected to much of the evidence from duplicated paralogous segments – McLysaght (2001) has described the debate over the 2R hypothesis as 'a war of words between the phylogeneticists and the cartographers', but this something of a simplification – phylogenetic methods and map-based methods are complementary, and a number of studies have begun to integrate both sources of data. In any case, more sophisticated map-based studies have become available – McLysaght et al. (2002) use the idea that, if genome duplication gave

rise to the ancestral vertebrate genome, there should be numerous pairs of contiguous blocks of duplicated genes that are the remnants of this process. They convincingly show that the pattern of these blocks in the human genome are more likely to have come from a 2R-style event than from individual gene duplications, but cannot exclude the possibility that regional duplications on a large, but subgenomic, scale could have produced the observed pattern.

Tree topology

One prediction of the 2R hypothesis is that vertebrate gene families will show a particular symmetrical tree topology (see figure 6.4d), caused by the two successive genome duplication events. This prediction has been widely used to test the 2R hypothesis using gene family phylogenies.

The earliest of these studies (Zhang and Nei, 1996) showed that the Hox clusters duplicated early in vertebrate history, but there was insufficient resolution to fully resolve the phylogeny of these genes beyond grouping HoxC and HoxD. A later phylogeny of the linked fibrillar-type collagen gene, however, supported a different grouping of the chromosomes carrying HoxB with that carrying HoxC, with the HoxA-linked gene forming a clade with these two to the exclusion of HoxD (Bailey et al., 1997), contradicting Zhang and Nei's analysis and not supporting the 2R pattern. Further work by Hughes et al. (2001) looked at 35 gene families with members on at least two of the Hox-bearing chromosomes, and found that only eight of these families had divergence times compatible with the duplication of the Hox clusters. Those families with members on three of the chromosomes disagreed on the phylogenetic history of the chromosomes, a result which Hughes et al. (2001) claimed rejected the 2R hypothesis, although Hughes et al.'s conclusion has been questioned recently by Larhammar et al. (2002).

Other studies have also focused on questioning claims about tetralogous relationships – showing that most of the gene families showing the four-to-one pattern do not display the expected, balanced topology (Hughes, 1999b; Martin, 2001), and questioning the relationships between the MHC-bearing chromosomes (Hughes, 1998). More extensive tests were possible following the availability of the complete genomes of *Saccharomyces cerevisiae* and *Drosophila melanogaster* as outgroups and the human genome sequence. These tests conclude that tree topologies for four-member families of human genes do not show the symmetrical pattern pre-

dicted by the 2R hypothesis (Friedman and Hughes, 2001). More intriguingly, the same authors (Friedman and Hughes, 2003) used tree topology to claim that there is no excess of highly conserved human gene families duplicating around the time expected for the '2R hypothesis' – in marked contrast to molecular-clock results (Gu et al., 2002).

Most recently, Furlong and Holland (2002) present a detailed review of previous attempts to test the 2R hypothesis, particularly focusing on phylogenetic tests. They include a number of additional gene families not considered previously, and conclude that the predominance of 1:2, 1:3 and 1:4 is 'entirely congruent' with the 2R hypothesis. Despite their use of explicit phylogenies for each gene, they ignore tree topology except to ensure the monophyly of vertebrate genes used. Furlong and Holland (2002) cast doubt on all of the tree-topology dependent methods used previously by arguing that, if vertebrates underwent a period of octoploid inheritance after two rounds of tetraploidisation, we would expect unbalanced or 'sequential' tree topologies for many loci. This argument relies on the two genome duplication events occurring in reasonably quick succession – certainly before diploidisation is complete following the first event, although Allendorf and Thorgaard (1984) report that 'residual tetraploidy' is observable in salmonid fish over 25 million years after the tetraploidisation event. Furlong and Holland conclude by stating that 'paralogy regions, asymmetrical tree and non-congruent linked trees are all compatible with two sequential rounds of autotetraploidy'. This is true, but all of these observations are also compatible with segmental duplications – Furlong and Holland's argument is not decisive in favour of the 2R hypothesis, but does urge caution in interpreting the results of phylogenetic studies.

Molecular clocks

As has already been mentioned, along with topological information - the expectation of a symmetrical tree topology, phylogenetic branch length information has also been used to question the 2R hypothesis. We would expect all genes that duplicated simultaneously to show compatible ages, and that the more recent duplications within a phylogeny should be simultaneous (Hughes et al., 2001; Martin, 2001). By using external calibration points, absolute dates can be used to reveal the entire pattern of gene duplication in a lineage, as has been attempted by Gu et al. (2002). Molecular clock estimates have also been used as supporting evidence in

map-based studies, for example by both Wolfe and Shields (1997) and McLysaght et al. (2002).

Despite over 15 years of intense research interest, and the availability of the complete sequence for the human genome, it is remarkable that even the most enthusiastic 2R believer (Spring, 2002) can only lament 'why is it so difficult to prove the obvious?'.

1.6 Phylogenetic consequences of gene duplication

In most molecular phylogenetic analyses it is assumed that the phylogeny of the genes analysed exactly parallels that of the organisms they are sampled from, so that the gene phylogeny or 'gene tree' is exactly the same as the 'species tree'. If the organisms represent different reproductively isolated populations, this assumption is met if the molecular sequences used in the analysis are orthologous – if they represent the same locus sampled from each organism. Gene duplication produces similar copies of a gene, members of a gene family. These copies are paralogues, begin related by a gene duplication event rather reflecting the relationship between species. Trees which include some paralogous sequences may not reflect the evolutionary history of the organisms, but the evolutionary history of the genes themselves (see chapter 4, figure 4.1). This can pose a serious problem to systematic biologists.

Although some authors (e.g. Brower et al., 1996) have claimed that problems of paralogy are of relatively little importance, and even that they can be overcome with sufficient 'weight of evidence' from multiple genes, there is no theoretical or empirical reason to think that this is the case (Slowinski and Page, 1999). Indeed, many of the more 'unconventional' results of molecular phylogenetic studies with nuclear genes may be due to paralogy. While certain features of sequences, such as intron structure and flanking regions can help distinguish orthologs and paralogs (e.g. Small and Wendel, 2000; see Sanderson and Shaffer, 2002), few authors make serious attempts to ensure the orthology of the gene sequences they use.

Most of the earliest molecular phylogenies were based on ribosomal RNA gene sequences (Woese, 2000; Woese and Fox, 1977), which are still very widely used. The many copies of rRNA genes are kept relatively uniform within the genome by frequent gene conversion events, and so do not suffer from paralogy, and they are

ubiquitous and easy to extract and sequence. More recently, mitochondrial genes have become a marker of choice – they are generally thought to be single-copy, and have a number of other properties that should make them very valuable for phylogenetics (Moritz et al., 1987). There may, however, be frequent sequence duplication (Broughton et al., 1998), or even recombination (Eyre-Walker and Keightley, 1999) in these genes, although there is debate over the evidence for the latter (Eyre-Walker et al., 1999; Macaulay et al., 1999).

There are, however, many important phylogenetic issues that are not resolved by these two genes, due either to these loci evolving at an inappropriate rate or being too short to provide sufficient evidence. Whole mitochondrial genomes are of the order of 16,000 bases long and the largest ribosomal RNA genes (23S) are under 3,000 bases. Nuclear, protein-coding gene sequences represent an enormous and ever-growing resource for phylogenetic reconstruction; there are as many as 585 genomic sequencing projects underway[2]. As we have seen, however, many nuclear genes are likely to show extensive gene duplication. The use of the ever-growing amount of nuclear gene sequence data to infer phylogenies will depend upon rigorous methods for dealing with paralogy. Reconciled trees represent one such method.

1.7 Reconciled trees

One natural way to consider the evolution of genes is to think of the gene lineages evolving independently within a species lineage. This leads naturally to considering the gene tree – the phylogeny of the gene sequences – as distinct from the phylogeny of the species the genes occur in (the species tree). A simple example is shown in figure 4.1. Understanding the evolution of a gene then becomes a problem of understanding the relationship of two associated trees, a problem of significant interest to systematic biologists. Similar problems occur in understanding the evolution of a number of associated systems – parasites and their hosts, organisms and the areas they inhabit and even languages and the people who speak them (Page and Charleston, 1998; Penny et al., 1993). These similarities have resulted in a fertile transfer of ideas between different disciplines (Page, 2003).

[2] in 350 prokaryotes and of 235 eukaryotes, including EST surveys; `http://www.ebi.ac.uk/research/cgg/genomes.html`; Kyrpides (1999).

The earliest quantitative attempts to solve these problems were pattern-based, coding the associate tree (i.e., the parasite or gene tree) as binary characters, which could then be used to infer the host phylogeny or could be optimised onto an assumed host phylogeny in an attempt to understand the evolution of a group (BPA – Brooks, 1981, 1990). There were a number of serious problems with this solution (Page, 1993a; Ronquist and Nylin, 1990), for example, BPA can produce results suggesting that associates travel back in time to infect hosts during transmission events, and there are other problems with interpreting the results of BPA analyses (Page, 2002b). These difficulties prompted the development of event-based methods, which attempt to explain the difference between the two associated trees in terms of the actual events that produced these differences (Ronquist, 2003).

Event-based methods consider that the phylogeny of the associate tracks the phylogeny of the host, but the fidelity of this 'tracking' depends upon how often events such as duplication, horizontal transfer and lineage sorting occur in the associate's evolution (Page and Charleston, 1998). In fact, only four such events need to be considered separately – cospeciation, duplication, lineage sorting (or extinction) and host switching (figure 1.8). These events will introduce differences between the trees that describe the hierarchy of the two entities, as in the example shown in figure 4.1), where a duplication in the gene tree and three gene losses (a sorting event) explains the difference between a gene tree and a species tree.

Reconciled trees were the earliest such method. Reconciled trees were first used to investigate the history of a gene family when Goodman et al. (1979) introduced the concept in investigating the evolution of globin genes in mammals. Page (1988, 1993b) recognised the analogy between Goodman et al.'s genes and host-parasite systems (table 1.1), leading him to formalise the concept of reconciled trees (Page, 1994a). The original method included only a subset of cophylogenetic events, ignoring host switches or lateral gene transfer – Page (1994b) later attempted to generalise the method to include this event, somewhat unsuccessfully. It turns out that dealing correctly with host switching is a rather difficult problem (Charleston, 1998; Charleston and Perkins, 2003; Ronquist, 2003). Reconciled trees have since been used extensively to investigate relationships between organisms and areas in biogeography (Linder and Crisp, 1995), but are now perhaps most often employed for studying coevolution between associated organisms, in associations as diverse as those between lice and their seabird hosts (Patterson

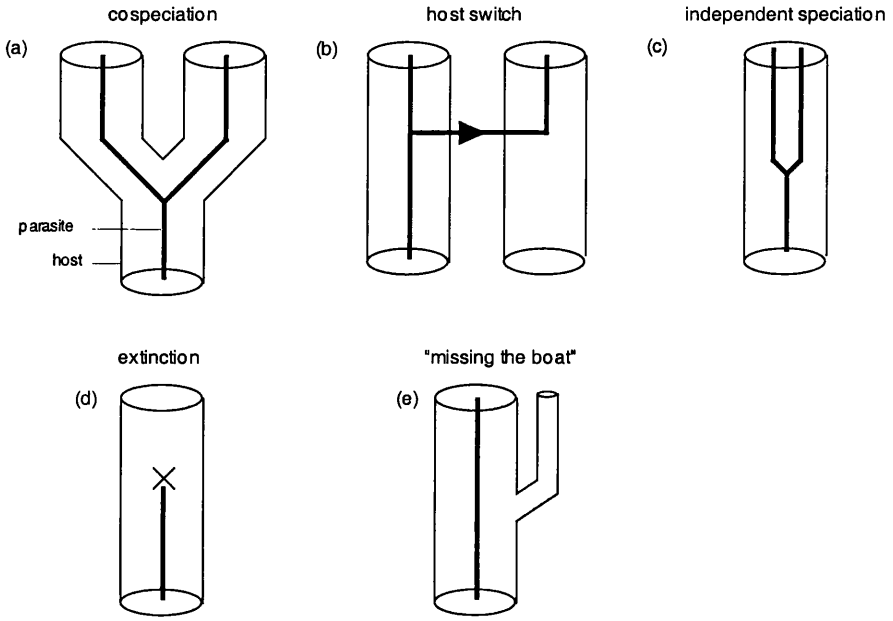

Figure 1.8: Possible events in a host-associate system. The host and associate may cospeciate (a), or the associate may speciate/diverge independently of its host (b,c), in which case the descendants my remain on the same host (b), or colonise a new host (c). Absence of an associate may be due to extinction of the associate (d), or due to a host lineage not inheriting the associate (e). Taken from Page (2003), figure 1.

Host-associate	Codivergence	Duplication	Horizontal transfer	Sorting event
Organism-gene	Codivergence	Gene duplication	Lateral gene transfer	Gene loss or deletion
Host-parasite	Cospeciation	Within-host speciation	Host-switch	Parasite extinction
Organism-area	Vicariance	Sympatry	Dispersal	Extinction

Table 1.1: Equivalent events in different historical associations. Modified from Page and Charleston (1998).

et al., 1993), between beetles and their plant hosts (Becerra and Venable, 1999) and between retroviruses and their hosts (Martin et al., 1999). The last example shows how blurred the line between investigating independently evolving associates and looking at events of molecular evolution can be.

The event-based nature of reconciled trees has another benefit crucial for the work in this thesis – by postulating biologically important events occurring along branches to explain incongruence, reconciled trees can be used to study these events themselves. An important aim of this thesis is to investigate the utility of reconciled trees in studying gene duplication, as well as in phylogenetics. Reconciled tree methods for comparing gene and species trees are implemented in the program GENETREE (Page, 1998), a software package that has continued to evolve over the course of this PhD project (Page and Cotton, 2000, Appendix A). Much more discussion about reconciled trees and gene tree parsimony is found throughout this thesis.

Part I

UNDERSTANDING GENE TREE PARSIMONY

Chapter 2

Tangled Tales from Multiple Markers

Reconciling Conflict Between Phylogenies To Build Molecular Supertrees[1]

Abstract

Supertree methods combine information from multiple phylogenies into a larger, composite phylogeny, resolving any conflict between them. On the other hand, there are many approaches to combined analysis of different data sets for similar or identical taxon sets, a subject that has been of interest to systematists over a long period of time. Gene tree parsimony is a method, related to supertree techniques but having a different conceptual background, which can combine data from molecular phylogenies for overlapping taxon sets and interprets conflict between these phylogenies in a biologically meaningful way. We review the method and discuss the relationship between gene tree parsimony and supertree methods.

[1]This chapter is currently in review for a forthcoming book on supertree methods, edited by Olaf Bininda-Emonds and to be published by Kluwer. It was co-authored with Rod Page.

2.1 Introduction

Combining data from different sources of phylogenetic evidence can be important for two different reasons – to increase the scope of the phylogenetic results by including a greater range of terminal taxa, or to improve the accuracy of the results by incorporating more data for these taxa. Supertrees seek to address both of these questions. Where source trees are rooted and compatible, supertree construction is relatively trivial – efficient algorithms exist to decide whether or not a set of trees are compatible and to construct a parent tree that contains all of these trees (Aho et al., 1981), to produce all of the possible parent trees (Ng and Wormald, 1996), to produce all of the minimally resolved parent trees (Semple, 2003) and to directly produce the strict consensus of these trees from the input trees (Steel, 1992). Unfortunately, the situation is far more difficult for unrooted source trees (Gordon, 1986; Steel, 1992; Steel et al., 2000).

However, most practical applications of supertree methods involve source trees that are incompatible, and supertree workers have been less successful in designing algorithms to combine information from conflicting trees. Such algorithms seek to either remove conflict by pruning trees (e.g. maximum agreement subtrees), represent the conflict through soft polytomies, resolve the conflict, or some combination of these.

In fact, the only supertree method that has been at all widely used by biologists is Matrix Representation with Parsimony (MRP, Baum, 1992; Ragan, 1992), with an increasing number of supertrees constructed using this method appearing in the literature (see Kennedy and Page, 2002; Pisani et al., 2002; Salamin et al., 2002, for three recent examples). MRP uses additive binary coding to represent the hierarchical structure of trees as a series of matrix elements - each node on the trees is represented by a column of the matrix, with missing data for those taxa not present on a particular source tree. This matrix is then analysed using maximum-parsimony methods to construct a single supertree. While MRP supertrees have played an important part in stimulating the field of supertree research, and may be reasonably successful in reconstructing relationships (Bininda-Emonds and Sanderson, 2001), there has been an increasing literature on the biases of MRP methods, and a similar number of proposed modifications to the original method (e.g. Bininda-Emonds and Bryant, 1998; Purvis, 1995b; Ronquist, 1996; Thorley,

2000). There are similar problems with other supertree algorithms too, such as the mincut supertree method (Semple and Steel, 2000), which has a number of undesirable properties (Page, 2002a). These problems have prompted a widening interest in other methods of supertree construction, such as shown in this volume and elsewhere (Page, 2002a).

2.2 The distance view of the supertree problem

In an effort to classify the growing number of supertree methods available to systematists, several authors have characterised the supertree problem in a distance framework (Chen et al., 2003; Lapointe and Cucumel, 1997; Thorley and Wilkinson, 2003). Both of these authors suggest that the supertree problem be seen as the problem of finding a tree, or set of trees, that minimises the distance from a set of input trees, under some measure of distance between trees. For example, as both sets of authors point out, MRP seeks to find the tree minimising the number of steps required on the MRP matrix. Other distance measures are certainly possible – such as distances based on nearest-neighbour interchanges (NNIs, Waterman and Smith, 1978). It has been suggested that the distance measure must be a metric (Thorley and Wilkinson, 2003), but we disagree – the supertree problem is inherently asymmetric, in that the supertree is more inclusive than its subtrees, so there seems no reason to require the distance used to be a metric. Bearing in mind this framework, we should note that any heuristic tree search is likely to be NP-complete (Wareham, 1993), including the maximum-parsimony problem used by MRP methods (Graham and Foulds, 1982).

We suggest a new distance measure for supertree inference, one based on the number of actual biological events that may have produced the differences observed between source trees. These events can be inferred using a co-phylogenetic method called reconciled trees.

2.3 Tangled trees, or cophylogeny

Evolutionary biologists have long been interested in the relationship between ecologically associated entities, particularly hosts and their parasites. One important question in host-parasite biology is the extent to which these organisms co-evolve,

and more specifically the extent to which they co-diverge – the extent to which speciation events in one lineage are mirrored by speciation events in the other. This lead to interest in comparing the phylogenetic trees of associated organisms, along with a parallel interest in relating the phylogenies of organisms to their biogeography (Page and Charleston, 1998). The most obvious solution to the problem was to use a binary coding of the dependent tree, similar to those used in MRP supertree methods. This matrix was then used either to reconstruct the host phylogeny, or to understand the pattern of evolution by optimizing the characters onto the second phylogeny (Brooks, 1981). Similarly to the problems with the binary coding used in MRP, various fixes failed to alleviate the fundamental problem that such characters are non-independent.

In cophylogeny, the solution has been to explicitly map the dependent phylogeny into the host phylogeny, directly postulating events that lead to the differences between the two phylogenies. This insight led to Page's 1994 formalisation of Goodman et al's 1979 idea of a reconciled tree – we can reconcile the differences between two trees that we would expect to be identical by postulating certain cophylogenetic events introducing differences. As shown in figure 2.1 these events can be extinction of a lineage, independent speciation of a lineage and horizontal transfer. While co-phylogeny methods were developed in the context of biogeography and host-parasite evolution, similar events occur in the evolution of a gene lineage within a species – lateral gene transfer, gene duplications and gene loss, so the same cophylogeny mapping can also be used to study this system.

The interest in supertree methods underlines the growing availability of reliable phylogenies, and this increasing amount of data reflects both an increase in width – in the taxonomic coverage of phylogenetic information – and in depth – in the amount of data available for particular organisms. This increasing depth is particularly due to the rise of genome-level sequencing efforts for an increasing number of organisms, and an important corollary of this work is the increasing realisation that phylogenies for different genetic loci for the same species frequently disagree, and the realisation that evolutionary events can cause the correct phylogeny for a gene to be different from the correct phylogeny for the species it is sampled from – a problem known as the gene tree-species tree problem (Doyle, 1992; Maddison, 1997). Reconciled trees are a natural solution to this problem (Page and Charleston, 1997a) – we can use the reconciled tree algorithm to score a species tree for

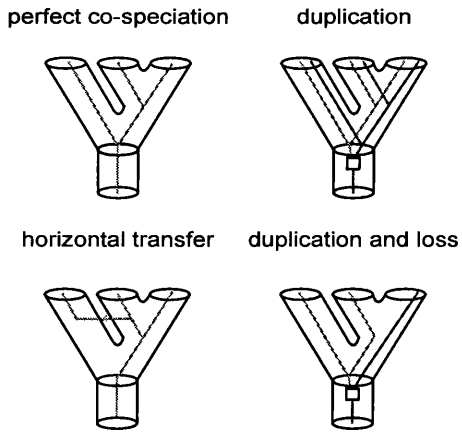

Figure 2.1: Some co-phylogenetic events, introducing differences between two associated phylogenies.

a particular gene tree in terms of the number of gene duplications, gene losses and other evolutionary events that have introduced differences between the two trees. This is a distance between the trees that has a natural, biological interpretation (Mirkin et al., 1996).

In principle, any of the events in figure 2.1 can be scored in this way, but it should be noted that dealing with horizontal gene transfer is complex, and existing implementations of reconciled trees in this context exclude this possibility (Page, 1998). In particular, dealing with horizontal transfer adequately is far more computationally intensive and requires us to make some assumption about the relative rates of gene duplication and loss and lateral gene transfer (Charleston, 1998). It is also often preferable to use the count of duplications alone (ignoring gene losses) as a distance function, because in some kinds of study, gene losses are confounded with failure-to-sample (simply the lack of a sequence in the sequence databases), and so do not represent a true biological cost.

2.4 From reconciled trees to supertrees

When we have multiple source gene trees, we can combine information from a number of these trees into a single species tree estimate by finding the species tree (or set of species trees) minimizing the number of co-phylogenetic events required

to reconcile the species tree with each source tree, or minimizing some weighted sum of these events (assigning a cost to each event category). The resultant species tree can be on a larger taxon set than any of the source trees, and is constructed using information from the topology of each source tree only, and so fits the definition of a conventional supertree. This method of combining data using reconciled trees has become known as "gene tree parsimony" (Slowinksi and Page, 1999).

If we restrict the source trees to be molecular trees, the duplication count (or duplication cost) is a biologically interpretable measure of the evolutionary difference between the source tree (or gene tree) and supertree (or species tree). By dealing with only gene trees, we have a better idea of what processes might introduce incongruence, and so can deal with this incongruence in a biologically attractive way. This contrasts markedly with other supertree methods – most supertree authors write off incompatibility between source trees as error that cannot be further dissected. If this noise is due to estimation errors, it could be further understood by reference to the character data underlying the source trees, but this is generally unavailable (or ignored) in supertree construction.

Finding an optimal species tree under either the duplication-only or duplication-and-loss score has been the focus of some attention by mathematicians and computational biologists. Linear-time algorithms exist for computing these scores for a particular pair of gene tree and species tree (Eulenstein, 1997; Zhang, 1997; Zmasek and Eddy, 2001), and while it is known (as expected) that finding the minimum-cost species tree is NP-complete (Ma et al., 1998), there is a fixed-parameter tractable algorithm to find this tree without heuristic searches of tree space (Hallett and Lagergren, 2000).

A number of papers have now used reconciled tree methods to infer species phylogenies (Cotton and Page, 2002; Martin and Burg, 2002; Page, 2000; Slowinski et al., 1997). One continuing concern is that gene tree parsimony methods treat conflict between the trees as a real, biological phenomenon, demanding a biological explanation. This is both a strength and a weakness, as much of this conflict may indeed be due to estimation error. A number of methods have been proposed for incorporating some confidence interval around a gene tree into the estimation process (Cotton and Page, 2002; Page, 2000; Page and Cotton, 2000). Interestingly, these suggestions mirror suggestions for incorporating similar information into MRP analyses, by using some form of "weighted MRP" (Bininda-Emonds

and Sanderson, 2001; Salamin et al., 2002).

2.5 Gene tree parsimony as a supertree method

As gene tree parsimony can be seen as a supertree method, it is of interest to see how gene tree parsimony resolves conflict between source trees when compared to different MRP methods. While many of these properties have been considered 'biases' in the literature, any method attempting to resolve some of the conflict between source trees, rather than simply representing these differences as polytomies, will show at least some of these effects. We can also defend some of these biases as biologically reasonable – for example, it is likely that larger trees are, on average, better supported (Bininda-Emonds and Bryant, 1998). Of course, if incongruence between the trees has been caused by gene duplication and gene loss, then the properties of gene tree parsimony supertrees reflect this correctly, and so should not be considered biases.

2.5.1 Correctly displays non-conflicting subtrees

Gene tree parsimony appears to correctly include non-conflicting subtrees in the supertree or species tree, a property shared by MRP methods, but not by the original formulation of mincut supertrees (Page, 2002a). Using Page's example (figure 2.2) we can see that gene tree parsimony correctly reconstructs these groupings under both duplication-only and duplication-and-loss criteria. In fact, gene tree parsimony performs somewhat better than the modified mincut method, in that it correctly places taxon a as sister-group to the clade (x1...x3) and taxon c as sister-group to the clade (y1...y4), rather than collapsing these relationships to a polytomy (Page, 2002a). Clearly, reconstructing clades that are non-conflicting is a desirable property for any supertree method.

2.5.2 Bias towards similarity with larger source trees

It has been noted that the original coding suggested for MRP supertree matrices (Baum, 1992; Ragan, 1992) produces supertrees biased towards including those relationships shown on larger source trees (Purvis, 1995b), because of redundant information in the matrix. We can use Purvis's example to show that gene tree

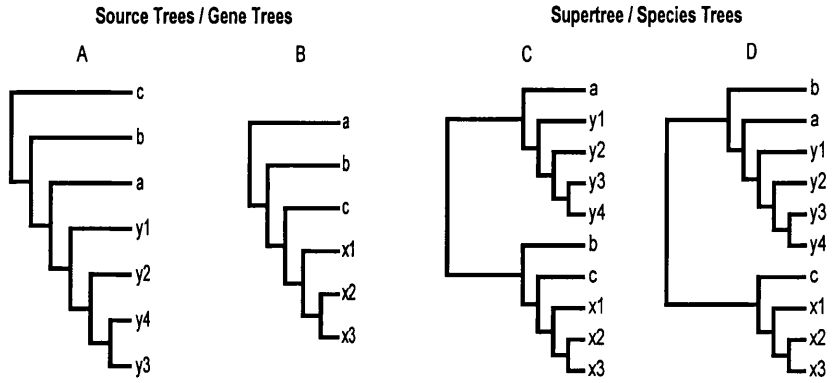

Figure 2.2: Trees C and D are the two supertrees for source trees A and B under both the duplication-only and duplication-and-loss costs (source trees taken from Page, 2002a).

parsimony also suffers from this bias when the duplication-and-loss criterion is used, but not under the duplication-only criterion. The two gene trees shown in figure 2.3 A and B support just a single species tree under the duplication-and-loss criterion, that of figure 2.3C. This tree places taxon d in the position supported by tree A, the larger of the two source trees, despite the very different position of this taxon in tree B, and so effectively ignores the conflicting signal from this smaller tree. Under the duplication-only criterion, an additional species tree (figure 2.3D) has an equal cost, and shows taxon d in the position suggested by the smaller input tree.

The reason for this bias under the duplication-and-loss criterion is clear – duplications inferred on larger gene trees will tend to infer more gene losses than those on smaller trees. Under this criterion, the species tree will thus be selected to minimize gene duplications on larger gene trees more than on smaller ones, and so will tend to reflect relationships in larger gene trees more accurately. This source of bias disappears under the duplication-only criterion.

2.5.3 Bias towards more crownward position of leaves

Several suggested variants of MRP appear to suffer from a bias towards placing species in the most crownward position displayed by the input trees. This bias was first noticed by Ronquist (1996) as being a problem with Purvis's 1995b suggested

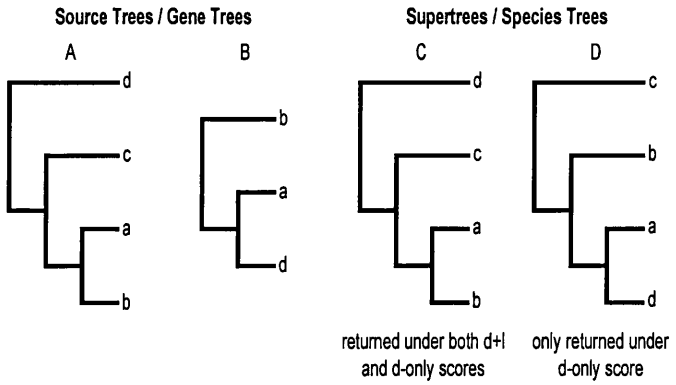

Figure 2.3: Trees C and D are the two supertrees for source trees A and B under the duplication-only cost. Tree C is the unique supertree under the duplication-and-loss cost.

modification to the original MRP encoding, as is shown by the example in figure 2.4 (from Thorley, 2000). The figure shows two source trees A and B. Under both the duplication-and-loss and duplication-only criteria, there is only a single optimal species tree (fig 2.4C). This places taxon e in the more crownward position, as suggested by source tree B, overruling the conflicting position suggested by tree A.

2.6 Biologists are interested in gene duplications

An additional desirable property of using this biologically meaningful cost function for supertree construction is that many biologists are interested in gene duplication and loss, and that reconciled tree methods can simultaneously reconstruct phylogeny and teach us something about these biological processes. Biological interest in gene duplication as a major source of evolutionary novelties dates back at least to 1933 (Haldane, 1933, see Prince and Pickett, 2002; Wagner, 1998 for more recent reviews). Of additional interest is the pattern of gene duplication through evolution, with particular attention being focused on the idea that entire genome duplication events have been important in structuring vertebrate genomes. This "2R hypothesis" was first proposed by Susumu Ohno in 1970 (Ohno, 1970), and has been the focus of intense research interest recently, as genome sequences have

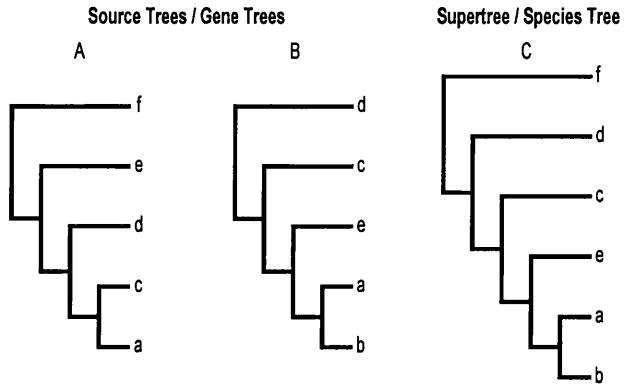

Figure 2.4: Tree C is the unique supertree for source trees A and B under both duplication-only and duplication-and-loss costs.

greatly increased the amount of data available to test this hypothesis (Gu et al., 2002; McLysaght et al., 2002, see Skrabanek and Wolfe, 1998 for some recent reviews). Reconciled tree methods may help us to test these ideas (Cotton and Page, 2002; Page and Cotton, 2002).

2.7 A probabilistic view of the supertree problem

We can usefully view the supertree problem in a probabilistic setting, a view which makes a number of the themes of this paper particularly clear. This is a fairly natural extension of the distance-based view expressed earlier – instead of seeking the closest tree to a set of source trees, we can look for the maximum likelihood supertree for this set. To do this, we need a likelihood function for the supertree, which is proportional to the probability that the source trees come from the supertree, i.e. for a supertree, T_s, from a set of n subtrees, $T_1 \ldots T_n$.

$$L(T_s|T_1, T_2..T_n) \propto p(T_1, T_2 \ldots T_n|T_s) \tag{2.1}$$

$$p(T_1, T_2..T_n|T_s) = \prod_{i=1}^{n} p(T_i|T_s) \tag{2.2}$$

There are some natural ways we can frame this likelihood function, based on how similar the source trees are to the subtrees of the proposed supertree induced

by their leaf sets. For example, if we assume every Nearest-Neighbour Interchange (NNI) needed to move from the induced subtree to the source tree is equally likely, it is relatively trivial to construct this function using a binomial distribution. To do this we need the NNI distance between source tree and induced subtree, and the diameter of the tree adjacency graph (Robinson graph) under this operation (Robinson, 1971). Unfortunately, calculating the NNI distance between any two trees is NP-complete (DasGupta et al., 1997), but DasGupta et al. present an exact algorithm to calculate the distance in reasonable time where this distance is small, and efficient approximation algorithms for the general case are available (Brown and Day, 1984). For the diameter under the NNI distance, upper and lower bounds are available, but exact values can only be computed using 'brute force' (Li et al., 1996). The probability of each NNI, q, must also be estimated from the data, but this adds only a single parameter to the model. If we represent the diameter of the adjacency graph as ΔG, the NNI distance between the source tree T_i and the subtree induced on T_s by the leaves of T_i as d_{T_i,T_s} then the probability of the source tree under the binomial model is simply:

$$p(T_i|T_s,q) = \binom{\Delta G}{d_{T_i,T_s}} q^{d_{T_i,T_s}} (1-q)^{\Delta G - d_{T_i,T_s}} \tag{2.3}$$

Constructing this likelihood function allows us to find a maximum likelihood supertree under this model, using standard heuristic methods or using methods such as Markov-Chain Monte Carlo, but it would also be easy to estimate the supertree in a Bayesian framework. To do this we need to propose a prior probability distribution on the supertree – either a 'flat' prior or one based on a model of the branching process of evolution, as has been done with earlier work on Bayesian estimation of phylogeny from sequence data. A Bayesian method would let us construct a credible interval of trees within which the true supertree lies with high probability. Alternatively, sampling from this posterior probability distribution of supertrees should allow the construction of probability distributions in the various evolutionary studies in which supertrees have been used (Huelsenbeck et al., 2000b).

More importantly, formulating the supertree problem in this way shows that any reasonable likelihood function relating a subtree to the supertree can be used to build supertrees – one based on the NNI distance seems a reasonable simple

null model (albeit one that is computationally difficult), but is an oversimplification. Other tree distances could be used, but will be similarly lacking in biological realism. If character data are available for all the subtrees, an obvious and valuable approach would be to calculate these probabilities using a model of sequence evolution, providing a natural way to incorporate uncertainty in the source tree estimates. The duplication-and-loss and duplication-only scores produced by reconciled tree methods are an attempt at a more biological reasonable distance score, and probabilistic models of gene duplication and gene loss are also being developed (Lindsey Dubb, pers. comm.) Even horizontal transfer can be incorporated, although this is more difficult to model mathematically (Charleston and Robertson, 2002). It seems likely that simplifying assumptions, like those of single base substitutions in DNA sequence phylogeny models, will be needed.

Conclusion

It is only now that realistic models for DNA sequence evolution are becoming widely used, as computational methods like MCMC become more widely understood and employed among biologists. This is some 20 years after the first tractable likelihood model for inferring phylogenies from sequences was introduced (Felsenstein, 1981). Supertree methods generally treat incompatibility between trees as noise, and treat this noise in a biologically unrealistic way. By considering gene tree parsimony alongside supertree methods, we can see that it is possible to treat such incompatibility in a more biologically realistic way. We hope that this chapter will encourage biologists to think more about how incongruence between trees can be investigated, and about the possible causes of this incongruence beyond simple estimation error. Lastly, we hope we have convinced readers that reconciled trees are a viable method for constructing supertrees for molecular data, and that we can make a first attempt to learn something about the causes of incongruence between source trees using these methods.

Chapter 3

Gene Tree Parsimony vs. Uninode Coding for Phylogenetic Reconstruction[1]

Abstract

Simmons and Freudenstein (2002) have suggested that there are important weaknesses of gene tree parsimony in reconstructing phylogeny in the face of gene duplication, weaknesses that are addressed by Simmons et al. (2000) method of uninode coding. Here, we discuss Simmons and Freudenstein's criticisms and suggest a number of reasons why gene tree parsimony is preferable to uninode coding. During this discussion we introduce a number of recent developments of gene tree parsimony methods overlooked by Simmons and Freudenstein. Finally, we present a re-analysis of data from Page (2000) that produces a more reasonable phylogeny than that found by Simmons and Freudenstein, suggesting that gene tree parsimony outperforms uninode coding, at least on these data.

[1]This chapter is currently accepted, pending minor revisions, for *Molecular Phylogenetics and Evolution*, and is co-authored with Rod Page.

3.1 Introduction

Two very different methods of using paralogous genes for phylogenetic inference have been proposed: gene tree parsimony (Slowinski and Page, 1999) and uninode coding (Simmons et al., 2000). The first step in gene tree parsimony is to identify where gene duplications and gene losses have occurred on a gene family phylogeny, or set of gene phylogenies. This can only be done with some knowledge of the phylogenetic relationship of those taxa the genes are found in, or species tree. Gene tree parsimony (named by Slowinski et al., 1997) methods then propose that, if the species tree is unknown or uncertain, we should prefer the species tree that minimises the number of gene duplications, or duplications and losses, across a set of gene trees. This species tree is the most parsimonious tree in that it minimises the number of ad-hoc assumptions of paralogy between sequences.

Uninode coding (Simmons et al., 2000) takes a rather different view – it circumvents the problem of including duplicate genes in a total-evidence analysis matrix by identifying clear orthology groups and coding them as separate columns in the matrix. This would leave a great deal of missing data, so a hypothetical ancestral sequence of all the duplicated copies – representing the sequence of the gene at the moment of duplication, reconstructed under maximum parsimony – is inserted into the matrix. Finally, a binary character representing the duplication event itself is added into the matrix. Figure 3.1 shows the uninode coding scheme. Simmons and Freudenstein (2002) present a list of further rules for the implementation of uninode coding.

Here we discuss the 10 criticisms of gene tree parsimony suggested by Simmons and Freudenstein (2002), and suggest that many of them have little force, also apply to the uninode coding method, or hail from a particular perspective on phylogenetic methodology. Of the few remaining criticisms, most are reflections of a wider debate, that between consensus and "total-evidence" methods for using multiple sources of evidence in phylogenetic reconstruction. We revisit this debate briefly, to suggest that these criticisms are not decisive in deciding between gene tree parsimony and uninode coding methods. A further subset of the criticisms are aimed at only a particular implementation of the gene tree parsimony method - that of the program GENETREE (Page, 1998), and overlook a number of recent algorithmic developments.

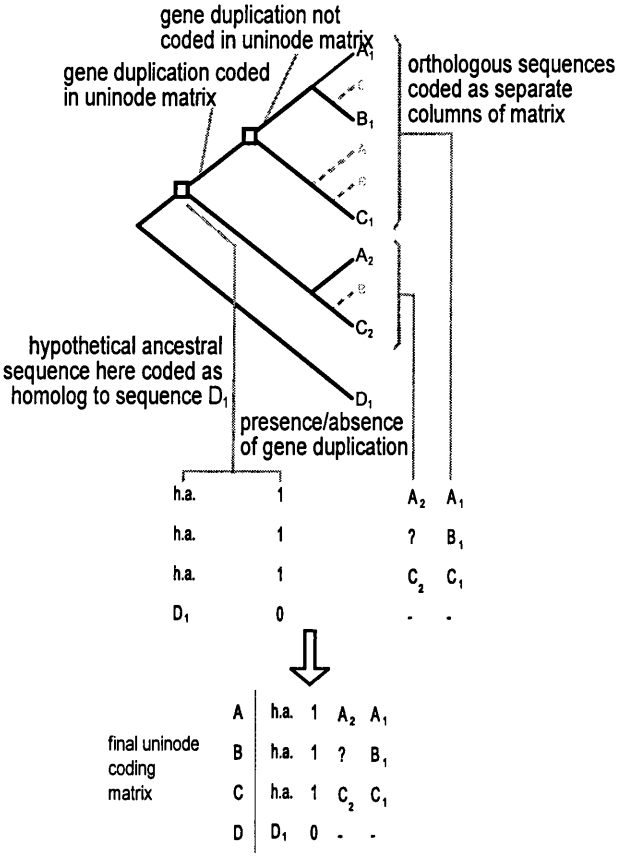

Figure 3.1: The uninode coding scheme for a gene tree for genes from species A-D (A_1 etc. are gene copies). If we assume a species tree (((A,(B,C)),D), reconciled tree methods would recognise 2 gene duplications and 4 gene losses. Only one of these duplications is recognised by uninode coding, as sequences are only present for one copy of the more recent duplication in any species. The uninode coding matrix for this gene tree is shown below - A1 etc. represent the aligned sequences of the respective genes, ? is missing data, - is inapplicable data and h.a. represents the hypothetical ancestral sequence of A_1, A_2, B_1, C_1 and C_2.

1	Problematic selection among variants
2	Non-independence of duplication events
3	Incomplete sampling of gene copies
4	Weighting of nucleotide/amino-acid characters
5	Partitioned data
6	Slow searching in GENETREE
7	Requires resolved gene trees
8	Assumes correct gene trees
9	Conflict between gene trees given equal weight
10	No branch support values

Table 3.1: Simmons and Freudenstein's criticisms of gene tree parsimony

Simmons and Fruedenstein's 10 criticisms of gene tree parsimony are listed in table 3.1. They appear in this table in the order they appear in the original manuscript – the titles given here are not from the original, but (hopefully faithfully) paraphrase the main point made by Simmons and Freudenstein. These criticisms are valuable in drawing attention to certain features of the gene tree parsimony method, and in highlighting the value of certain new developments in gene tree parsimony techniques, but we disagree with Simmons and Freudenstein's conclusion that these criticisms imply that "uninode coding be used instead of gene tree parsimony for phylogenetic inference from paralogous genes".

3.2 Different algorithms and new techniques

GENETREE is a single implementation of reconciled tree methods to infer phylogeny from gene families, but Simmons and Freudenstein confuse the limitations of the GENETREE program with the conceptual limitations of the reconciled tree methods themselves. This is particularly clear in the case of criticism #6 – about the slowness of GENETREE's heuristic searches – GENETREE currently implements the algorithm of Eulenstein (1997), a development of the original mapping algorithm (Page, 1994a), and then uses heuristic searches through tree space to find the optimal species tree. More efficient search strategies are available – Hallett and Lagergren (2000) present a fixed-parameter tractable algorithm for finding

the optimal species tree for a set of gene trees under the duplication-and-loss criterion without the need for this heuristic search. This is likely to be implemented in a future version of GENETREE, and certainly demonstrates that slowness is not a property of the gene tree parsimony method itself. Simmons and Freudenstein's use of Page and Charleston (1997a) search strategy to claim that "GENETREE is too slow to thoroughly search the tree space" is particularly misleading given that Hallett and Lagergren (2000) demonstrate that Page and Charleston do indeed identify species trees with the globally best cost for Guigó et al.'s (1996) data.

The same algorithms also answer criticism #7 – both Eulenstein, and Hallett and Lagergren suggest that their algorithms can be easily extended to cases where gene trees contain polytomies. One easy way to include polytomies, which we have implemented in a version of GENETREE, is to allow a set of gene trees to be input, and to minimise duplications or duplications and losses across this set of trees. If a polytomy is considered to be a "soft" polytomy (Maddison, 1989), it represents uncertainty between a number of different possible bifurcating trees, differing in the order of branching above this node. A set of trees could thus include all the possible dichotomous resolutions of any polytomies in the input gene tree, but equally could be a set of most-parsimonious trees from a parsimony analysis, or some similar representation of the uncertainty in the gene tree estimate.

Simmons and Freudenstein suggest that no branch support values can be provided for reconciled trees (criticism #10) – this is untrue. One way to incorporate branch support values is to use a bootstrap profile of gene trees as an input to the gene tree parsimony step, generating a set of species trees. The proportion of these trees containing a clade of interest would then be a direct analogue of standard bootstrap proportions, as suggested by Page and Cotton (2002), and recently used in Cotton and Page (2002). In fact, this method also helps answer criticism #8 – using a set of bootstrap trees effectively provides a confidence interval around the best estimate of each gene tree, relaxing the requirement for correct, fully resolved gene family trees. This should also improve inferences about the patterns of gene duplications and losses. In fact, we need not use a set of bootstrap trees – using a Bayesian credible set of trees might give a more statistically rigorous confidence interval (Huelsenbeck et al., 2000b).

3.3 Selection among variants of gene tree parsimony

The choice between different analysis methods is not unusual in scientific methods, and is hardly a substantive criticism – in parsimony methods generally (including analysis of uninode coded data) we must choose between different weighting schemes (e.g. weighting transitions higher than transversions) and we frequently have to make choices between methods of phylogenetic reconstruction. Beyond this, a number of methods of phylogenetic analysis are available when faced with a sequence alignment. Flexibility in analytical method only seems a problem under the view that there is only a single "true" method of phylogenetic inference, a philosophy not shared by all systematic biologists. We see the availability of a range of analytical tools as a positive thing, not a negative one.

In any case, the fact that Simmons et al. (2000) and Simmons and Freudenstein (2002) only suggest a single uninode coding method does not imply that other variants cannot be proposed. For example, Simmons et al. (2000) make no defence as to why the binary gene duplication characters need to be included in the matrix at all – uninode coding would still be logically consistent without these characters, or with these characters weighted twice, or three times or any number at all. This problem was recognised more than 20 years ago (Fitch, 1979) – there is no logical way to decide how to weight a duplication character relative to a nucleotide or amino-acid substitution. Uninode coding methods suffer from the same 'problem' of multiple variants as gene tree parsimony methods.

A final point is that it seems that duplication-and-loss and duplication-only scores will always give compatible results, but that the duplication-and-loss result will be better resolved. Using the duplication-only criterion is, in this case, merely more conservative, avoiding the risk of grouping some taxa together by sampling failure. This is a corollary of conjecture 3 of Page and Charleston (1997b, p. 63), which is still formally unproven. Even if this conjecture is shown to be formally false, there is certainly a close relationship between the different cost functions used in gene tree parsimony – both duplication-only and duplication-and-loss scores will be highly correlated with the deep coalescence cost (as Zhang, 2000, has shown for a slightly different cost to that implemented in GENETREE).

Duplication-and-loss results can be misleading in certain circumstances. If the sampling of genes is incomplete, the absence of a gene copy from the sequence

database could be for two different reasons – because the gene copy does not exist in the species' genome or because it has not been sequenced. Duplication-and-loss costs risk conflating these two costs, and so supporting relationships on the basis of the uneven sampling of molecular biologists. In some studies, such as that of Martin and Burg (2002), where sampling is known to be fairly complete, duplication-and-loss costs are appropriate. However, studies using only a small selection of sequences taken from the public sequence databases, and including taxa that are not fully sequenced (e.g. Cotton and Page, 2002), such as the data used here, are likely to produce biased results under this criterion.

3.4 Consensus methods vs combined analysis

The debate over whether to combine data from multiple different sources of evidence in a single data matrix for phylogenetic analysis has been on-going for over a decade (for reviews see de Queiroz et al., 1995; Huelsenbeck and Bull, 1996). Three different opinions have been reflected in the literature – taxonomic congruence, which supports separate analysis and the use of consensus methods to investigate similarities between them (Miyamato and Fitch, 1995; Swofford, 1991), "total evidence" or combined analysis, which supports combining separate datasets before analysis (Barrett et al., 1991; Kluge, 1989) and an intermediate position, which advises combining data when statistical tests suggest they are compatible (Bull et al., 1993; Huelsenbeck and Bull, 1996). There has been a long debate between proponents of these methods for dealing with multiple data sources in systematics.

We believe that, in the context of this debate, a number of Simmons and Freudenstein's criticisms of gene tree parsimony merely reflect differences between these positions, These criticisms have thus been addressed in previous discussions, and are, in any case, not decisive criticisms of the gene tree parsimony method. Simmons and Freudenstein suggest that both reconciled trees and uninode coding are "total-evidence" or "simultaneous-analysis" approaches, in the sense of Kluge (1989). However, Kluge uses "total-evidence" to apply to methods that seek to find the hypothesis that maximises total "character congruence" rather than "taxonomic congruence" – by including all possible evidence in analysis of a single data matrix. Gene tree parsimony is not a total-evidence method in this sense – as Page

(2000, p.99), explicitly states "It should be emphasized that the topology of this species tree depends entirely on the topology of the 9 gene trees (and the constraint tree); no reference is made to the underlying sequence data".

In fact, gene tree parsimony methods have something in common with both consensus methods and total evidence approaches. Gene tree parsimony is a total-evidence method in the sense that it seeks the best explanation for all the available data, but the data it uses are the phylogenies for the gene families rather than the sequence alignments themselves – effectively applying total evidence under the parsimony criterion to higher-level characters, namely gene trees. On the other hand, if we use the terminology of de Queiroz et al. (1995), gene tree parsimony is clearly a consensus method, in that 'characters in two (or more) data sets are not allowed to interact directly with one another in a single analysis, but instead interact only through the trees derived from them'. Gene tree parsimony is not a traditional consensus method, however, in that rather than seeking to summarise the a set of source trees, it seeks to find a tree best representing the evolution of a set of gene trees in a biologically meaningful way.

Traditional consensus methods are likely to be a poor choice for studying historically associated lineages such as genes and their species, as discussed by Page (1996), and acknowledged by authors on both sides of the debate (e.g. Cognato and Vogler, 2001). Consensus methods seek to represent incongruence between source trees, whereas reconciled tree methods attempt to resolve this incongruence by explaining it in terms of evolutionary events such as gene duplication and gene loss - effectively taking this incongruence 'at face value' as needing a biological explanation. The uninode coding method simply makes the minimum variation to simple combined analysis needed to incorporate multiple gene copies - any incongruence is treated as statistical error, to be submerged by the weight of combined data from multiple loci. By relaxing the requirement of gene tree topologies to be exactly correct (e.g. by using a bootstrap profile or Bayesian credible set of trees, as discussed above), we effectively allow gene tree parsimony methods to find evolutionary explanations only for significant incongruence. In fact, the difference between combined analysis and methods relying only on the reconstructed phylogeny (such as consensus methods and gene tree parsimony) reflects a statistical trade-off between reducing bias (by combining all data) and correctly estimating variance in the estimate of phylogeny (by partitioning data) – a trade-off widely

accepted in the statistical literature (Holmes, 2003). In the sense that one uses the sequence data directly and the other considers trees from the separate data partitions, gene tree parsimony and uninode coding represent alternative sides of the debate over combined analysis vs. consensus methods. Simmons and Freudenstein's criticisms #4 and #5 reflect this debate – a debate that is still active (Levausser and Lapointe, 2001) and can hardly be considered a decisive criticism of gene tree parsimony.

In fact, for practical purposes, the debate over consensus methods vs. total evidence is probably not of crucial importance. Simmons and Freudenstein, in common with other advocates of total evidence methods, suggest that total evidence methods may be more successful in that they allow "hidden support" for certain nodes to emerge from the combined matrix (Gatesy et al., 1999; Nixon and Carpenter, 1996). Hidden support refers to support across data partitions for relationships that are not evident in the most-parsimonious tree for the partitions analysed separately. While a number of studies have identified hidden support, they do not demonstrate that the hidden support is truly hidden in the sense of not being evident in a number of the trees from a bootstrap profile, or being excluded from the credible interval of trees in a Bayesian framework. Relaxing the dependence of gene tree parsimony on a single estimate of the gene trees would be expected to identify most significant hidden support.

3.5 Non-independence of gene duplications

The potential non-independence of gene duplications on trees has been recognized by a number of authors – some of the earliest theoretical work presented a method for identifying larger-scale genome duplications on a tree (Guigó et al., 1996). Most authors have followed Guigó et al. in considering independence of gene duplications as a valid simplifying hypothesis which can later be tested by comparing the distributions of duplications under this assumption and under the assumptions that the individual duplications are clustered into the minimum number of larger-scale episodes (Page and Cotton, 2002). This parallels a common assumption of phylogenetic methods, where nucleotide substitutions are considered independent because modeling dependencies between substitutions at different sites would be intractable except in simple cases where this dependency is clear, such as in the

stems of RNA molecules (Jow et al., 2002). In particular, uninode coding also makes the same assumption – the "gene duplication characters" are duplications coded as independent characters. A pragmatic reason that we do not attempt to find the species tree minimising the number of gene duplication episodes is that this is demonstrably NP-hard (Fellows et al., 1998).

3.6 Hidden paralogy

The main criticism we have of the uninode coding method is that it ignores the possibility of hidden paralogy – paralogy that is not obvious due to the presence of both gene copies existing in extant genomes (Figure 3.2).

How frequent hidden paralogy will be depends upon rates of gene duplication and loss – as gene families evolve under a birth-and-death process (Nei et al., 2000). Hidden paralogy may be more common than would be suggested by single average rates of duplication and loss, as duplicate genes are complementary, so one copy will rapidly go extinct if a mutation renders one of the copies non-functional – there is no selective pressure to retain both copies of the gene (Lynch and Conery, 2000). If a speciation event occurs during this process, then different paralogous copies could easily go extinct in each lineage – in the simple case in which the two lineages have an equal chance of survival this will occur 50% of the time. Where gene duplications are frequent, and gene silencing and subsequent loss relatively slow, hidden paralogy will be very common. Apparent hidden paralogy could also pose a problem for the uninode coding method – even where multiple gene copies from a duplication exist in the genomes of some species, there will be situations in which no species shows both gene copies because of the incomplete sampling of genomes.

Uninode coding also ignores the possibility that the gene duplications present on the most-parsimonious gene tree (in stage 1) are incorrect – these duplications will be incorporated into the uninode coding matrix. This matrix pseudo-replicates some of the data by incorporating hypothetical ancestral sequences many times into the matrix, which are entirely dependent on the sequences they are calculated from. This pseudo-replication has two effects – it makes it very unlikely that the phylogenetic groups supported by gene duplications on the original parsimony trees will not be present in the final parsimony trees, particularly for duplications ancestral

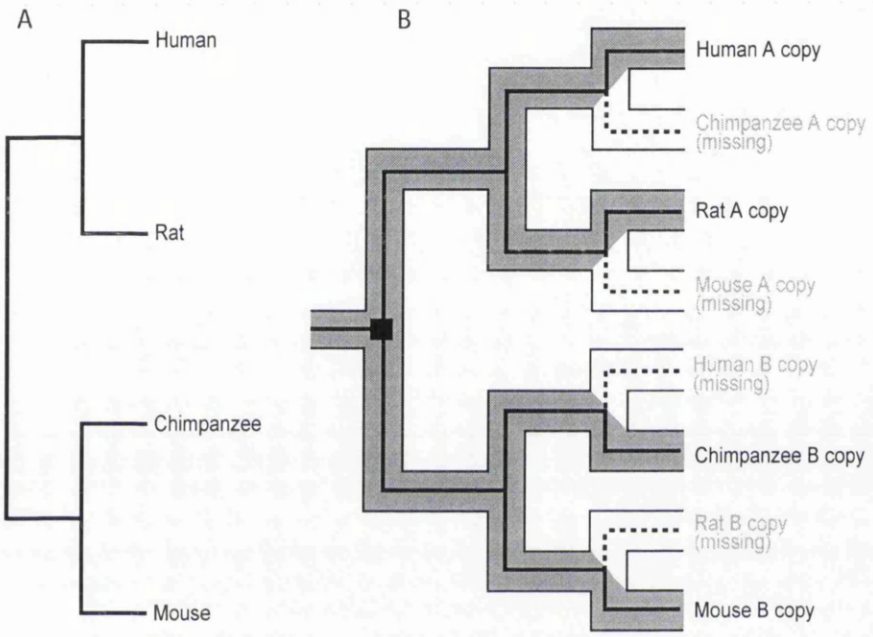

Figure 3.2: Hidden paralogy. The gene tree (a) shows no duplicated genes that would be coded as such in the uninode coding method, but any reasonable assumption about the relationships between these four species would suggest that the true pattern of evolution in this gene family is as seen in the reconciled tree (b). (b) shows a duplication at the base of the gene tree, followed by four losses (or failure to sample four of the genes), suggesting that the rat and human genes are orthologues, and are paralogous to the mouse and chimpanzee orthologues.

to large numbers of species, and it makes bootstrap values for these nodes very difficult to interpret.

3.7 An empirical example

Simmons and Freudenstein present a re-analysis of data from Page (2000) using uninode coding, and find a substantially different result. We use this data again to demonstrate some of the more recently developed methods discussed above. Page originally used the neighbour-joining method to generate gene family trees for the 9 genes used, while Simmons and Freudenstein use parsimony trees to infer the locations of gene duplications in stage 1 of the uninode coding process. To investigate how much the differences between the results of these two studies was due to the use of parsimony rather than neighbour-joining, and to demonstrate how multiple most-parsimonious trees can be used in gene tree parsimony, we also use parsimony gene trees here.

3.7.1 Methods

The gene trees for this analysis were generated from the ClustalX alignments used by Page (2000). These alignments are freely available from `http://taxonomy.zoology.gla.ac.uk/rod/data/vertebrates/`. These alignments were converted to the NEXUS format and then analysed using PAUP 4b10 (Swofford, 1998) under the parsimony criterion, with 50 random addition-sequence replicates and TBR branch-swapping to completion, keeping multiple trees. All most-parsimonious trees found were incorporated into a GENETREE format NEXUS file and analysed using a specially-written version of the GENETREE program, which treats multiple gene trees as equally-parsimonious gene trees, searching for the species tree that minimises the cost across the set of trees, by, for each iteration of branch-swapping during the heuristic search, reconciling the species tree with each gene tree in turn, and recording as the correct cost the minimum cost across all the trees for that gene family. As discussed by Page, constrained searches are needed for this data to address the limited taxonomic coverage of most gene families, and the same constraints as used by Page (and Simmons and Freudenstein) were used in all analyses shown here.

Because of the complexity of searching across the profiles of most-parsimonious trees for each gene family, for every postulated species tree during the heuristic search, the searches for this data were very slow. The inclusion of multiple MPTs for each gene family also greatly increased the numbers of equal-cost trees found, so a two-step search strategy was employed. For the first step, a large number of starting tree replicates were used, but branch-swapping was performed on only a single tree during the search, thus preventing the searches becoming trapped on plateaus of equally-parsimonious trees. The shortest trees from these searches were then swapped on to exhaustively sample from the island of trees identified during the first stage. This two-stage procedure gives us a reasonable chance of locating the shortest trees, and ensures that we sample adequately from the island (or islands) of trees found.

For both duplication-only and duplication-and-loss criteria, 100 searches starting from random addition-sequence replicate trees were performed. Under the duplication-only criterion, 7 of these searches found the lowest detected cost of 92 duplications, finding 7 different species trees. Several additional searches, holding multiple trees, were also run under this criterion, which were not run to completion but found over 15,000 trees of this cost without finding any lower-cost solutions. Under the duplication-and-loss criterion, 21 searches found trees with the lowest detected cost, of 383 duplications and losses. All seven of the duplication-only optimal trees found in these searches, and a randomly chosen sample of 10 duplication-and-loss optimal trees were used as starting points for searches swapping on multiple trees. Each of these searches was run until at least 1000 trees had been found, and in many cases were left for much longer, with none of the searches finding shorter trees than were identified in the first stage searches. The Adams and strict consensus for each of the 7 duplication-and-loss results and each of the 10 duplication-only sets of trees were identical, or differed only in the degree of resolution of a single node within the reptiles (for the duplication-and-loss data), confirming that each search had successfully sampled from across the island of minimal trees. The strict consensus trees are shown in figure 3.3.

As pointed out by Simmons and Freudenstein, the standard gene tree parsimony analyses described above use only a single fully-resolved phylogeny for each gene family, and so can take no account of weaknesses in the gene family trees. For example, many gene families may be unable to resolve particular rela-

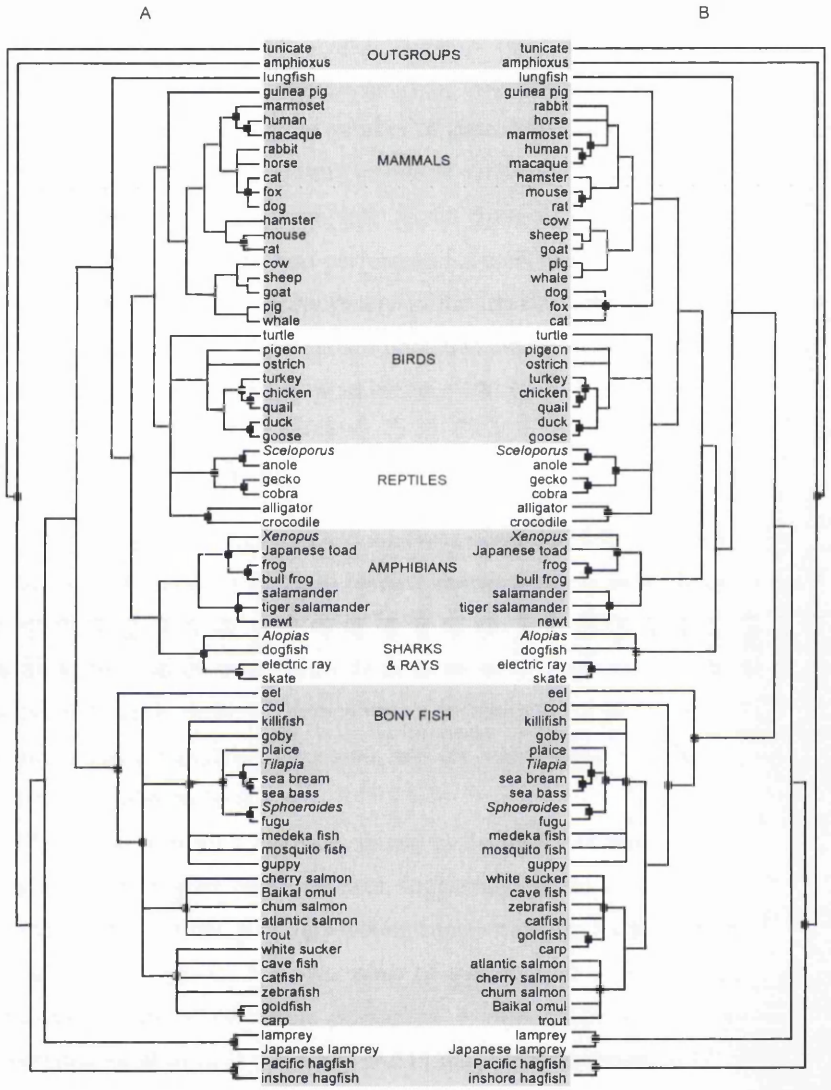

Figure 3.3: Results of a gene tree parsimony search finding the species tree minimising the number of (a) duplications and losses and (b) gene duplications across the most parsimonious trees from the gene families of Page (2000). The trees shown are the Adams consensus of all minimal cost trees found during the searches. Nodes marked with a square were constrained during the search.

tionships or show only limited support for a particular resolution. To incorporate this information, we have adopted a gene tree bootstrapping protocol (Cotton and Page, 2002; Page and Cotton, 2000). A set of 100 bootstrap trees for each gene family in the dataset, using the fast heuristic bootstrapping method of Paup 4b10. The species tree minimising the number of gene duplications were then found for successive trees from the bootstrap profile of each gene family, producing 100 sets of species trees. A single, complete search from a single random starting tree, keeping multiple solutions, was performed for each replicate, with multiple equal solutions down-weighted appropriately in the final calculation of support values. Support values analogous to standard bootstrap values could then be calculated as the number of times nodes appeared in these 100 species tree.

3.7.2 Results and discussion

The full phylogenetic results of the analyses described here are shown in figures 3.3 and 3.4. A summary of these results, showing relationships between the major vertebrate groups and comparing these results with the results of Page (2000) and Simmons and Freudenstein (2002) is shown in figure 3.5. We restrict this discussion to relationships between major vertebrate groups, all of which are unconstrained in the analyses discussed, and for which there is a clear idea of what the expected relationships are.

We can see that all 4 analyses shown in figure 3.5 support different relationships among the higher vertebrate taxa, suggesting (as our bootstrap values reflect) that these genes do not give very strong support for any picture of vertebrate relationships. As figure 3.5 shows, none of the analyses correctly reproduces the traditional picture of vertebrate phylogeny, a view supported by a great weight of morphological work (e.g. Bishop and Friday, 1988; Løvtrup, 1977) and by gene tree parsimony analysis of a much larger data set (Cotton and Page, 2002). Furthermore, none of the results are wholly congruent with phylogenies based on whole mitochondrial genome data (Rasmussen and Arnason, 1999; Zardoya and Meyer, 2001b).

All of the four results share some weaknesses – all misplace the sharks and rays, placing them in too derived a position in the vertebrate tree. The trees also all fail to resolve relationships within the reptiles, or present a somewhat unusual

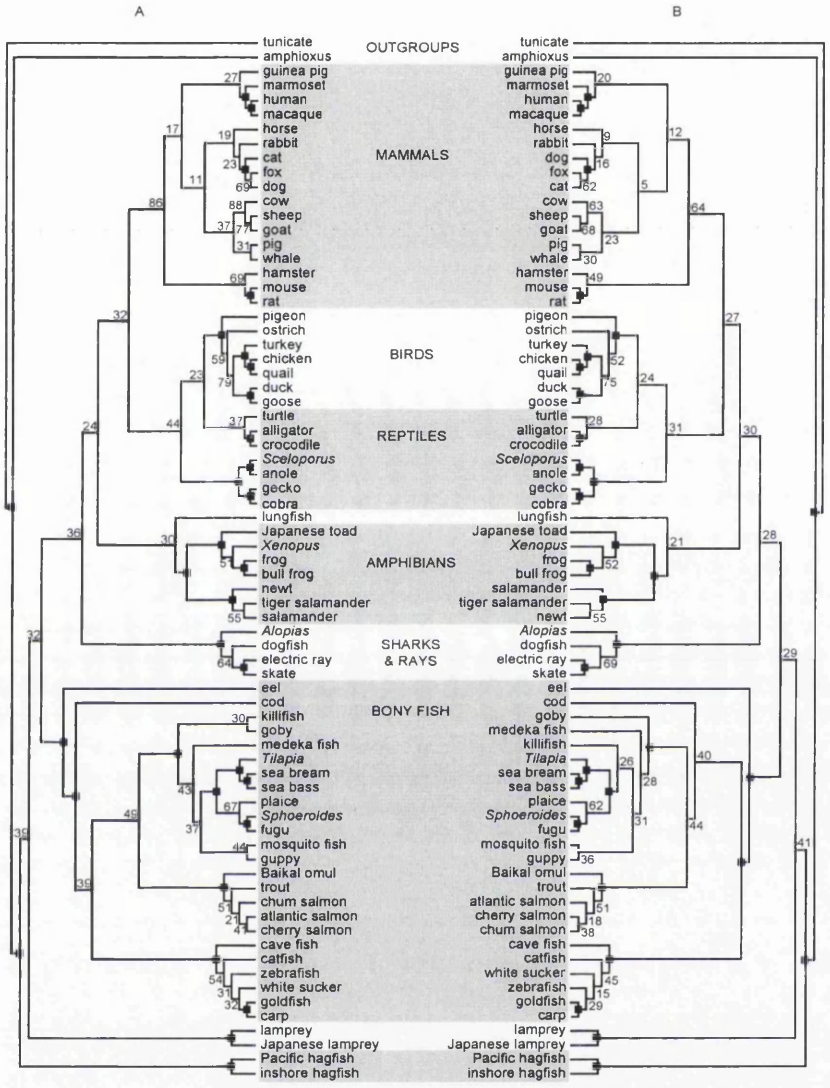

Figure 3.4: Results of a gene tree parsimony bootstrap analysis. Show are majority-rule consensus trees (including compatible groups present in less than 50% of the trees) of 100 species trees obtained by minimising the number of (a) gene duplications and losses and (b) duplications only, for each of 100 bootstrap trees for the gene families of Page (2000). Figures at nodes represent the number of times this node appeared in the 100 resulting species trees. Nodes marked with a square were constrained during the search.

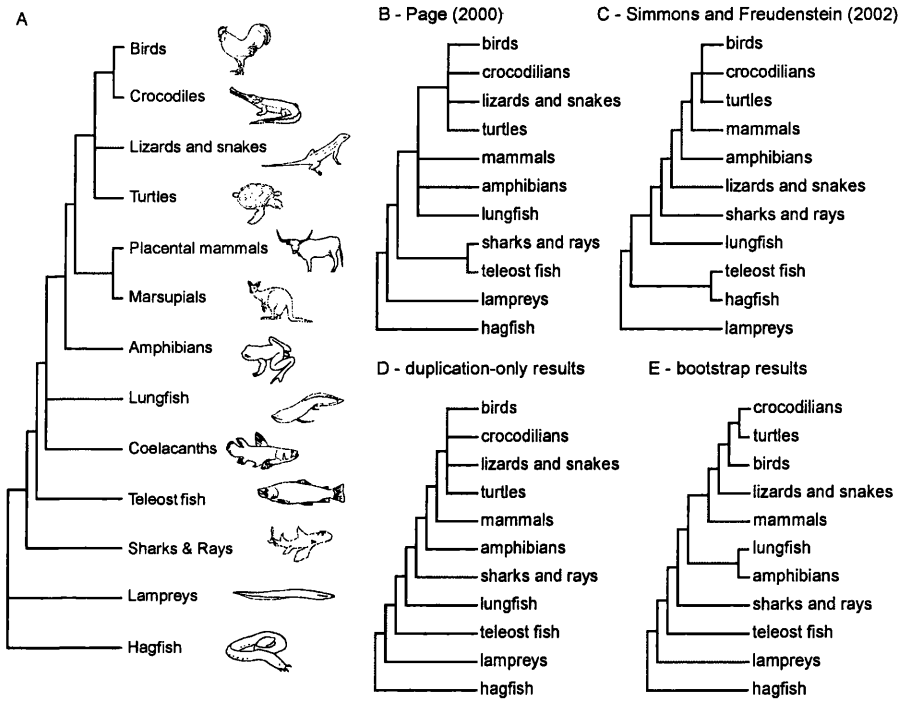

Figure 3.5: Summary of the results of (b) Page (2000), (c) Simmons and Freudenstein (2002) and this study: (d) shows the strict consensus of duplication-only optimal trees, (e) the majority-rule consensus of the bootstrap replicates. Part (a) shows a traditional picture of vertebrate phylogeny based on morphological and paleontological evidence (Bishop and Friday, 1988). Part a is from chapter 4, figure 4.2.

phylogeny within this group. While most workers would agree that the turtles are the most basal of the extant reptiles, with lizards and snakes (the lepidosaurs) forming a sister-group to an archosaur clade of crocodiles and birds, relationships within the group have become somewhat uncertain in the light of molecular evidence, which tends to place turtles as relatives of the archosaurs (Hedges and Poling, 1999; Rieppel, 2000), as suggested by Simmons and Freudenstein's result – the placement of turtles within the archosauria as shown in figure 3.4d isn't supported by other evidence.

Simmons and Freudenstein's result shows some problems not present in any of the gene tree parsimony results. Their results fail to correctly unite the lizards and snakes with the other archosaurs, and fail to place the hagfish as a basal vertebrate lineage. There is no doubt that lizards and snakes form part of a monophyletic radiation of diapsid reptiles, although there has been some debate about the exact relationships between the different extant lineages within this radiation, as discussed above. Similarly, there has been debate about the exact relationships between hagfish, lampreys and gnathostomes (Delarbre et al., 2002; Janvier, 1996), but the only hypotheses supported by recent work are that lampreys and hagfish form a monophyletic cyclostome group, or that hagfish are the most basal vertebrates, with lampreys a sister-group to the gnathostomes. In conclusion, the results of this study are a better estimate of correct vertebrate phylogeny than those of Simmons and Freudenstein. It is striking that Simmons and Freudenstein find high bootstrap support for some clearly erroneous relationships, such as 87% support for a monophyletic clade of amphibians and tetrapods, but excluding the lizards and snakes, and 90% support uniting the hagfish and teleost fish.

3.8 Conclusion

Differences between uninode coding and gene tree parsimony are largely ones of perspective – uninode coding is a combined analysis method, modified to allow the use of multiple genes for each taxon. The relative effectiveness of gene tree parsimony methods and uninode coding will partly depend on the extent of hidden paralogy – the extent to which the signal from different clades coded in the uninode matrix conflict – and to what extent noise makes the individual gene trees inaccurate. This is an empirical issue, and not one decided by Simmons and Freudenstein's

criticisms of gene tree parsimony methods. For the data analysed here, gene tree parsimony gives a more reasonable vertebrate phylogeny, suggesting that for these data it is important to correctly identify hidden paralogy. Finally, gene tree parsimony methods can identify gene duplications despite widespread gene loss, and so are valuable tools in the study of the pattern and process of gene duplication itself (Page and Cotton, 2002).

Part II

INFERRING SPECIES PHYLOGENIES

Chapter 4

Going Nuclear: Gene Family Evolution and Vertebrate Phylogeny Reconciled[1]

Abstract

Gene duplications have been common throughout vertebrate evolution, introducing paralogy and so complicating phylogenetic inference from nuclear genes. Reconciled trees are one method capable of dealing with paralogy, using the relationship between a gene phylogeny and the phylogeny of the organisms containing those genes to identify gene duplication events. This allows us to infer phylogenies from gene families containing both orthologous and paralogous copies. Vertebrate phylogeny is well understood from morphological and palaeontological data, but studies using mitochondrial sequence data have failed to reproduce this classical view. Reconciled tree analysis of a database of 118 vertebrate gene families supports a largely classical vertebrate phylogeny.

[1]This chapter has been published as *Proc. Roy. Soc. Lond.* B (2002) **269**, 1555-1561, co-authored with Rod Page.

4.1 Introduction

The central assumption of molecular systematics is that a phylogeny estimated from a set of gene sequences tells us something about the phylogeny of the organisms the genes have been isolated from. In fact, systematists generally assume that the gene phylogeny (or gene tree) is isomorphic with the organism phylogeny (or species tree), so that a correct estimate of the species tree can be obtained by simply re-labeling the leaves of the tree with the appropriate species names. In this case, differences between phylogenies from different loci – or differences between a gene tree and the commonly accepted species tree – are due to either the method by which gene phylogenies have been constructed or sampling error in the estimate of gene phylogeny. In the latter case, more sequence data should produce the correct species tree.

However, gene trees are not species trees, and a number of evolutionary processes can introduce differences between a correctly estimated gene phylogeny and the correct species phylogeny (Doyle, 1992; Maddison, 1997). These processes are horizontal transfer, duplication and loss, and deep coalescence (Doyle, 1992; Slowinski and Page, 1999). Because these events introduce differences between the gene tree and species tree, we can use incongruence between these two trees to infer the past occurrence of the events (Page and Charleston, 1997a). This is the motivation behind reconciled trees. Reconciled trees are a general method for analysing historical relationships where one entity tracks another, with the fidelity of this 'tracking' dependent on how often events such as duplication, horizontal transfer and lineage sorting occur (Page and Charleston, 1998). These events will introduce differences between the trees that describe the hierarchy of the two entities, as in figure 4.1, where a duplication in the gene tree and three gene losses explains the difference between the gene and species trees. Where all these different events are allowed, it can be very difficult to correctly reconstruct potential evolutionary scenarios (Charleston, 1998), but if we restrict the analysis to consider only duplications and losses then finding the most parsimonious reconstruction of events is relatively trivial and can be computed in linear time (Zhang, 1997).

As we consider all the gene trees to be independent estimates of the underlying species phylogeny, the most parsimonious species tree is that which implies the minimum number of gene duplication (or duplication and loss) events over the

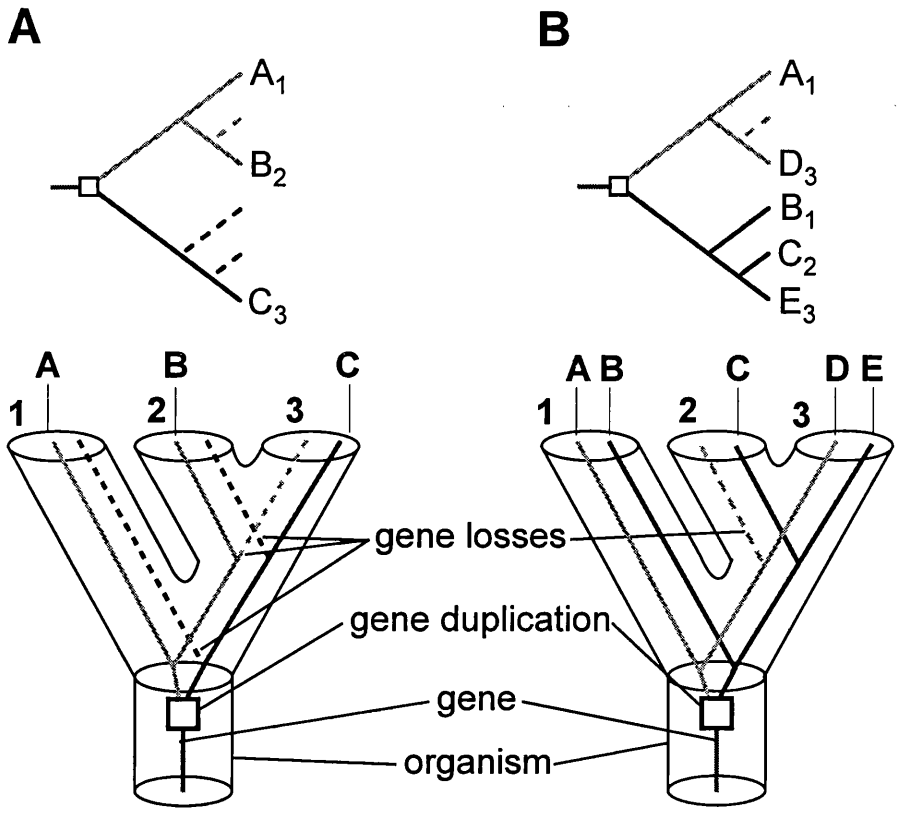

Figure 4.1: Gene duplication and loss can introduce incongruence between gene phylogenies and species phylogenies. (A) With three genes (A-C) sampled from three different species (1-3), the difference in topology between the gene and species trees can be explained by one gene duplication and three losses. The same approach also applies where multiple genes are known from each species – (B) shows a gene tree requiring one duplication and one loss. Reconciled trees can be seen as representing the simplest embedding of a gene phylogeny inside a given species phylogeny.

set of gene families, and we can use simple and standard heuristic methods to find an optimal species tree topology (Page and Charleston, 1997b). Using the number of gene duplications as an optimality criterion to choose between competing phylogenetic hypotheses in this way has become known as "gene tree parsimony" (Slowinski and Page, 1999). Gene tree parsimony thus treats gene trees as characters of species, in contrast to conventional phylogenetic methods using molecular sequences as characters of organisms, conflating organismal and gene phylogenies.

The evolution of the vertebrates represents an ideal case for testing the utility of reconciled tree methods (Page, 2000). Vertebrate classification has been of interest since antiquity, and a great deal of morphological data from both extant and fossil taxa has produced a well-supported outline of vertebrate phylogeny (figure 4.2). Vertebrate workers have a keen sense of where the vertebrate tree is fairly robust and where relationships are much less clear – and all of these areas have attracted a great deal of debate. There is thus an opportunity for new techniques to both prove themselves, by successfully reconstructing those parts of the tree that are more-or-less beyond doubt, and to make a real contribution to resolving areas of contention.

Given the great deal of support for much of the current pattern of vertebrate relationships, it is surprising how poorly molecular methods have fared in reconstructing the broad outline of vertebrate evolution. This is particularly worrying in the case of mitochondrial genome sequences, which are relatively large markers that have been thought of as ideal for phylogenetic work and are certainly very commonly used. Figure 4.3 shows two recently published phylogenies based on mitochondrial genome sequences, showing the unusual relationships between major groups of basal vertebrates typical of analyses based on these data.

Some of the errors in mitochondrial phylogenies have been due to incorrect rooting of the gnathostome part of the tree (Takezaki and Gojobori, 1999), but other unusual placements occur. These errors occur despite mitochondrial loci having increasingly good taxon sampling. Explaining these erroneous results has become a major concern in the literature, particularly because several studies show high bootstrap support for unusual relationships (Naylor and Brown, 1997; Zardoya and Meyer, 1996), which some have taken at face value as providing strong evidence for these relationships. Other studies have sought to explain the unorthodox relationships as artefacts due to a low signal-to-noise ratio (Zardoya and Meyer, 2001b)

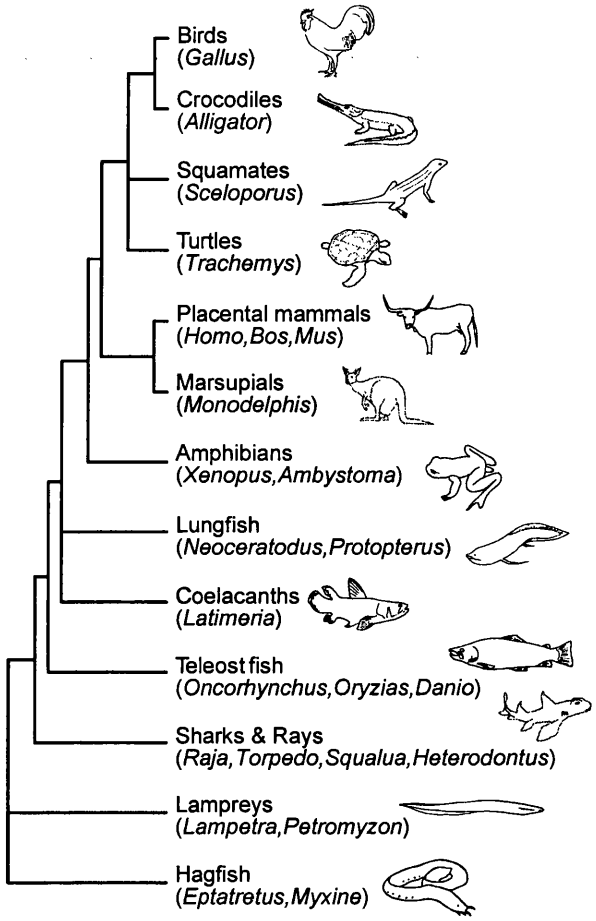

Figure 4.2: A traditional view of vertebrate phylogeny, based on morphological and palaeontological data. Based on Bishop and Friday (1988). The names of all genera included in the gene tree analysis (see figure 4.4) are listed.

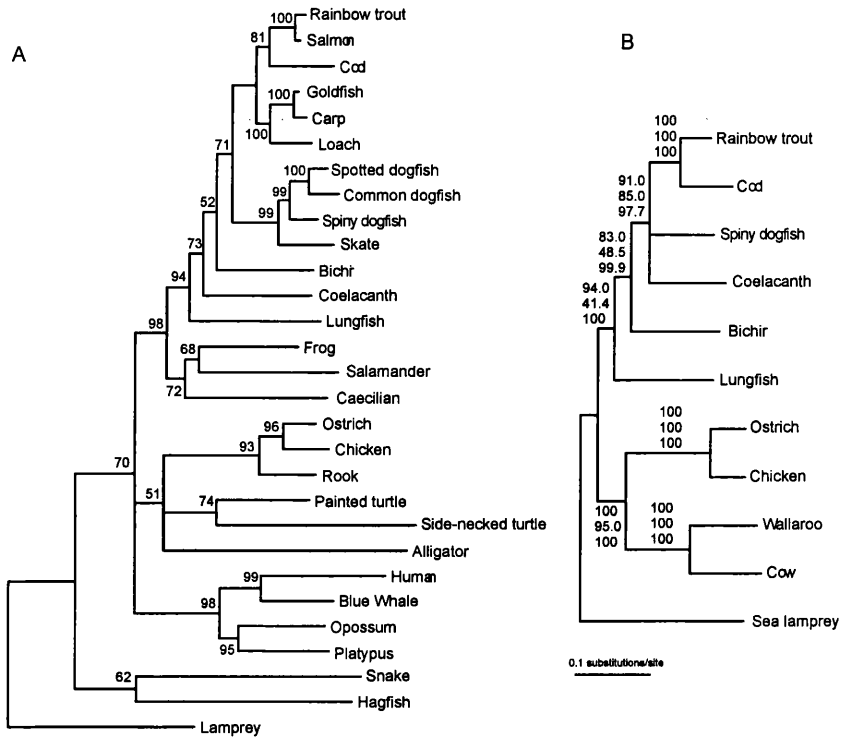

Figure 4.3: Vertebrate phylogenies based on whole mitochondrial genome data. (A) is a maximum-likelihood tree from Zardoya and Meyer (2001b), with numbers on nodes being bootstrap percentages based on 100 pseudo-replicates. Zardoya and Meyer do not accept this tree of vertebrate relationships, but are unable to reconstruct a more reasonable phylogeny. (B) is the maximum-likelihood tree from Rasmussen and Arnason (1999). Figures on branches are neighbor-joining (top) and maximum-parsimony (middle) bootstrap values based on 100 replicates, and maximum-likelihood (bottom) support values from 1000 puzzle replicates. Both trees were constructed using PUZZLE (Strimmer and von Haeseler, 1996) and the mtREV-24 model.

and wide differences in substitution rates between lineages (Takezaki and Gojobori, 1999), between classes of amino acids (Naylor and Brown, 1997) and between sites (Takezaki and Gojobori, 1999). Most authors agree that phylogenetic results from recent analyses of whole mitochondrial genomes 'need to be confirmed with data from nuclear genes' (Curole and Kocher, 1999; Takezaki and Gojobori, 1999; Zardoya and Meyer, 2001b).

We have used gene tree parsimony to reconstruct vertebrate phylogeny based on a database of 118 vertebrate gene families. These analyses demonstrate the utility of reconciled trees in inferring phylogenies from gene family data, supporting most of the conventional vertebrate phylogeny and adding to the evidence for some more controversial relationships, such as a monophyletic cyclostome clade of lampreys and hagfish.

4.2 Materials and methods

4.2.1 Gene family phylogenies

We chose those representatives of the major vertebrate groups present in the largest number of gene families in the HOVERGEN (Duret et al., 1994) database. We assumed the monophyly of genera, grouping genes from all species in a genus together. Where no genus in a particular group was well represented, an additional genus was used, so that data from both could help to accurately determine the relationship of the larger group. Genera included are listed on figure 2. Gene families sampling at least five vertebrate classes were selected from HOVERGEN, with additional families chosen if they provided evidence about the relationships of those genera that were poorly sampled in the initial selection. Outgroups for each gene family were found using sequence similarity searches against a number of sequence databases to identify related genes – either invertebrate orthologues or vertebrate paralogues. Due to the size of the dataset, amino acid sequences were aligned in ClustalW (Thompson et al., 1994) using default parameters and neighbor-joining phylogenies constructed in ClustalW, including gapped positions and using uncorrected distances. Alignments were also examined by eye to ensure they were reasonably sensible, and so that small sequence fragments that might reduce alignment quality and be difficult to place phylogenetically were removed. Several gene fam-

ilies were excluded at this stage, and some large gene families split into subsets. This rapid approach was chosen to allow our methods to be scaled-up to much larger amounts of data. It is important to note that many gene families only contained sequences from a few species, and that some pairs of genera never co-occurred in the same gene family.

4.2.2 Gene tree parsimony

The species phylogeny minimising the total number of duplications on the gene family trees was found using GENETREE (Page, 1998), constrained to only consider trees supporting the monophyly of the two genera each of lampreys, hagfish, lungfish and rays. 50 heuristic searches were performed from random starting trees, with the 'steepest ascent' option and using alternate NNI and SPR branch swapping (Page and Charleston, 1997b). The same analysis but minimising the total numbers of duplications and losses was also performed. Note that because each of the gene family trees is rooted, the species tree found by this procedure is also a rooted tree.

4.2.3 Confidence in species tree nodes

Current implementations of reconciled trees have lacked any method to take account of uncertainty in gene family trees and express confidence levels in the reconciled species tree (Page and Cotton, 2000). To calculate support values on nodes, 100 pseudoreplicate alignments were generated for each gene family using the bootstrap (Felsenstein, 1985), and phylogenies for each replicate constructed exactly as described above. The species tree minimising the number of gene duplications was then found for successive trees from the bootstrap profile of each gene family, producing 100 species trees. Each search was performed from a single random starting tree, using the same options as the main gene tree parsimony analysis but only finding a single shortest tree for each replicate. Support values analogous to standard bootstrap values could then be calculated for nodes in the species tree.

4.3 Results

The results of our gene tree parsimony analysis are shown in figure 4.4. 50 heuristic searches found the same island of three equally-parsimonious shortest trees 19 times. Figure 4 also shows the majority rule consensus tree of the 100 species trees from gene tree parsimony analysis of the bootstrap profile of gene trees. Our phylogenies differ very little from traditional views of vertebrate relationships. Relationships within the major terminal groups are reconstructed identically to recent phylogenetic analyses for the teleosts (Nelson, 1994) and chondrichthyes (Maisey, 1984). Interestingly, we get very good support for the 3-taxon relationship between *Mus*, *Bos* and *Homo*, agreeing with the largest study of mammalian phylogeny (Liu et al., 2001) but disagreeing with a recent molecular study (Murphy et al., 2001). There is ongoing difficulty in resolving many ordinal-level relationships within the placental mammals (Waddell et al., 1999).

There are two main competing hypotheses about the relationship between hagfish, lampreys and the higher, jawed vertebrates or gnathostomes. Our analysis very strongly supports a close relationship of hagfish and lampreys, with these groups together forming a sister clade to the gnathostomes, called the cyclostomes. The other popular alternative unites lampreys and vertebrates as a 'Vertebrata' group, which together with the hagfish forms the 'Craniata'. Traditional classifications included the cyclostome group, but the first cladistic studies of the group led to a new view of the group (Janvier, 1981; Løvtrup, 1977) and eventually to a consensus among morphologists supporting the alternative Vertebrata group (Forey and Janvier, 1993; Janvier, 1996). In contrast, molecular phylogenies have consistently supported a cyclostome group, with evidence from 18S and 28S rRNA molecules (Mallatt and Sullivan, 1998; Stock and Whitt, 1992) and a number of nuclear loci (Kuraku et al., 1999). Evidence from mitochondrial genomes has been somewhat equivocal – a maximum-likelihood analysis of the hagfish mitochondrial genome sequence (Rasmussen et al., 1998) supported the lamprey and gnathostome clade, and a subsequent analysis (Delarbre et al., 2000) found that the position of the hagfish depended on the method of analysis used. Recent evidence from additional sequence data strongly supports cyclostome monophyly (Delarbre et al., 2002). There is also some other molecular evidence supporting a lamprey and gnathostome clade (Gursoy et al., 2000; Page, 2000; Suzuki et al., 1995), but our results

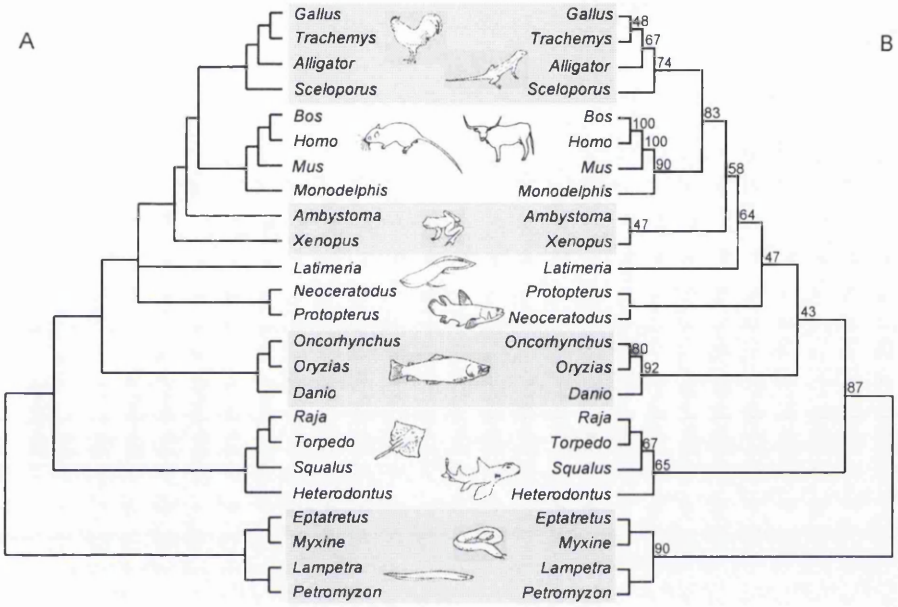

Figure 4.4: Phylogenies of vertebrates reconstructed using gene tree parsimony on a set of 118 nuclear genes. Alternate bands of shading and non-shading identify traditional higher taxonomic groups of vertebrates. Part (A) shows the strict consensus of 3 most parsimonious trees, each requiring 1380 gene duplications to fit the gene family trees. Part (B) is the majority-rule consensus of 100 bootstrap replicated as described in section 4.2.3, with figures on nodes being bootstrap percentages from this analysis.

show that nuclear gene loci strongly support a cyclostome clade, adding weight to a recent morphological re-evaluation of basal vertebrate relationships (Mallatt, 1997).

Another area of considerable debate is the relationship between lungfish, coelacanths and the tetrapods. The traditional taxonomy placed the fossil coelacanths as the closest relative of tetrapods, uniting them in the paraphyletic group Crossopterygii along with a number of other fossil taxa, but the discovery of the extant coelacanth *Latimeria* revealed many un-tetrapod like features (Forey, 1988), casting doubt on how conclusive the morphological data really is (Janvier, 1998). We find the coelacanths as closest relative to the tetrapods, but bootstrap support below 50% shows that this node is essentially unresolved. Evidence from mitochondrial genome sequences has been ambiguous, depending on the phylogenetic method used (Zardoya and Meyer, 1997) and often misplacing both lungfish and tetrapods completely (see figure 4.3a and b).

Finally, we have an unusual result for the phylogeny of the reptiles (taken to include the birds). The bulk of morphological and palaeontological evidence groups alligators and birds with the extinct dinosaurs as the archosauria, with lizards forming the sister group to this clade and turtles most basal. This has been challenged by data placing turtles as the sister group to the lepidosaurs (Rieppel and deBraga, 1996) and molecular data, which seems to unanimously place turtles as relatives of archosaurs (Hedges and Poling, 1999; Rieppel, 2000). A number of recent reviews (Rieppel, 2000; Zardoya and Meyer, 2001a) have concluded that relationships within the reptiles are still uncertain. The results of our analysis are unconventional in placing turtles as the closest relative of birds, but add to the molecular evidence placing turtles within crown-group diapsids.

4.4 Discussion

The gene tree parsimony method makes a number of assumptions about the process of gene duplication that may be important in this context. Firstly, the correct inference of gene duplications and losses on a gene tree requires that the gene tree is known without error. This is a potentially important problem that has been widely recognized (Page, 2000; Page and Cotton, 2000) which we have dealt with by using a bootstrap profile of trees for each gene family.

We also make some assumptions about the process of gene duplication, as the number of duplications and losses is assumed to be the minimum required to fit the gene tree into the species tree. If duplications and losses are frequent, there may be lineages that originated in a duplication event and were then lost, leaving no trace in extant genomes. These numbers could thus be a significant underestimate of the true number of duplication and loss events, but should not introduce any systematic bias in the optimal species tree.

Another important issue is that failure to sample – where a gene has simply not been sequenced from an organism – is conflated with gene loss (where the gene is actually deleted from the genome). This has no effect on the optimal species tree under a duplication-only criterion, but could lead to artefacts under the duplication and loss criterion, where species can cluster on the basis of this failure to sample (Page, 2000; Page and Charleston, 1997a). We would advise against duplication and loss as an optimality criterion in data where this problem is likely to be very significant, although in fact the optimal species tree under the duplication and loss criterion for our data differs little from the minimum-duplications tree, placing *Latimeria* as sister taxon to an amphibian clade at the base of the tetrapods, and grouping *Trachemys* with *Alligator* rather than *Gallus*.

Finally, our method assumes that gene duplication and gene loss are the only processes introducing disparity between gene and species trees. Gene duplications have clearly been important in vertebrates, as shown by the existence of many complex gene families in vertebrate genomes (Page, 2000), but we cannot rule out that other processes might introduce incongruence between gene and species trees. The frequency with which genes will fail to coalesce between speciation events (deep coalescence) will depend on both the effective size of the population the alleles are present in, and the time between speciations. If we imagine the width of branches to be effective population size, long, thin branches should show few, if any, failures to coalesce, while short, fat branches should show many failures to coalesce (Pamilo and Nei, 1988). We have no information about effective population sizes, but all the branches on our phylogeny are very long in population genetics terms – molecular clock divergence dates suggest that the split between *Homo* and *Bos* is probably around 92 million years old, and that between birds and crocodilians about 222 million years (Kumar and Hedges, 1998). There are very few reliable reports of horizontal gene transfer in eukaryotes (Syvanen, 1994), so we can rule

out any large-scale effect from horizontal transfer in our data set.

Any study attempting to infer species phylogenies from gene phylogenies of multiple loci needs to take into account the potential problem of paralogy. As large-scale sequencing projects produce genomic sequence data from an increasing number of taxa, we believe that the issues discussed in this paper will become of increasing importance to systematists, and that reconciled tree methods will become more widely used. Gene tree parsimony is fast enough to scale-up to analysis of whole genomes and even whole genetic databases, raising the possibility of effective automated phylogenetic reconstructions from molecular data (Page and Cotton, 2000).

4.5 Conclusion

We have shown that reconciled trees can successfully reconstruct phylogeny in the presence of a mixture of orthologous and paralogous genes. In contrast to evidence from mitochondrial sequences, our results largely agree with traditional views on vertebrate phylogeny, but add new evidence to support some controversial ideas, such as a monophyletic cyclostome group. The techniques described in this paper should scale-up to genome-scale comparisons, so we hope that this success will encourage systematists struggling to reconstruct credible phylogenies from the vast amounts of genomic data now accumulating (Brown, 1996).

4.6 Supplementary Material

The data used in this study is available from `http://darwin.zoology.gla.ac.uk/~jcotton/vertebrate_data`. This includes a complete list of the gene families used in this paper, with phylogenies and alignments for each, along with the GENETREE input file for the analysis.

Chapter 5

Primate Gene Family Evolution

An Integrated Study of Phylogeny and Evolution by Gene Duplication

Abstract

Reconciled tree methods enable the inference of species trees from a set of gene family trees. A less well-described property of these methods is that they allow us to study the events that introduce incongruence between phylogenies from different loci. In current reconciled tree methods, these events are gene duplication and gene loss. Phylogenetic inference and an understanding of these evolutionary events are closely related, so integrated studies investigating both of these aspects are possible. This study represents the first attempt to integrate phylogenetic inference from a substantial set of gene families with investigation of gene family evolution in these gene families. We focus on primates, using 69 gene families to construct a framework of primate phylogeny from molecular data. Our results confirm current ideas about primate relationships and establish a molecular timescale for primate evolution. We also present the first data on the temporal pattern of gene duplications during primate evolution.

5.1 Introduction

The explosion in the availability of molecular data over recent years has led to a similar explosion of interest in methods for combining multiple sources of molecular data in evolutionary studies. In molecular systematics, methods for combining results from multiple loci include combined analysis, whether in a parsimony (Kluge, 1989) or likelihood (Sullivan, 1996) framework, consensus methods (Swofford, 1991) and supertree methods (Sanderson et al., 1998). Gene tree parsimony is a less well-known alternative, which uses reconciled trees to explicitly study the differences between a set of gene family trees and an estimated species tree.

Reconciled trees are a general method for reconstructing the evolutionary history of an association between two evolving entities. The method emerged first in molecular systematics (Goodman et al., 1979), but later found applications in biogeography and in parasitology (Page, 1994a). Its application in molecular systematics has grown from the increasing realisation that the relationship between gene trees and species trees is more complex than simple equivalence – one cannot simply re-label the leaves of a gene tree with the names of the equivalent species (Doyle, 1992; Maddison, 1997). Molecular processes such as gene duplications, gene losses or lateral gene transfer can introduce differences between the gene tree topology and the correct topology for the species included. By interpreting incongruence with a proposed species tree as being due to these processes, reconciled tree methods allow us to study the processes themselves. They also give a biological meaningful measure of similarity between two trees – a property taken advantage of in the method that has become known as 'gene tree parsimony' (Slowinski et al., 1997). From a set of gene family trees, we can find a species tree minimizing the number of evolutionary events needed to explain the difference between each gene tree and the species tree. Heuristic methods are then used to find a species tree that is most compatible with the gene family trees. The utility of reconciled tree methods to infer species trees has been shown recently by a number of authors (Cotton and Page, 2002; Martin and Burg, 2002; Page, 2000).

As primates are the group of organisms containing our own species, there has long been a considerable interest in their biology, including their systematics and evolution. Morphological and palaeontological data together has contributed to a

view of primate systematics that splits the group into two taxa, the Prosimii, including the bushbabies, lorises, lemurs and tarsiers, and the Arthropoidea – the monkeys, apes and humans. There is a considerable amount of molecular evidence on primate phylogeny (see table 1 in Page and Goodman, 2001), which has been extensively reviewed (e.g. Goodman, 1999; Goodman et al., 1998). Most of this work is based on sequences of various globin genes (e.g. Porter et al., 1997), but other genes such as von Willebrand factor (Porter et al., 1996), Alu repeats (Zietkiewicz et al., 1999) and even non-coding DNA (Page and Goodman, 2001) have also been used as markers for the relationships between major primate clades. Despite this, there has been little work on integrated studies incorporating this evidence, with the exception of one supertree study (Purvis, 1995a), and our study is probably the largest amount of molecular data on primate phylogeny integrated into a single analysis.

Molecular data have largely confirmed previous classification within the group, recognising an Old World group of great apes and monkeys, the Catarrhini and a New World sister group, the Platyrhinni. Perhaps the most serious conflict between molecular evidence and the morphological data is over the relationship of the tarsiers – this group has alternatively been placed as a sister-group to the Strepsirhini, as sister-group to both the Strepsirhini and Haplorhini, or as sister-group to the Haplorhini (Shoshani et al., 1996), with support for the first two alternatives being largely from the fossil record, and support for the latter being largely molecular (Goodman et al., 1998; Koop et al., 1989; Zietkiewicz et al., 1999), and from the morphology of *Tarsius* (Groves, 1975). This conflict between neontological and palaeontological data may largely be due to the fact that *Tarsius* is something of a living fossil, as the sole extant, derived representative of a once diverse group of Eocene Tarsiiformes. Shoshani et al. (1996) list 8 potential morphological synapomorphies of a prosimian group of tarsiers and the lemurs and lorises, but argue that many of these characters are plesiomorphic characters retained by both of these groups but lost in the Haplorhines.

There is less disagreement about most other primate relationships – the debate about the closest living sister group to humans has fairly decisively concluded that the two chimpanzee species are our closest living relatives, with the *Gorilla* species being a more distant outgroup (see references in Goodman, 1999; Koop et al., 1989). All evidence suggests that these three taxa are very closely related. There is

comparatively little data on lemur relationships, but with this exception the primate tree represents a fairly well-known phylogeny to study using a relatively untested method.

Reconciled tree methods rely on making an explicit statement about how molecular events such as lateral gene transfer, gene duplication and gene loss have introduced incongruence between gene trees and a species tree. In fact, in available reconciled tree methods, only gene duplication and gene loss are included, as the action of lateral gene transfer is rather more difficult to account for (Charleston, 1998). The reconciled trees approach can thus provide an insight into the pattern of gene duplications in the gene families used, a matter of increasing interest to evolutionary biologists. Page and Cotton (2002) discuss how the reconciled tree method identifies nodes in the gene family trees as representing either speciation or duplication events. This identification of each node as representing a particular sort of splitting event allows separate investigation of the evolutionary pattern of these two events.

There have been some recent attempts (Gu et al., 2002; Page and Cotton, 2002) to reconstruct the historical pattern of gene duplications in vertebrates, but no investigation has focused on the pattern of duplications on a smaller phylogenetic scale. There has been some interest in evolution by gene duplication as being the creative force behind evolutionary innovations in human evolution (Bailey et al., 2002; Ohno, 1970), and there is evidence that adaptive selection of particular gene families may have been important in the emergence of humans and the African great apes (Johnson et al., 2001). Work on the pattern of gene duplications in recent human evolution has focused on the spatial pattern of duplications within the genome rather than on the timing of duplications (Bailey et al., 2002). We attempt to reconstruct the pattern of gene duplications in the gene families used here.

An additional context in which an appreciation of paralogy is important is in molecular dating of evolutionary events on phylogenetic trees. Using nodes representing gene duplications to date splits will overestimate the age of lineages (Figure 5.1), so ignoring possible paralogy will lead to a general bias towards overestimating divergence dates. This is particularly interesting given that most large analyses published to date (Kumar and Hedges, 1998) have produced estimates much older than estimates from the fossil record for a number of different groups, sometimes remarkably so (e.g. Heckman et al., 2001). There is thus a serious discrepancy

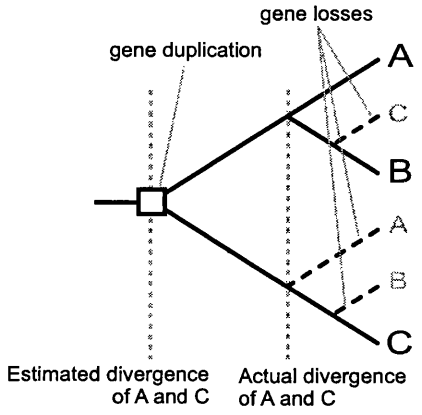

Figure 5.1: How paralogy can alter estimates of divergence dates. Gene duplication and gene loss events can affect tree-based estimates of divergence dates if the date of the gene duplication event rather than the actual speciation event is estimated. This will lead to over-estimates of divergence dates.

between dates from fossil and molecular sources – a discrepancy that seems too large to be easily explained by the poverty of the fossil record (Morris, 1999).

A number of studies have suggested that this may be due to problems with date estimation techniques, such as a bias towards over-estimating divergence times (Rodriguez-Trelles et al., 2002), which of course will only ever be approximate, as the rate of molecular evolution is ultimately 'dependent on the fickle process of natural selection' (Ayala, 1999). Here, we estimate dates for divergences within the primates, using reconciled trees to explicitly identify duplications nodes and so removing the potential bias due to mistaken orthology of gene copies.

Another major potential error in molecular clock dating is poor use of calibration points – these have often been fairly arbitrarily selected point estimates, not employing statistical methods to correctly estimate earliest common ancestors from the fossil record (Tavarè et al., 2002). There are no generally accepted calibration points within the primate tree, so we compare results from two different calibration points – one based on the fossil record and one from a previous molecular study. There have been several previous studies looking at dating evolutionary events within the primates (Penny et al., 1998; Stauffer et al., 2001; Yoder et al., 1996; Yoder and Yang, 2000), allowing us to compare our results with other studies

that do not take gene duplications into account.

Here, we aim to show how reconciled tree methods can both reconstruct the phylogeny of a group and give some insight into the pattern of gene duplications and speciation events across that phylogeny.

5.2 Materials and methods

5.2.1 Data gathering

We selected primate genera to ensure maximum taxonomic coverage of the primate groups, and to maximise the number of gene families that would be available for analysis. To this end, we parsed the HOVERGEN database (Duret et al., 1994) and selected from each primate family the genus that was present in the largest number of nuclear gene families. Where a few different genera within a family were similarly well sampled by HOVERGEN families, we included several representatives of the family. The genera included in the paper are listed in table 5.1. We used all nuclear gene families from HOVERGEN that contained sequences from at least 3 of the selected primate genera and at least one of our outgroups (*Mus*, *Rattus*, *Xenopus*, *Gallus* and *Bos*). One additional gene family (MHC class I-related protein, FAM003540 in HOVERGEN) was included that lacked any of these outgroups, for which *Sus* was used as an outgroup. In this way, we found data on representatives of all the primate families except the Megalapidae, or sportive lemurs, which has only a single genus (*Lepilemur*, Cowlishaw and Clutton-Brock, 2001) and is represented very poorly in the HOVERGEN database. Because of the significant interest in hominid relationships, we included all 4 extant genera of the Hominidae. The large number and diversity of outgroups was used to increase the chance that non-primate sequences would be present to reveal paralogy of primate sequences and to increase the length of evolutionary time sampled by gene families, so that midpoint rooting would be likely to establish the correct root for each family (Figure 5.2).

5.2.2 Gene family phylogenies

For each gene family, specially written software was used to extract the appropriate amino acid sequences from flat files of the HOVERGEN database to FASTA

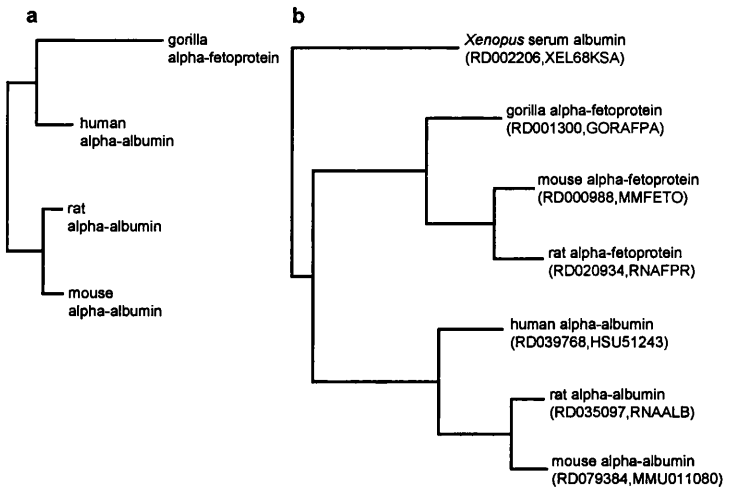

Figure 5.2: Multiple outgroups can help resolve complex paralogy relations. (a) and (b) show phylogenies for α-fetoprotein sequences from gorilla, human, rat and mouse. In tree (a), only a single set of outgroup sequences were used and gorilla and human sequences appear to be orthologous, but including additional sequences reveals the paralogy of the two primate sequences, as shown in tree (b). Both trees are midpoint rooted. Figures in brackets are locus codes from the HOVERGEN database (Duret et al., 1994) and locus names from the EMBL sequence database.

Order Primates		
	Suborder Haplorhini	
		Infraorder Platyrrhini
			Family Callithricidae	*Callithrix, Saguinus*
			Family Cebidae	*Aotus, Saimiri*
		Infraorder Catarrhini
			Family Cercopithecidae	*Cercopithecus, Macaca, Papio*
			Family Hominidae	*Gorilla, Homo, Pan, Pongo*
			Family Hylobatidae	*Hylobates*
		Infraorder Tarsiiformes
			Family Tarsiidae	*Tarsius*
	Suborder Strepsirhini	
			Family Cheirogaleidae	*Microcebus*
			Family Daubentoniidae	*Daubentonia*
			Family Galagonidae	*Galago, Otolemur*
			Family Indriidae	*Propithecus*
			Family Lemuridae	*Eulemur, Lemur*
			Family Loridae	*Nycticebus, Perodicticus*
non-primate outgroups		
	Mammalia, Artiodactyla	*Bos, Sus*
	Mammalia, Rodentia	*Mus, Rattus*
	Aves, Galliformes	*Gallus*
	Amphibia, Anura	*Xenopus*

Table 5.1: Simplified classification of genera included in this study. Primate classification follows Cowlishaw and Clutton-Brock (2001).

format files. Short sequence fragments may have an adverse effect on automatic sequence alignment algorithms, so sequences less than 20% the mean length of the sequences from their gene family were removed, except in a few instances where this would lead to the removal of a sequence from a poorly-sampled genus. Alignments were then constructed using ClustalW (Thompson et al., 1994) with the default parameters of gap opening penalty 10, gap extension penalty 0.1 for the pairwise alignment step, and gap penalties of 10 and 0.2 respectively for the multiple alignment step.

Gene family phylogenies were constructed from these alignments using a number of different methods. Neighbor-joining (Saitou and Nei, 1987) trees were constructed using ClustalW using uncorrected, gapped distances. Parsimony trees were constructed using PAUP* 4b10 (Swofford, 1998). These trees were found using 10 separate heuristic searches from different random addition-sequence starting points, holding multiple trees but setting a maximum of 5000 trees to be held in memory at any time. To produce the single, fully-resolved tree needed by the reconciled tree algorithm, the maximum parsimony trees found in the first stage were rooted and the majority-rule consensus tree, incorporating all compatible components, of these rooted trees used. The majority-rule consensus tree has the desirable property of being the median tree of the trees found in stage 1, under the symmetric difference metric (Margush and McMorris, 1981). Finding the binary median tree is NP-complete (McMorris and Steel, 1993), but the consensus tree incorporating compatible components will be close to this tree, and so is a good summary of the trees found. This search strategy was designed to allow a thorough search of tree space, while only returning a single fully-resolved phylogeny for each gene family. Finally, a maximum-likelihood estimate of distances between each pair of taxa was found using TreePuzzle v.5.0 (Schmidt et al., 2002), using the model selected by the program, with amino-acid frequencies estimated from the data and using an 8-category approximation to a gamma distribution to model rate heterogeneity between sites. These distances were then used to produce a minimum-evolution tree in PAUP 4b10.

All gene family trees were midpoint rooted initially, then checked manually to ensure that the root was in a sensible position relative to our outgroups – the midpoint root was accepted if the root was either between the outgroup and ingroup, or between clades of orthologs containing both outgroup and ingroup gene copies.

In a few gene families, outgroup rooting was performed manually to ensure this.

5.2.3 Gene tree parsimony

The species phylogeny minimising the total number of duplications (or duplications and losses) on the gene family trees was found using GENETREE v1.3, constrained to only consider trees supporting the monophyly of the primates as a whole, and of the Lemuridae, Galagonidae and Loridae (the constrained nodes are shown on figure 5.3). All searches were from random starting trees, with the 'steepest ascent' option and using alternate NNI and SPR branch swapping. Other details of the GENETREE analyses varied slightly. For most analyses, multiple trees were swapped on simultaneously, allowing a more thorough search of the tree landscape, and only the number of random starting points was varied, to ensure that the minimum-cost island of trees was found sufficiently often to give some confidence that it was not merely a local optimum. In one analysis – under the duplication-only criterion for parsimony gene trees – too many equal-cost trees were present to allow this strategy, as the search became trapped in islands of sub-optimal trees. For this analysis, 500 random starting-point replicates were used, but keeping only a single tree at a time for branch-swapping. 101 of these searches found trees of the shortest length identified (582) finding 100 different trees. To sample a wider region of this tree island, 10 of these trees were used in searches, holding multiple optimal trees but terminated once 500 trees had been found. None of these searches found trees shorter than 582 duplications, and the Adams consensus of the search results from these 10 searches were essentially identical – differing only in how well resolved the position of *Propithecus* was, suggesting that we have successfully sampled from across this island.

Note that because each of the gene family trees is rooted, the species tree found by this procedure is also a rooted tree.

5.2.4 Bootstrap analyses

The standard gene tree parsimony analyses described above use only a single fully-resolved phylogeny for each gene family. These single trees may be a poor summary of the phylogenetic information for a gene family if there are many similarly good trees. To incorporate this information, we adopted a gene tree bootstrapping

protocol. We constructed a set of 100 bootstrap trees for each gene family in the dataset, using either neighbor-joining or parsimony methods in PAUP 4b10. For the parsimony bootstrap, trees were found using a single addition-sequence replicate, swapping on only a single tree to completion. The species tree minimising the number of gene duplications was then found for successive trees from the bootstrap profile of each gene family, producing 100 sets of species trees. Each search was performed from a single random starting tree, using the same options as the main gene tree parsimony analysis but only finding a single shortest tree for each of the 100 replicates. Support values analogous to standard bootstrap values could then be calculated as the number of times a particular node appeared in these 100 species tree.

5.2.5 Gene duplication distribution

Gene duplication events can occur at any scale, from small pieces of DNA to the entire genome duplicating in a polyploidisation event. Duplications on different gene family trees may thus be the result of the same multiple gene duplication event. To investigate this, we clustered gene duplications from individual gene families into the minimum number of sets that may represent these larger gene duplication episodes. A minimum set cover algorithm was used to find the smallest set of species nodes that could accommodate all the duplications required by the 69 gene families, identifying which gene duplications took place at each node in the species tree. This clustering algorithm is fully described in Page and Cotton (2002). This clustering can be thought of as the distribution of duplication events if we assume that duplications of any size occur with similar frequency

To examine the history of gene duplications without clustering them into large episodes of duplication, we also reconstructed the most probable distribution of duplication events under the assumption that duplications occurred independently. For each branch of the species tree, the most probable number of duplications actually occurring at that location was found by summing the number of duplications that were reconstructed as occurring on that branch weighted by the uncertainty in the duplication's position. A duplication that was reconstructed as occurring at a particular location added 1 to the number of duplications occurring at that location, while a duplication that could have occurred on any of three different branches ad-

ded $\frac{1}{3}$ to the estimate for each of the three branches.

5.2.6 Molecular dates for gene duplications and phylogeny

Trees produced using maximum-likelihood distances, as described above, were used for all dating analyses. Ultrametric trees were produced from these phylogenies by using the non-parametric rate smoothing method (Sanderson, 1997) implemented in the r8s software package, v1.50, with calibration based on dates for two alternative nodes – the divergence of rodents and primates and the divergence of Humans and Old-World monkeys. The date used for the first calibration point was a molecular estimate of 110 mya (Kumar and Hedges, 1998). The date for the second calibration point was based on an estimate that Humans and Old-World monkeys (Hominoidea and Cercopithecidae) diverged at the Oligocene-Miocene boundary at about 23 mya, a date just prior to the earliest known fossils of both groups (see discussion in Stauffer et al., 2001) and very close to Kumar and Hedges' estimate for the same event. Only a single calibration point was used at one time, allowing the results from these two different dates to be compared. For a particular calibration point, all nodes representing the relevant speciation event were constrained to the same age, so there were multiple calibration points in a number of gene families. Similarly, some gene families had no nodes mapping to that particular speciation, and so were not available for estimating dates based on that calibration point. These ultrametric trees were then analysed in a special version of the GENETREE program, where dates were output separately for each node on the species tree, and for duplications mapped onto each branch on the species tree, showing the pattern of gene duplication events through evolutionary time. Analyses were performed assuming the phylogeny shown in figure 5.3b, the best-resolved phylogenetic result, except that the effect of substituting the generally accepted relationships within the Hominidae was also assessed.

5.3 Results and discussion

5.3.1 Phylogenetic results

Our phylogenetic results are based on six different analyses – using source gene trees built by neighbor-joining on uncorrected distances, parsimony and minimum-

evolution on maximum-likelihood distances, and combining these estimates using gene tree parsimony under both duplication-and-loss and duplication-only optimality criteria. The results of all these analyses are shown in figures 5.3 and 5.4. For ease of interpretation, figure 5.3a identifies the higher primate taxa shown in table 5.1. Note that it seems likely that duplication-and-loss results will always be more resolved and compatible with duplication-only results (Page and Charleston, 1997b).

Figure 5.3a shows the strict consensus of 2 optimal trees found under the duplication-only criterion on NJ gene trees. These trees differ only in the relationships within Cercopithecidae. The same 2 trees were found by 24 out of 25 addition-sequence replicates, and imply 564 gene duplications in the gene trees. Figure 5.3b shows the single optimal tree, of cost 1123, found by 19 out of 25 replicates under the duplication-and-loss criterion on the same gene family phylogenies. Figure 5.4a shows the single optimal tree under the duplication-and-loss criterion on ME-ML trees. It has a cost of 1462, and was found by 16 out of 100 replicates. Figure 5.4b shows the Adams consensus of 156 trees requiring 602 duplications, found by 15 out of 25 addition-sequence replicates. Figure 5.4c shows the 2 optimal trees (of cost 1212) found by 24 out of 100 replicates performed under the duplication-and-loss criterion on parsimony gene trees. The search strategy used to construct figure 5.4d is described in detail under the methods section above. The trees found have a cost of 582 duplications. Figure 5.5 shows the results of the gene tree bootstrap analysis

The results of our analyses are broadly congruent with each other, but we will highlight differences between the results of similar analyses. Firstly, there is some instability both between and within analyses in the position of *Daubentonia*, *Microcebus* and *Propithecus*, which is probably explained by the relative paucity of data for these three taxa – they are represented by only 1, 2 and 3 gene families, respectively. Of more interest is the fact that *Tarsius* is grouped with the Haplorhini in both NJ and ME-ML analyses, but grouped with the Strepsirhini in the parsimony analysis, so that our three analyses have produced both of the two most previously supported relationships for the tarsiers. Bootstrap figures should be taken to indicate relative support for different nodes, and may not be entirely comparable to traditional bootstrap values. Note that instability of certain taxa is probably responsible for the low bootstrap values across this tree, as leaf stability

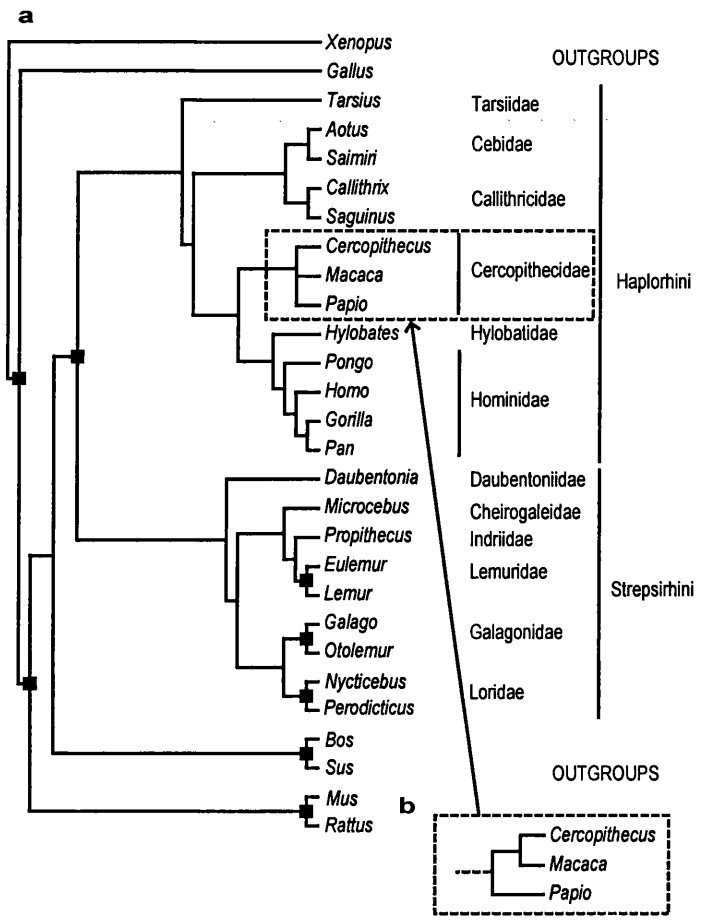

Figure 5.3: Results of analysis with gene family phylogenies inferred using neighbour-joining on uncorrected distances (NJ-D). (a) shows the results of the analysis under the duplication-only criterion, while (b) shows the resolved relationship between genera of the Cercopithecidae under the duplication-and-loss criterion. Nodes marked with a square were constrained to be present in the results of the analysis.

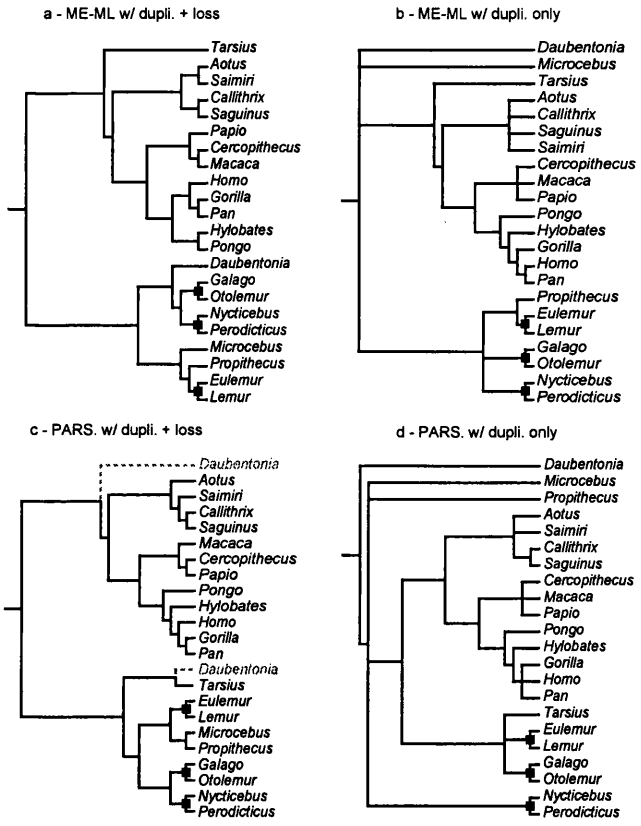

Figure 5.4: Results of analysis of gene family phylogenies inferred using minimum-evolution on maximum-likelihood distances (ME-ML) and parsimony (PARS) gene family trees. (a) and (b) are from analyses of ME-ML gene trees. (a) The single optimal species tree under the duplication-and-loss criterion. (b) Adams consensus of 156 optimal trees under the duplication-only criterion. (c) and (d) are trees from analyses of PARS gene family trees. (c) The 2 optimal trees under the duplication-and-loss criterion. The 2 optimal trees differ only in the position of *Daubentonia* and the two alternate positions are indicated by dashed grey branches on this tree. (d) Adams consensus of 5000 optimal trees found under the duplication-only criterion. In all figures, nodes marked with a square were constrained in the analysis – these are the same constraints as used in figure 3. To save space, outgroup taxa have been removed from all trees, as relationships within the outgroup were, in all cases, identical to those shown in figure 3.

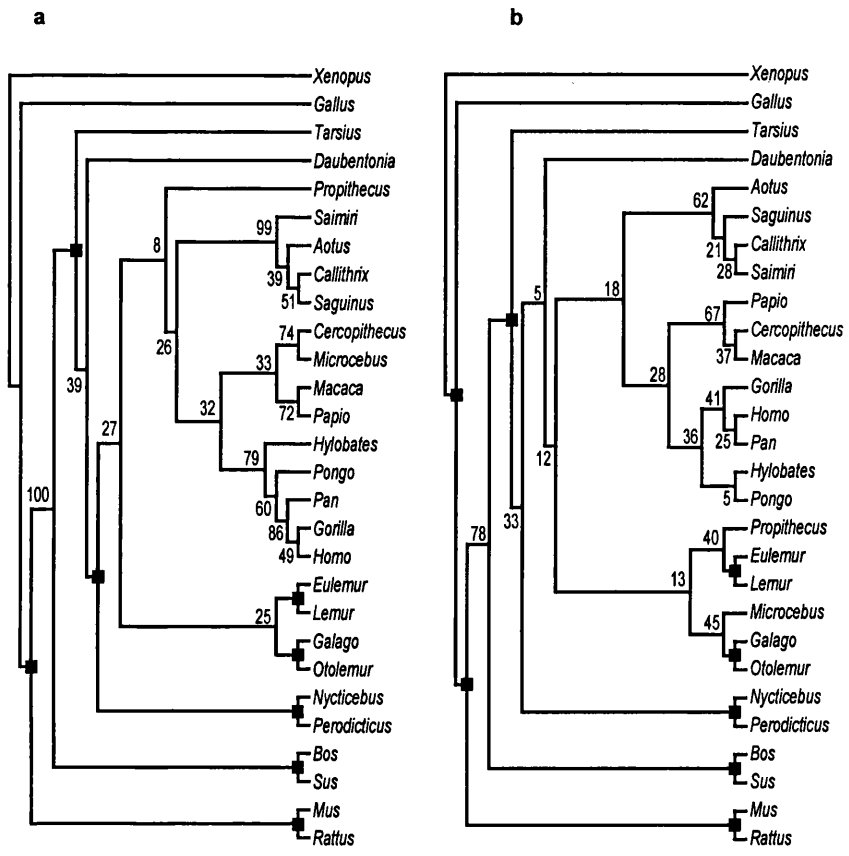

Figure 5.5: Gene tree bootstrapping analyses, with (a) NJ bootstrap gene trees and (b) Parsimony bootstrap gene trees. The trees shown are the majority-rule consensus of the 100 species trees found, incorporating compatible minority components, with bootstrap values shown at each node. Nodes marked with a square were constrained to appear in all species trees from the analysis, and so have no meaningful bootstrap values.

analysis (Thorley and Wilkinson, 1999) identifies *Daubentonia*, *Microcebus* and *Propithecus* as being significantly less stable across the two bootstrap profiles of trees than other taxa, followed by the other lemurs.

In conclusion we suggest that the results from the Neighbor-Joining gene trees represent a fully-resolved picture that roughly summarises our results, with the caveat that we have little information about the correct phylogenetic position of *Daubentonia*, *Microcebus* or *Propithecus* – the position of these taxa in particular differ between the parsimony, ME-ML analyses and on the bootstrapping results.

Perhaps the most disappointing result is our failure to resolve the correct relationship between human, chimpanzee and gorilla. This may be because of the extremely close relationship between these three taxa – humans and chimpanzees are only around 1.2% divergent (Chen et al., 2001) at the DNA level, and around 95% of base pairs are exactly shared between humans and chimpanzees when indels are taken into account (Britten, 2002). We would expect differential sorting of alleles, or deep coalescence, among three lineages to be relatively common over the short divergence time between humans, chimpanzees and gorillas, leading to incongruence of a different kind to that from gene duplication and loss, and potentially misleading our analysis.

Apart from the closest human relatives, our phylogenetic results, where resolved, are largely congruent with previous work, but there are some small differences. Page and Goodman (2001) find ((*Macaca*,*Papio*),*Cercopithecus*) from noncoding DNA evidence, agreeing with the supertree of Purvis (1995a). Purvis has presented the only previous phylogenetic hypothesis seeking to integrate evidence from a large number of different sources, in an MRP supertree of the primates. His aim was rather different from ours, in that he wanted to establish a species-level phylogeny for the group, but the results largely agree with our own. There are two other differences – Purvis's tree places *Daubentonia* with the Indriidae and Megalapidae – a position supported by some workers (Cowlishaw and Clutton-Brock, 2001), but in conflict with other molecular and morphological data (Yoder et al., 1996), and shows the Callithricidae nested inside a paraphyletic Cebidae, which contradicts most other evidence (Shoshani et al., 1996).

Our results agree with Shoshani et al.'s morphological work (Shoshani et al., 1996) on relationships within the Haplorhini, but clearly differ in rarely finding *Daubentonia* as part of a clade with the other Madagascar lemurs. Our results lend

some support to the view that this unusual primate may be the sister-group to other extant strepsirrhines (Groves, 1989), a view which is in conflict with more recent evidence on strepsirrhine phylogeny (Yoder et al., 1996) and with biogeography, and may reflect rather poor sampling of nuclear genes from this organism. Interestingly, our difficulty in resolving the relationships of *Microcebus*, the only cheirogaleid in our analysis, reflects the other major difficulty in Strepsirrhine taxonomy – early workers placed this taxon with the Afro-Asian loris group, but molecular work has placed the Cheirogaleidae in a clade with the other Malagasy species (Yoder et al., 1996).

5.3.2 History of gene duplications

Figure 5.7 shows the history of gene duplications obtained by mapping individual gene duplication nodes onto one of the species trees. This gives a representation of the history of gene duplications through primate evolution independent of branch-length or molecular-clock calculations. We can see that branch-lengths in the clustered and unclustered distributions are similar, suggesting that the clustering procedure itself has had relatively little impact on the distribution of gene duplications across our tree, and is fairly reliable. Figure 5.6 shows the pattern of gene duplications along the human lineage through evolutionary time, using information from the branch lengths in the ME-ML gene family trees. These plots appear to show a pattern of roughly continuous gene duplication through time, with peaks of duplications at 40-50 mya and 80-90 mya. They certainly do not look like the exponential-like curves we would expect from a constant-rate process of lineage birth and extinction (Nee et al., 1995), as seen in recent work on vertebrate genome evolution (Gu et al., 2002), although there does seem to be some sign of an increasing number of duplications occurring more recently, which would be consistent with these models. A number of difficulties arise in interpreting these plots. Firstly we should be cautious of the effect that the taxonomic sampling of the gene families can have on the rate of duplication observed at a particular time. Secondly, and more fundamentally, only a single gene is ever available to estimate the age of single duplications, while experience from molecular clock studies of speciation suggests that individual genes can give very inaccurate estimates of dates. Accurate studies of the pattern of gene duplication through time will thus be

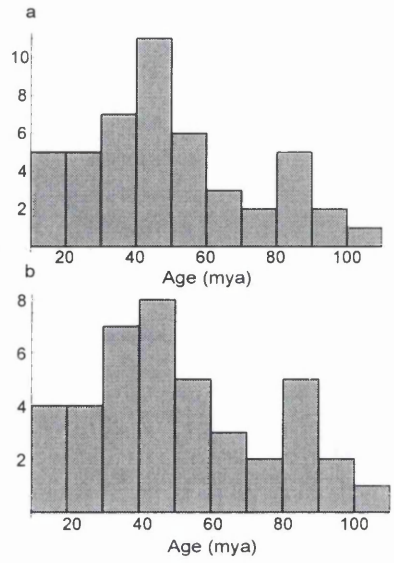

Figure 5.6: The distribution of gene duplications during primate evolution, inferred from the ME-ML phylogenies for our gene families. (a) Assuming the phylogeny is ((*Pan*,*Gorilla*),*Homo*) (b) assuming ((*Homo*,*Pan*),*Gorilla*).

difficult, highlighting the need for branch-length independent methods such as the clustering algorithm used here (Page and Cotton, 2002).

Figure 5.8 shows the distribution of molecular clock age estimates for two different divergences in the species tree, and for inferred gene duplications mapped to occur along the branch leading up to the divergences. It shows that the age distributions for these speciation events are approximately normal, as would be expected for the combination of multiple sources of error affecting the estimates in a non-biased way. The age distributions for duplication events look quite different. We interpret this to be because these individual dates are not multiple estimates of the same date, but estimates of the dates of independent gene duplication events. We would thus expect this kind of multi-modal distribution.

5.3.3 A molecular timescale for primate evolution

Figure 5.9 shows the ages of different events during vertebrate evolution. Note that different numbers of observations are included for different calibration points, as some gene families will fail to sample one or other of the calibration nodes.

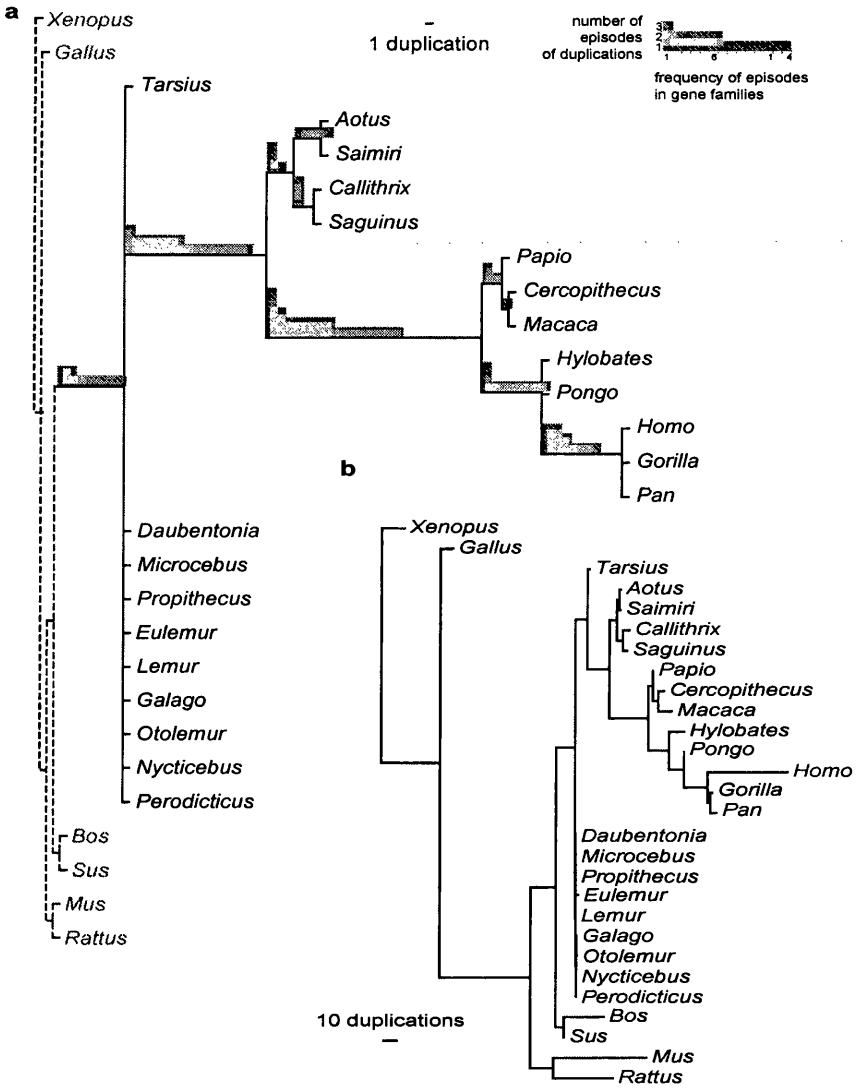

Figure 5.7: Distribution of gene duplications during primate evolution. The species tree is that of Figure 3b. In part (a), branch lengths represent the number of separate gene duplications inferred to have occurred along each branch. Stacked bars represent the number of distinct episodes of gene duplication in each of the gene families that have duplicated along the branch. For clarity, bars have been omitted where only a single duplication episode is inferred for each gene family. In (b) the branch lengths represent the most likely duplication distribution if all duplications are independent.

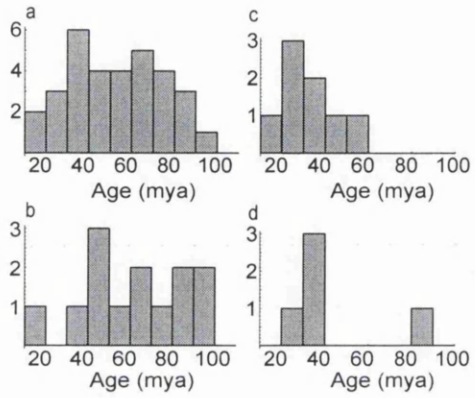

Figure 5.8: Distributions of age estimates for selected nodes on the primate tree. The figures are histograms for age estimates, in mya for the divergence of (a) the Old-World monkeys and apes (Cercopithecidea, Hominidae and Hylobatidae) and New-World monkeys (Callithricidae and Cebidae), (b) the dates of gene duplications mapped to the branch below this divergence; (c) the Hylobatidae and Hominidae, (d) the dates of gene duplications mapped to this branch.

Our two calibration points clearly give very different results from each other – inspection of the dates for the deepest split in our tree shows that the second calibration significantly over-estimates this date. In fact, the accuracy of our two calibrations seems to vary over the tree, when compared to existing date estimates and to the fossil record – the more recent calibration has performed well at estimating the dates of recent events, while the more ancient calibration gives more credible date estimates for more ancient events. For example, the first calibration estimates for the split between humans and chimpanzees, at 20.2 mya, is clearly much too high, whereas the 7.3 mya estimate from the second calibration seems fairly reasonable, particularly given recent human fossil finds (Wood, 2002). In contrast, the first calibration estimate of 78.1 mya for the common ancestor of the primates is only a little too recent – Tavarè et al. (2002) estimate 81.5 mya from a sophisticated analysis of the fossil record, whereas the second calibration estimate of 247.3 mya is clearly absurd (despite the very large standard error, an approximate 95% confidence interval for this estimate of ± 2 standard errors would still not include the other estimates). This pattern is to be expected, of course, but does underline the importance of calibration methods, and highlights the failure of this

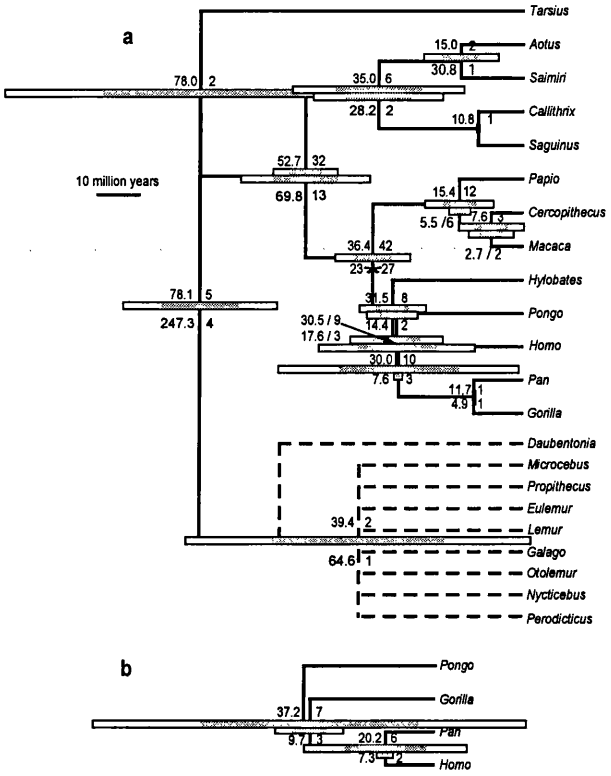

Figure 5.9: Age estimates for major divergences within the primates. (a) Using ME-ML source trees, and assuming the best-resolved phylogenetic result (Figure 3b). Tree drawn with branch lengths proportional to estimated dates under the 110 Mya Rodent-Primate calibration. Figures to the left of nodes are mean date estimates in Myrs, and figures to the right of nodes represent the number of gene tree nodes used to estimate these figures, above nodes based on the Rodent-Primate calibration, below based on the Hominidae-Cercopithecidae calibration. Shaded bars represent ± 1 standard error around the mean, white bars represent ± 2 S.E, based on the Rodent-Primate calibration point. The shaded branch lengths within the Strepsirhines indicate that estimates could only be produced for a single divergence within this clade, between *Daubentonia* and the other taxa, rather than suggesting that Strepsirhine taxa evolved simultaneously in an explosive radiation event. (b) is as above, but showing different estimates where the correct relationship within the Hominidae is assumed. For reasons of space, error bars are excluded from the second estimate for the divergence between Strepsirhines and Haplorhines – one SE around the mean is ± 49.2 myrs for this estimate.

study to accurately estimate divergence dates. It is possible that using multiple calibration points simultaneously would help resolve this issue.

Much work has focused on the error variance of estimates from single genes, but less has looked at the variance in estimates of the same dates from multiple genes – although a number of multi-gene analyses have been presented (Heckman et al., 2001; Kumar and Hedges, 1998). The most striking results from our analysis is that date estimates from different genes are extremely diverse, so that very little confidence should be placed in estimates based on a single gene or small set of genes.

Our results seem to suggest that sophisticated methods need to be used to determine correctly the confidence intervals around each estimate, but this has rarely been done in large-scale studies using many markers (e.g. Heckman et al., 2001; Kumar and Hedges, 1998). Methods such as non-parametric rate smoothing (Sanderson, 1997) and linearised trees (Takezaki et al., 1995) may be far more susceptible to non-uniform molecular clock dates than other methods (e.g. Rambaut and Bromham, 1998; Thorne et al., 1998; see review by Wray, 2001).

We suggest that a more sophisticated treatment of paralogy, such as we have attempted, might be expected to improve molecular clock estimates from such studies, but we have little evidence of that from these data. Further study investigating how much variance there is between apparently orthologous estimates of divergence dates from different markers, and investigating the reasons for this variance, is sorely needed. It is still unclear how much of the variance can be put down to the complexities of the evolutionary process, which can only ever be modeled in a simplified way, and how much is due to errors like paralogy and lineage sorting. This distinction is important because certain sources of error (such as paralogy) could produce biases in the estimates, rather than simply decreasing the accuracy of estimates, and so could help explain the frequent conflict between fossil-record and molecular dating techniques.

5.3.4 Reconciled trees

The strength of reconciled tree is that they can interpret differences between gene trees and a species phylogeny in terms of biologically meaningful events. This strength can also be a weakness, as all differences between the two trees is inter-

preted as being due to gene duplication and gene loss, ignoring inevitable errors in phylogenetic reconstruction for the gene families. This means that some of the gene duplications inferred on the gene trees will be false – due to incongruence from other sources – and this will, in turn, add noise to the gene duplication patterns and will reduce to number of nodes identified as speciation events, and so reduce the power of molecular clock date estimates.

Reconciled trees provide a framework to study both gene duplication and speciation events together, in a consistent manner, for both phylogenetic inference and molecular dating. Wherever nuclear genes are used for either of these two purposes, the species phylogeny and pattern of gene duplications need to be taken into account – they are mutually dependent and cannot be properly appreciated without reference to the other. A species phylogeny must be assumed to properly appreciate where gene duplications have occurred in gene families, when gene loss or incomplete sampling has complicated this inference. Knowledge of where gene duplications have occurred on gene family phylogenies has the potential to both improve the accuracy of molecular dating and to allow us to explicitly study the pattern of gene duplications through evolutionary time. This represents the first attempt to comprehensively cover both aspects in a single study.

Methods to integrate evolutionary evidence from a number of loci will increasingly be needed. Reconciled trees are at one extreme of a potential spectrum of methods, in that they take any incongruence at face value – as representing real evolutionary events, rather than as an error. Methods such as using a bootstrap profile of trees, as used here, can avoid this, but are somewhat unsatisfactory. If the flood of genomic data is to be fully utilised, systematic biology needs a more sophisticated understanding of the causes of incongruence between different genes, which will need more complex (and so more realistic) models of evolution to properly allow us to understand how much of this incongruence is due to interesting evolutionary events and how much is due to statistical sampling error. Such models are now becoming more widely available (although extending their use to the scale of even the smallest genomes remains a distant possibility).

5.4 Conclusion

Existing estimates of primate phylogeny appear to be correct, and are well-supported by data from a number of nuclear gene families. We also report data on the pattern of gene duplication and speciation events during primate evolution. Our results suggest that reconciled tree methods can provide credible estimates of phylogeny, but that reconstructing the pattern of duplication and speciation is more difficult.

Part III

INVESTIGATING PATTERNS OF GENE DUPLICATION

Chapter 6

Imbalance of Human Gene Family Phylogenies

Abstract

The shape of cladograms or other phylogenies has often been used in attempts to make inferences about the evolutionary process. Most of this work has involved comparing the shapes of actual phylogenies with expectations from simple models of the speciation process. Previous studies have focused almost exclusively on speciation events, but gene duplication is another lineage splitting event, analogous to speciation, and gene loss or deletion is analogous to extinction. Measures of the shape of gene family phylogenies can thus be used to investigate the processes of gene duplication and loss. I make a first attempt to use tree shape measures to study gene duplication, and investigate the "2R hypothesis" of two rounds of genome duplication in vertebrate evolution, using phylogenies for human genes. I find that gene duplication has produced gene family trees significantly less balanced than expected from a simple model of the process, but more balanced than for species phylogenies, which I suggest is due to regional duplications or genome duplications making individual duplication events on a tree non-independent.

6.1 Introduction

Traditional cladograms represent the evolutionary history of a group of organisms, with the leaves representing species or higher taxa group. In such trees, each node represents a speciation event. Molecular systematists generally assume that the phylogeny for a set of molecular sequences (a gene tree) is identical to the phylogeny of the organisms the sequences were obtained from (the species tree). This may not be the case, however – a number of processes can introduce differences between a correctly estimated gene tree and the correct species tree (Doyle, 1992; Maddison, 1997). These processes can affect any molecular phylogeny, but this is particularly obvious where gene duplication (Holland, 1999) has produced multiple sequences of a gene for a particular species. Gene duplication events generate families of related genes in genomes (Henikoff et al., 1997) leading to difficulties in inferring species relationships – the problem of paralogy (Slowinski and Page, 1999).

Molecular phylogenies for gene families (e.g. Figure 6.1a) usually display sequences from a number of species for different orthologous groups of sequences (Mindell and Meyer, 2001). These trees thus show a complicated tapestry of orthology and paralogy, and nodes on such trees may represent both gene duplications and speciations. Gene duplication events affect the form of such phylogenies in the same way as speciation events – both are splitting events, producing daughter lineages that henceforth have independent evolutionary histories (at least in the absence of gene conversion or introgression). Because both gene duplication and speciation are splitting events, represented by the internal nodes of a molecular phylogeny, we can use similar tools to study the two analogous processes, and in particular a number of techniques developed to investigate speciation and extinction may give some insight into the pattern of gene duplication and gene loss.

Tree shape has been used to make inferences about the processes of speciation and extinction that govern the birth and death of organism lineages (Mooers and Heard, 1997). Such inference requires an assumption that all of the nodes on the trees examined represent speciation events – i.e. that there is no hidden paralogy (Page and Charleston, 1997a). Similarly, in order to use tree shape to investigate the processes of gene duplication and gene loss, or deletion, we need to use phylogenies where all the nodes represent gene duplication events. As shown in Figure

6.1b, phylogenies containing homologous genes from a single genome have this property – here, sampling just the zebrafish genes will produce the tree on the right, which includes all but one of the duplication events present on the more complete tree (figure 6.1a). As figure 6.1 makes clear, we cannot be certain that such phylogenies will include all the duplications that have occurred during the evolutionary history of an organism, due to gene loss or deletion. To minimise this problem, and allow inferences about the processes of loss and deletion to be made without confounding these processes with the absence of a gene from the sequence databanks, it is preferable to use gene sequences from a completely sequenced genome.

The large literature on tree shape has focused on the cladistic balance of trees – how comb-like or bush-like the shape of the tree is – and has largely ignored information from branch lengths. In particular, a great deal of the work has investigated how closely the balance of real phylogenies (measured with one of a number of different indices) taken from the literature matches the balance expected under more-or-less simple models of the speciation process (Mooers and Heard, 1997). The simplest realistic model has become known as the Equal-Rate Markov model (ERM), based on models of the diversification process suggested by Yule (1924). Under the ERM model every lineage has an identical and constant rate of splitting to form new lineages. This is often contrasted with the proportional-to-distinguishable arrangements (PDA) model (Rosen, 1978, called the equal-probabilities model or EP model in Rogers, 1993, 1994, 1996), under which every different labelled tree is equally probable. Previous authors have suggested that no biological model of the evolutionary process leads to this distribution (Mooers and Heard, 1997). Recently, however, Steel and McKenzie (2001) have shown analytically that this distribution results from a model in which, unless a lineage has undergone a speciation within a certain time period, it will never speciate, providing the time period specified is sufficiently small. Notwithstanding this biological model, the PDA model is useful because it represents the case in which a tree-building method is simply selecting randomly from the set of possible result trees. A third simple model has been suggested, in which every possible cladogram shape (i.e. every unlabelled tree) is equally probable (the equiprobable-types model of Simberloff et al., 1981). Since no biological process has been proposed that could produce trees following this model (Mooers and Heard, 1997), I do not consider this model further. More complex models can usually be described as

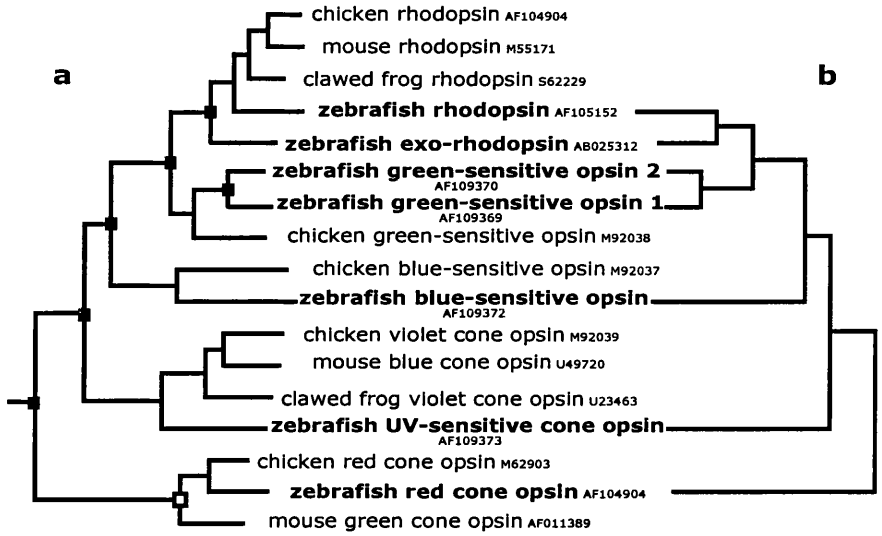

Figure 6.1: A gene family tree: Opsins from four vertebrate species - mouse, chicken, zebrafish and clawed frog. (a) including all 4 taxa. Some nodes represent speciation events, others (marked with a black rectangle) gene duplication events. (b) Including only zebrafish sequences. All the nodes in this tree represent gene duplications, so this sort of tree can be used to study the gene duplication process alone. Identities of gene duplications inferred using reconciled trees (Page and Charleston, 1997a) and the commonly accepted relationship between mouse, chicken, zebrafish and frog. Note that some duplication events, such as that splitting the green cone opsin from the two red cone opsin sequences, are inferred only from differences between species tree and gene tree (Maddison, 1997; Page and Charleston, 1997a), and, as there is no zebrafish ortholog (Mindell and Meyer, 2001) of the green cone opsin, this duplication event is not represented on tree (b).

relaxations of the assumptions of the ERM model, while bias in tree estimation towards randomness will produce deviations towards the PDA model.

The most widely used index of tree imbalance is Colless's (1982) coefficient of imbalance (I_m), largely because it has proved to be the most mathematically tractable (Rogers, 1994). This index takes the sum, over every node in the tree, of the absolute difference in the number of leaves descended from its two descendant nodes. In fact, I_m is usually used normalised to range from 0 (for a completely balanced topology) to 1 (for a completely unbalanced topology) by dividing by $\frac{(n-1)(n-2)}{2}$, where n is the number of leaves on the tree. Colless's index was the first to be studied analytically, resulting in the availability of recursion equations for the expected mean, variance, skewness and probability distribution of this measure under both the ERM and PDA models (Rogers, 1993, 1994). Similar results are also available for several other measures (Rogers, 1996). For its mathematical tractability, and to allow easy comparison with previous data sets, we employ the normalised form of Colless's I_m in this paper.

Previous investigations of tree shape have established that actual phylogenetic trees are significantly more unbalanced than expected under the ERM model (Mooers and Heard, 1997). A number of different explanations for this have been put forward, falling into two categories, with the first set of explanations claiming that this deviation from the null model is an artefact due either to errors in phylogenetic reconstruction or bias in data collection. Previous work has found that poorly supported maximum-parsimony trees tend to be less balanced than well-supported ones Mooers (1995). Random data will produce trees from the PDA distribution. The effect of poor data on phenetic trees is rather different – as data deteriorates, UPGMA trees change little in balance, despite being as prone to error as cladistic trees (Huelsenbeck and Kirkpatrick, 1996). There has been some debate as to whether there are differences in balance of trees from real data produced using phenetic and cladistic methods, but it seems likely that there is no significant difference between the balance of phenograms and cladograms for fairly robust data (Heard and Mooers, 1996). Finally, Mooers (1995) has demonstrated that complete trees (that include all extant members of a taxon) are more balanced than incomplete trees, a result expected if taxon selection is non-random and the selection of included taxa is clumped (Guyer and Slowinski, 1991).

The second category of explanations claim that deviation from the ERM model

reflects the true pattern of speciation, suggesting that the speciation process is more complex than this model allows. This has lead a number of authors to propose more complex, and perhaps more realistic, models of the speciation and extinction process. Heard (1996) and Kirkpatrick and Slatkin (1993) both propose models in which diversification rates evolve through time and found that while this can produce unbalanced trees, extremely large amounts of rate variation are required to produce the degree of imbalance observed in real data. Losos and Adler (1995) proposed a model with a 'refractory period' after speciation where a lineage cannot speciate further, which generally produces trees more balanced than ERM expectations, although Rogers (1996) has pointed out that extremely long refractory periods can produce unbalanced trees. Much work remains to be done on finding macroevolutionary explanations for the imbalance of real phylogenetic trees.

It is worthwhile placing the gene family trees used in this study in the context of previous studies. Gene family trees from complete genomes will be complete in the sense of Mooers (1995) in that they include all the extant sequences of a clade. Gene family trees will, of course, only sample currently extant gene copies. If the trees are evolving under an ERM process, they will match the expectations of the simple ERM model (the ERM-TS model of Harcourt-Brown et al., 2001).

Here, we make an initial attempt to use cladistic balance to make inferences about the process of gene duplication. To do this we grouped protein sequences from the human genome into gene families and constructed phylogenetic trees for these families. We show the imbalance of these trees, comparing different methods of tree construction. To put these values in context, we compare the imbalance of these trees with expectations from both the ERM and PDA models and with the imbalance of species phylogenies collated from the literature by other workers. Differences between the balance of gene family trees and species trees may highlight important differences in the branching processes of gene duplication and speciation. This should be a useful comparison, as, if species trees and gene family trees are constructed similarly, differences between the balance of these two types of tree will be due to differences in the evolutionary process alone.

We also examine one hypothesis about the shape of human gene families. Imbalance of gene family trees has been previously used to test the idea that there have been two episodes of whole-genome duplication during vertebrate evolution (the "2R hypothesis", Holland et al., 1994). In the absence of gene deletion, two

consecutive genome duplications should amplify a single gene into a 4-member gene family, with a perfectly balanced tree topology (Furlong and Holland, 2002; Hughes, 1998, 1999b; Martin, 2001). Martin and Hughes have both found that most 4-member gene families are unbalanced, and hence rejected the 2R hypothesis.

6.2 Materials and methods

6.2.1 Building gene family trees

Protein sequences extracted from the NCBI's annotation of the human genome sequence were grouped into gene families using a strategy based on BLAST searches (Altschul et al., 1990). The *blastclust* program was used to cluster amino-acid gene sequences from the NCBI reference sequence of 20/03/2002. These sequences were then matched with invertebrate outgroups by blast searches against the entire invertebrate section of Genbank. A database of all the sequences was compared with the outgroup database using the *blastp* option of the *blastall* program. Alignments were generated for all families with more than 3 and less than 500 member sequences using ClustalW (Thompson et al., 1994) with default parameters, and phylogenetic trees constructed using the neighbour-joining algorithm (Saitou and Nei, 1987) implemented in the same software package, using uncorrected, gapped distances. A second set of unrooted trees were generated using Tree-Puzzle v5.0 (Schmidt et al., 2002), followed by neighbour-joining using these distances. Some trees were discarded due to difficulties in alignment or tree reconstruction. Trees were constructed separately both with and without the outgroup sequences, and either midpoint rooted or rooted using the outgroups. Colless's I_m was calculated using a purposely written C++ program. All specifically-written software is available from the author on request. This process produced 4 sets of trees – constructed using either uncorrected distances or maximum-likelihood distances, and rooted either using midpoint rooting or with an outgroup sequence.

6.2.2 Simulating genome duplications

To establish the effect that non-independent gene duplication has on tree balance, the effect of the most extreme non-independent event, a whole-genome duplication,

was simulated. A C++ program was used to evolve trees under the ERM model, but with every lineage duplicating simultaneously as the final cladogenesis event. A separate simulation simulated two consecutive genome duplications as the first cladogenesis events in a gene family, followed by subsequent evolution under the ERM model.

6.2.3 Statistical tests of imbalance

In common with previous workers, statistical tests were based on ${}_pI_m$ scores (Heard, 1992; Kirkpatrick and Slatkin, 1993). To calculate these scores, each tree's I_m was compared with the expectation based on 10000 trees of the same number of leaves simulated under the ERM model, and the ${}_pI_m$ score was taken as the number of these 10000 simulated trees with the same or more extreme I_m scores: i.e. the p-value of observing a tree this unbalanced under the ERM model. Such ${}_pI_m$ scores have previously been considered to be independent of tree size (Mooers, 1995), but this is probably due to the limited power of small datasets (and particularly, datasets of small trees) to detect this statistics relationship to tree size (Stam, 2002). To ensure greater homogeneity of variance within tree sizes, ${}_pI_m$ scores used in statistical tests were transformed using the arcsine transformation (Sokal and Rohlf, 1995, p.421).

6.3 Results

Our single-linkage clustering approach divided the protein-coding genes from the human genome into 32,995 gene families, including families of a single gene. The distribution of gene family sizes was roughly consistent with previous work (Lander, 2001; Li et al., 2001). We constructed midpoint-rooted trees for 700 gene families that had more than 3 members, excluding one large family of 3314 sequences that was rejected because of the difficulty of aligning such a large set of sequences. Colless's index cannot be calculated for polytomous trees, so a few trees were excluded from the final datasets because they contained zero-length internal branches, representing polytomies. There are 661 outgroup-rooted trees and 657 midpoint-rooted trees built using uncorrected distances, and 680 outgroup-rooted and 672 midpoint-rooted trees built using maximum-likelihood distances.

Figure 6.2 shows the imbalance of our trees in comparison with expected values under the ERM and PDA models. Clearly, gene family trees are more unbalanced than expected under the ERM model but substantially more balanced than expected under the PDA model. This can be confirmed for the ERM model because the individual ${}_pI_m$ scores can be combined using Fisher's method to yield an overall p-value that the trees have been drawn from an ERM distribution (Sokal and Rohlf, 1995, pp.794-797). This test significantly rejects this possibility for my best-quality data, but cannot for the other three sets of trees (for ML distances, outgroup rooted, $\chi^2 = 1831.34$, df = 1322, $P < 0.0001$; for ML distances, midpoint rooted, $\chi^2 = 1357.38$, df = 1314, P = 0.1976; for uncorrected distances, outgroup rooted, $\chi^2 = 1348.69$, df = 1304, P = 0.189900; for uncorrected distances, midpoint rooted, $\chi^2 = 1043.93$, df = 1308, $P \approx 1.0000$).

Because our four sets of trees are tree for the same gene families, we can use paired methods to compare the imbalance of these different trees. Examining the differences in I_m between the outgroup-rooted maximum-likelihood distance trees and the other three sets for each gene family, we see that the distributions of these differences are underdispersed with respect to a normal distribution, but are symmetrical, so we can use the nonparametric Wilcoxon signed-rank test to show that the medians of our three comparisons are all significantly different to zero, showing that the outgroup-rooted maximum-likelihood trees are the least balanced of the four sets (vs. midpoint-rooted, uncorrected distance trees, median = 0.1078, test N = 379, Wilcoxon statistic = 59284; vs. outgroup-rooted, uncorrected distance trees, median = 0.0417, test N = 316, Wilcoxon statistic = 40588; vs. midpoint-rooted, maximum-likelihood distance trees, median = 0.0762, test N = 373, Wilcoxon statistic=53531; for all three comparisons, $p < 0.001$). The direction of this difference is somewhat surprising – we would expect that less sophisticated methods of tree construction, such as using uncorrected distances, would produce trees closer to the PDA model expectations, and so produce less balanced trees. The difference between midpoint rooted trees and outgroup rooted trees is as would be expected if evolutionary rates were increased immediately after a duplication event, as has been suggested by a number of studies (e.g. Lynch and Conery, 2000). All subsequent analyses were carried out using results for what should be the most accurate estimates of the gene family trees – using maximum-likelihood distances and outgroup rooting.

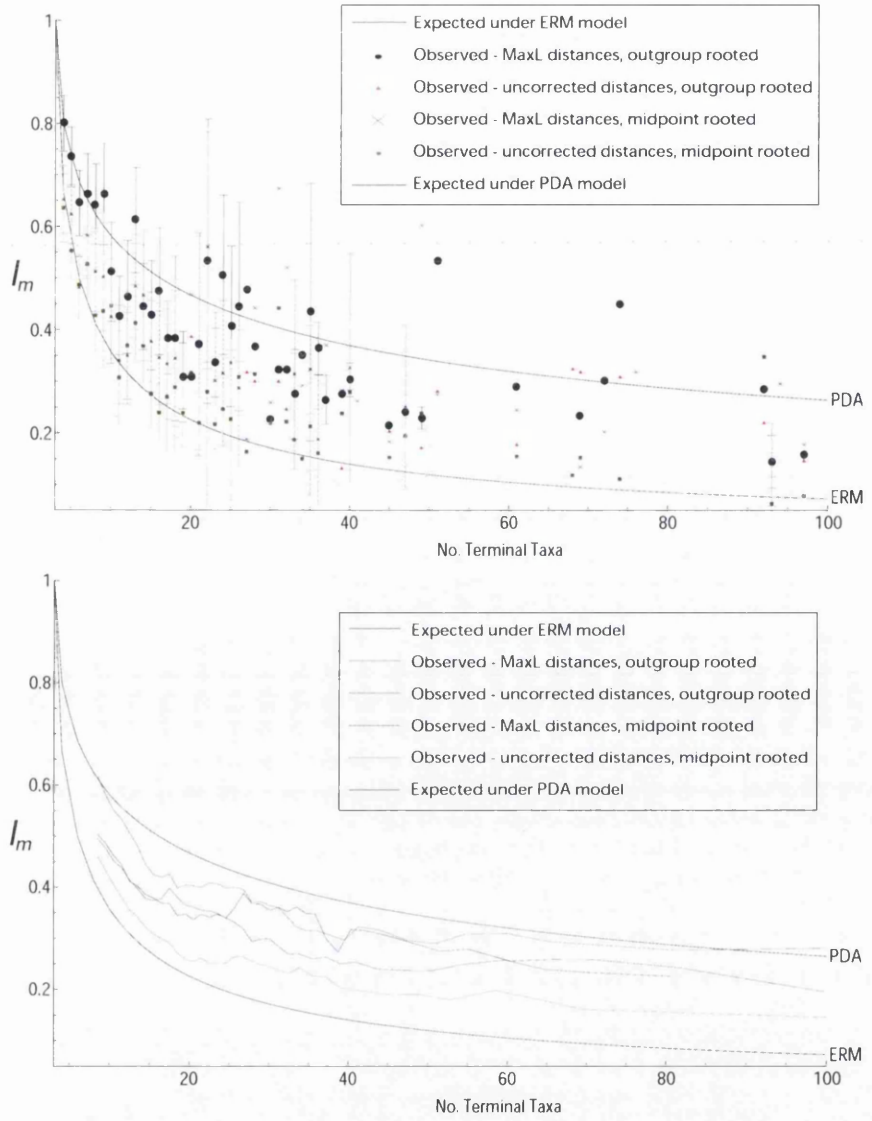

Figure 6.2: Imbalance of human gene family trees against number of leaves, comparing values for the four different sets of gene family trees used here. On the upper figure, points represent the mean I_m values of human gene families for each leaf number, with error bars representing 2 standard errors around these means. On the lower figure, the lines connect 10-term moving averages of I_m values. Smooth lines connect expected mean I_m values under the ERM model (lower line) and PDA model (upper line), found using the method of Rogers (1994).

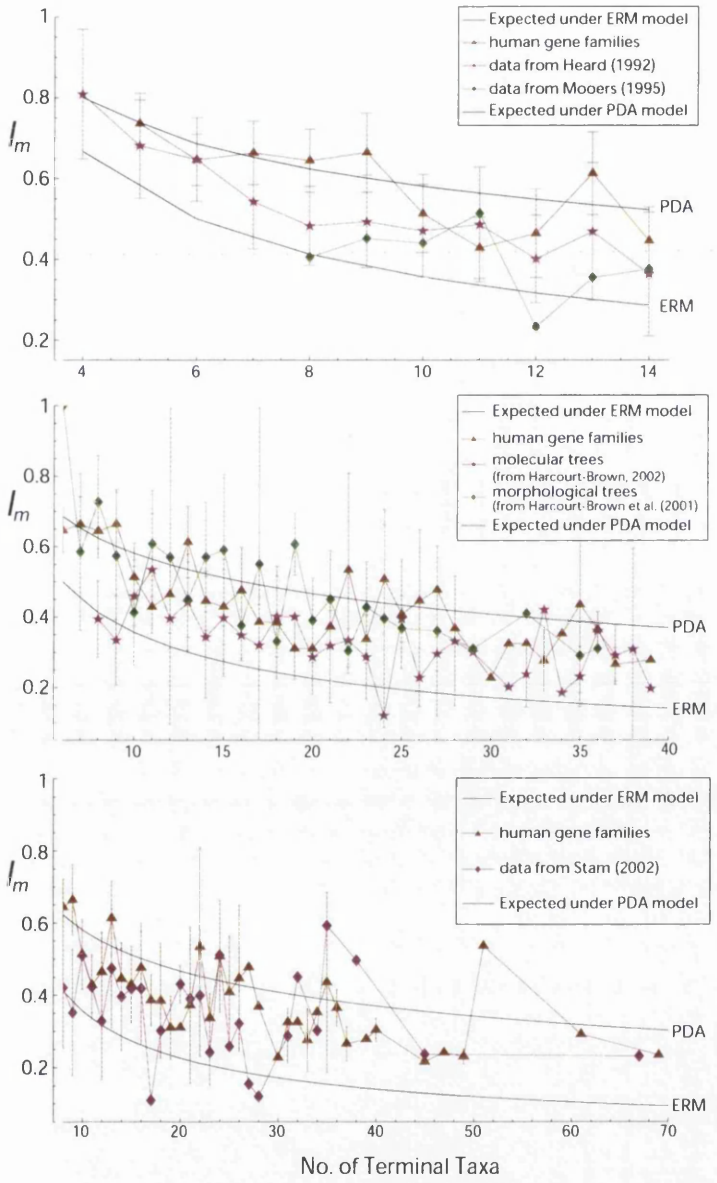

Figure 6.3: Comparison of imbalance, measured by Colless's I_m, between human gene family phylogenies (estimated using maximum likelihood distances and outgroup rooting) with species phylogenies collected from the literature (Species phylogeny data from Harcourt-Brown, 2002; Harcourt-Brown et al., 2001; Heard, 1992; Mooers, 1995; Stam, 2002). Smooth lines connect expected mean I_m values under the ERM model (lower line) and PDA model (upper line), found using the method of Rogers (1994).

For comparing between datasets, previous authors (e.g.. Mooers, 1995) have reported that using Arcsine transformed ${}_pI_m$ scores show no correlation with number of leaves on a tree. The human gene family tree data reported here support the finding of Stam (2002) that this is not the case. I suspect that this is because of the greater statistical power of these data and Stam's dataset, which include much larger trees than Mooers (1995). Mooers' finding that he could not reject an effect of number of leaves appears to imply that this effect was not present in his dataset, but this may be due to a lack of power in his test. Because of this, we have used statistical tests including the number of leaves in our models, but, because the relationship between ${}_pI_m$ and number of leaves is non-linear, we have adopted a conservative method, treating number of leaves as a categorical variable so that ${}_pI_m$ values are only compared for trees with the same leaf numbers.

Using these methods, we find significant differences between our data and that from real trees in each of the three incomplete-tree datasets for which full information was available (for Harcourt-Brown (2002) [100 molecular trees], GLM of arcsine-transformed ${}_pI_m$ scores with number of leaves and dataset as factors: Nleaves (number of taxa) F = 12.44, df = 49/729, P < 0.0001, dataset F = 14.24, df = 1/729, P < 0.0001; Harcourt-Brown et al. (2001) [100 morphological trees], GLM of arcsine-transformed ${}_pI_m$ scores with number of leaves and dataset as factors: Nleaves F = 12.69, df = 48/779, P < 0.0001, dataset F = 0.05, df = 1/779, P = 0.823; for Heard (1992) data [249 trees], GLM of arcsine-transformed ${}_pI_m$ scores with number of leaves and dataset as factors: Nleaves F = 13.18, df = 47/928, P < 0.001, dataset F = 8.38, df = 1/928, P = 0.002). Imbalance measures for all of these sets of trees are shown on figure 6.3. The only dataset that is not significantly rejected by this test is that of morphological trees from Harcourt-Brown et al. (2001). This is probably because half of the trees from this dataset include fossil taxa as leaves – Harcourt-Brown et al. (2001) show that this is likely to make these trees more unbalanced than equivalent trees containing only contemporaneous leaves.

I have obtained data for complete trees from two different compilations – those of Mooers (1995) and Stam (2002). For the Mooers (1995) data, I_m scores for individual trees were not available, so a statistical comparison was not possible. However, it appears that our trees are slightly more unbalanced than the ones compiled by Mooers – he reports that the median ${}_pI_m$ scores or his data, which varies in leaf number from 8 to 14, is 0.429, and that these scores are not significantly correlated

with leaf number. If we consider the set of the most ultrametric gene family trees of between 8 and 14 leaves, we have 152 trees and the median ${}_pI_m$ score is 0.556, slightly higher than the figure Mooers reports. Stam (2002) collected a larger set of 69 complete species trees, including larger trees than Mooers (1995), and so potential allowing more powerful statistical comparisons. Using similar statistical tests as above confirms that human gene family trees show significantly different balance to the trees collected by Stam (GLM of arcsine-transformed ${}_pI_m$ scores with number of leaves and dataset as factors: Nleaves F = 12.56, df = 49/748, P < 0.001, dataset F = 13.05, df = 1/748, P < 0.001).

For four-member gene families, 176 out of 220 gene families, or 80%, are unbalanced. However, if we assume that gene families generally evolve under an equal-rate Markov model, with a genome duplication superimposed on this background of lineage birth and death, we would expect many of the 4-taxon gene trees to be unbalanced whether or not a genome duplication occurred. Specifically, two-thirds of such trees under the pure ERM model should be fully unbalanced. Using Fisher's method confirms that these trees are significantly more unbalanced than expected than the ERM expectation (χ^2 = 145.724, df = 442, P < 0.001), reflecting the general trend of the human gene family trees.

6.4 Discussion

We find that our trees are more unbalanced than species trees compiled from the literature (figure 6.3) and than expected under the PDA model, and significantly less balanced than under the ERM model (figure 6.2). These results suggest that the process of gene duplication occurs similarly, but not identically to that of speciation. The difference in balance between gene family trees and species trees invites us to look for differences between trees showing speciation events and trees showing gene duplication events that might explain it.

This difference in balance may be due to different biases acting on these trees than on species trees. Taxon sampling is perhaps the most obvious explanation – our trees are complete in that they sample all the extant members of a gene family from the human genome, so there is no effect on balance from non-random taxon sampling (Mooers, 1995). The difference between my trees and published cladograms (many of which are based on morphological data) could also be due

to some differences between trees from morphological and molecular data, but my trees are more balanced than the molecular trees from Harcourt-Brown (2002), suggesting that this explanation is not sufficient alone. However, it is possible that a combination of these two factors could explain the relative balance of the gene family trees used here without invoking any difference between the processes of gene duplication and speciation.

If we are seeking an evolutionary explanation for the different balances, a number of differences between the processes of gene duplication and speciation might explain the different balance of the trees produced. In principle, any of the models that have been invoked to explain the deviation of observed species trees from Markov (ERM) expectations could be acting on gene duplications, but to a lesser extent than on speciation. For example, if the model of evolving rates suggested by (Heard, 1996) applied to both speciation and duplication rates, but with less variation in the rate of duplication than in the rate of speciation, this would predict the sort of difference observed.

The fact that gene family trees appear to be significantly less balanced than species trees is particularly surprising given that we would expect gene family trees to be more balanced, as gene duplications within a single gene family are not always independent events. Many gene duplications are caused by the copying of a stretch of DNA from one part of the genome to another. This can occur due to a number of different molecular mechanisms (Ohno, 1970, pp. 89-109). Several of these mechanisms may copy fairly large quantities of DNA in a single event – duplication by processes of polysomy (the multiplication of a single chromosome pair) and polyploidy (the multiplication of the entire genome) will copy many or all genes. When multiple members of a gene family are duplicated by a single event of these kinds, this will produce more symmetrical trees than expected under the ERM model (figure 6.4). If these processes are occurring as part of a birth-death process for gene duplications that tends to generate trees less balanced than the ERM they will shift the trees towards greater balance, the opposite trend to the observed pattern of tree balance in our data. This non-independence of gene duplications might explain the lower imbalance of gene family trees even if the birth-death processes of gene duplication-loss and speciation-extinction are otherwise identical.

Our suggestion that non-independent gene duplications could reduce tree imbalance is formally equivalent to the suggestion by Kirkpatrick and Slatkin (1993)

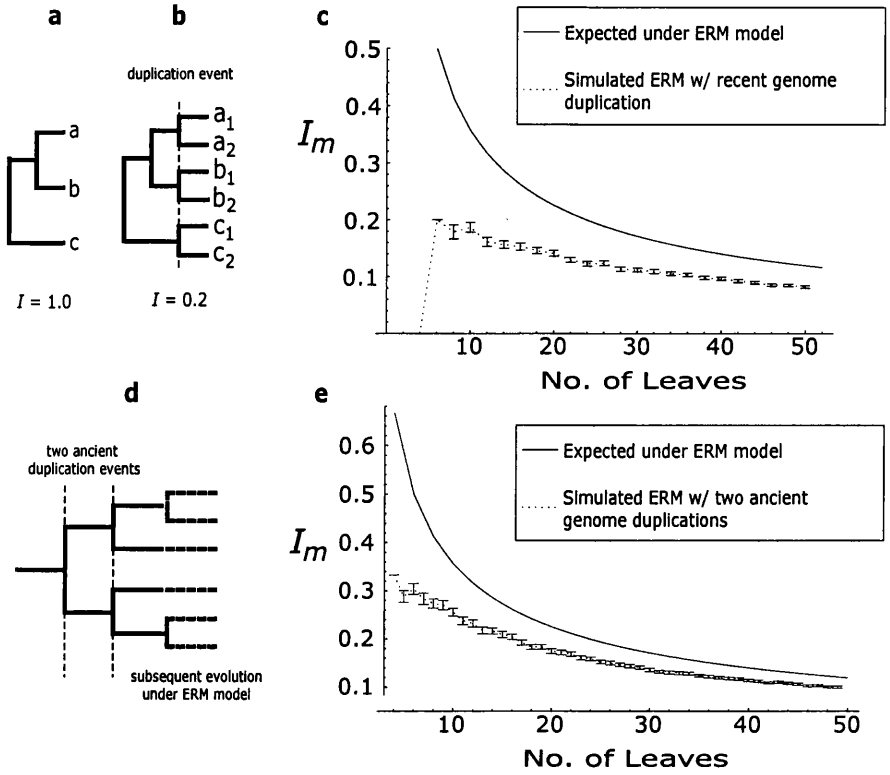

Figure 6.4: Episodes of gene duplications increase tree balance. A gene family phylogeny (a) before and (b) after a genome duplication event. I_m for tree (b) is 0.2. Tree (a) has I_m 1, but evolving to 6 taxa under the ERM model, the expected mean I_m is 0.5 (Rogers, 1994). There is only a one-in-three chance of producing a tree as balanced as (b) under this model. For larger trees, duplication episodes will produce I_m values outside the 95% confidence interval for the ERM model. (c) and (e) Show results of simulations of genome duplications on trees evolving under the ERM model, based on 500 trees each of sizes from 4 to 50 leaves, showing mean I_m and 2 standard errors around the mean. (d) Shows the effect of a single, recent genome duplication – note that this only produces trees with even numbers of leaves. (e) Shows the effect of two consecutive ancient episodes of genome duplication, as shown in (d), and matching the assumptions of the 2R hypothesis. Note that recent duplications leave a larger signal in I_m values, despite I_m giving higher weight to basal branches (Agapow and Purvis, 2003).

that trees more balanced than expected under the ERM model may be produced by "synchronous speciation caused by vicariance events that affect most or all of the species in a clade". However, such events would seem unlikely to be sufficiently common or wide-ranging to have a very major effect on the shape of resulting phylogenies – sampled phylogenies rarely include many taxa from the same area. Only large-scale (and presumably unusual) geographic events would lead to simultaneous speciation in a number of related lineages. This will depend to some extent on the size of the tree – clearly the 'tree of life' relating all species to one another would show a number of large cladogenesis events relating to large bio-geographic changes.

Gene duplications are rather different. Gene duplication events can potentially duplicate multiple members of a gene family, or even all members in a polyploidy or genome duplication event. Regional duplications are particularly likely to duplicate multiple members of a gene family where families have been produced by tandem duplication, producing many lineages tightly linked in the genome (Li, 1997). The fact that gene family trees show significantly greater imbalance than species trees of the same size suggests that regional duplication has not played a sufficiently large role in gene family evolution to leave any signal on the balance characteristics of gene family trees, or that the rate of gene shuffling after tandem duplication is high enough to move duplicated genes apart before regional duplication occurs (McLysaght et al., 2000; Seoighe and Wolfe, 1998).

Another peculiarity of gene duplication will have the opposite effect on tree balance, tending to produce less balanced trees. Tandem gene duplications, where a piece of DNA is duplicated adjacent to the original copy, will produce arrays of related genes, such as observed in the developmental Hox clusters of metazoans (Garcia-Fernandez and Holland, 1994). These repeats of similar sequence will themselves tend to increase the rate at which illegitimate recombination or replication slippage occurs, and so lead to further tandem duplications (Ohno, 1970, pp.62-64). This tendency for the rate of duplication to increase following a duplication will produce imbalanced tree topologies – it is the opposite situation to that modeled by Losos and Adler (1995) and Rogers (1996) (see figure 6.5). In fact, the problem is rather more complex than this, as phylogenetic trees from tandem duplicated loci are highly constrained – only a small proportion of possible tree shapes could actually represent the history of tandem-duplicated genes. Tech-

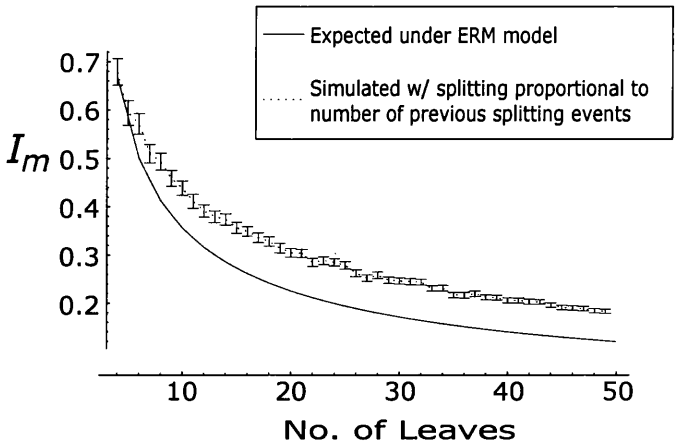

Figure 6.5: If arrays of tandem duplications duplicate at increasing rates, this could produced highly unbalanced trees. Results of a simulation of a branching process where the probability of a particular branch splitting is proportional to the number of splitting events leading to that branch, based on 500 trees each of sizes from 4 to 50 leaves, showing mean I_m and 2 standard errors around the mean.

niques for randomly generating these trees are available (Gascuel et al., 2003), so it should be possible to construct PDA and ERM distributions for tandem duplication trees.

The balance of four-taxon trees from our data seems to weakly support previous work by Hughes (1999b) and Martin (2001) suggesting that such trees do not show the fully balanced picture that would be expected from two consecutive genome duplications (Skrabanek and Wolfe, 1998). In the wider context of tree balance for human gene families this is not surprising, and we would expect that most four-taxon trees would be unbalanced. As discussed above, any genome duplications would have shaped the phylogenies for gene families of every size – for larger trees, they will be a product of genome duplications superimposed over the regular birth-death pattern, so there seems to be little reason to focus exclusively on these small and so relatively uninformative families.

6.4.1 Inferring evolutionary processes from tree imbalance

Studies of tree imbalance have moved on little from the situation summarised by Mooers and Heard (1997), who argue that we have too little understanding of the extent to which methodological biases shape phylogenies, making it hard to establish "how much of the deviation of estimated trees from the Markov model needs macro-evolutionary explanation". We believe this is particularly problematic because of the difficulty in acquiring large sets of trees without time-consuming literature searches. Complete genomes are one place where many complete phylogenies are easily collected, and methods for studying tree shape could usefully be developed in the context of studying gene duplications using this rich dataset, before being applied to the more uneven data relating to the speciation process.

We have suggested some additional complications for models of the duplication process – the non-independence of gene duplications and the process of tandem duplication – which do not apply in the analogous process of speciation. More sophisticated models of regional gene duplications would show the different effects that the size, number, and timing of such events could have on the balance of phylogenetic trees – here, I have modeled only two very simple situations (figure 6.4). The balance of gene family trees will reflect the relative rates of large-scale duplication and tandem duplication, so tree balance could provide a method to study these processes, although this will require a better understanding of the basic birth-death processes of speciation-extinction and gene duplication-loss. It may also require better statistics of tree shape – as Agapow and Purvis (2003) point out, I_m is a powerful statistic for trees from a variety of models, but different measures have different properties, and may be more useful in different situations.

One major drawback in using phylogenies for genes from a single genome is the lack of any external calibration for the rate of molecular evolution within each gene family. The rate of the molecular clock is very variable between different gene families (Li, 1997, p.191), so it is unsound to assume that similar amounts of sequence divergence on different gene family trees represent similar lengths of evolutionary time. Branch lengths on different trees are then incomparable, making it difficult to use measures of tree shape beyond cladistic measures for these kinds of data. Only human sequences were included in this study, so there is no external calibration available to relate branch lengths between different gene trees.

One approach that might allow us to use absolute branch length information is to include sequences from additional species, providing both speciation and duplication nodes. It is possible to distinguish between nodes representing speciation and duplication events on such a phylogeny (Page and Charleston, 1997a). By constraining speciation nodes (which all represent the same speciation event) to occur at the same time within all gene family phylogenies (for example, to be consistent with the fossil record), it should be possible to directly compare absolute times of gene duplications across many gene families. We will address this in future work.

6.5 Conclusion

Gene family trees are significantly less balanced than would be expected under the equal-rate Markov (ERM) model and are even more unbalanced than published species trees. The different balances of gene family trees and species cladograms suggests some difference between the processes of gene duplications and speciation. This could be due to some some quantitative difference in how rates of speciation and duplication evolve. This difference is surprising, given the non-independence of gene duplications, suggesting that relatively few gene duplications have occurred as segmental duplications affecting multiple loci.

Supplementary Information

The rooted trees used to generate this data are available from `http://kimura.zoology.gla.ac.uk/human_genetrees`. Also available from this site is a text file listing the number of taxa and Colless's index of imbalance for each family. A Mathematica notebook for calculating expected values of this index under the ERM and PDA models is also available from this site or from the author, along with tables of expected values for Colless's Index under these two models for trees of between 3 and 500 leaves. A C++ program to calculate ${}_pI_m$ scores is available from the author.

Chapter 7

Interpreting the Pattern of Vertebrate Gene Duplications

Abstract

A number of recent papers have looked at the pattern of gene duplications during the course of vertebrate genome evolution, focusing on both the pattern in space within the genome (McLysaght et al., 2002) and the pattern over evolutionary time (Gu et al., 2002). Here we re-examine Gu et al's data on the dates of gene duplications during vertebrate evolution. We show that similar data, collected by us, seem to confirm the pattern presented by Gu et al., both when analysed similarly and when examined using a complimentary method that does not rely on molecular clock dating techniques. However, we disagree with Gu et al.'s interpretation of their results – mathematical models of the birth-death process show that there has been no recent increase in the rate of gene duplication.

7.1 Introduction

There has been much recent interest in the pattern of gene duplications in vertebrate evolution. This interest stems originally from Susumu Ohno's seminal 1970 book (Ohno, 1970), which introduced the idea that whole-genome duplications had occurred during vertebrate evolution, making extant vertebrates 'degenerate polyploids'. This idea re-emerged more recently as the '2R hypothesis'

that 2 genome duplications occurred during vertebrate evolution. This hypothesis was based largely on evidence that there were four vertebrate Hox clusters for the single cluster in invertebrates (Garcia-Fernandez and Holland, 1994; Holland et al., 1994). It has been difficult to test this proposal (Skrabanek and Wolfe, 1998) because most of the additional gene copies generated by this evolutionary event have been lost in the subsequent process of diploidisation (Wolfe, 2001) or have diverged to form new loci with different functions (Walsh, 1995).

The arrival of genome-scale sequence data in the last few years has prompted a number of attempts to prove or disprove the 2R hypothesis. Gu et al. (2002) use a dataset of 749 gene families from the HOVERGEN database, and molecular clock estimates of the dates of gene duplications on these phylogenies, to show the timing of 1739 gene duplications in the human lineage. In previous work on vertebrate phylogeny (Cotton and Page, 2002), we collected a similar dataset of 118 gene families, which include 947 human-lineage gene duplications. Applying different but related methods to our data, we find a similar pattern of gene duplication (figure 7.1)

There is however, a potential problem with both our analysis and that of Gu et al. (2002). Both of these analyses use branch lengths on molecular phylogenies to infer the timing of duplication events, so that they are dependent on assuming at least an approximate molecular clock. Given the theoretical concerns about the rate constancy of molecular clocks (Ayala, 1999; Rodriguez-Trelles et al., 2002) and the possibility that previous molecular dating analyses (e.g. Heckman et al., 2001; Kumar and Hedges, 1998) may have substantially over-estimated the dates of evolutionary events (Morris, 1999), this is cause for some concern, as has been discussed by Friedman and Hughes (2003).

Fortunately, other information about the dates of gene duplication events is available, as gene duplications are constrained by speciation nodes above and below them (figure 7.2). There are more reliable dates for these speciation nodes than is possible for duplications, as much data can be used to reconstruct them. A large number of genes can be used simultaneously to estimate the date of a speciation event (Heckman et al., 2001; Kumar and Hedges, 1998), while a gene duplication might only affect one or a few genes, and so only these genes are available for date estimates (Li, 1997, p.289). The dates of speciation events can also be estimated using fossil data (e.g. Tavarè et al., 2002). Kumar and Hedges (1998) represents

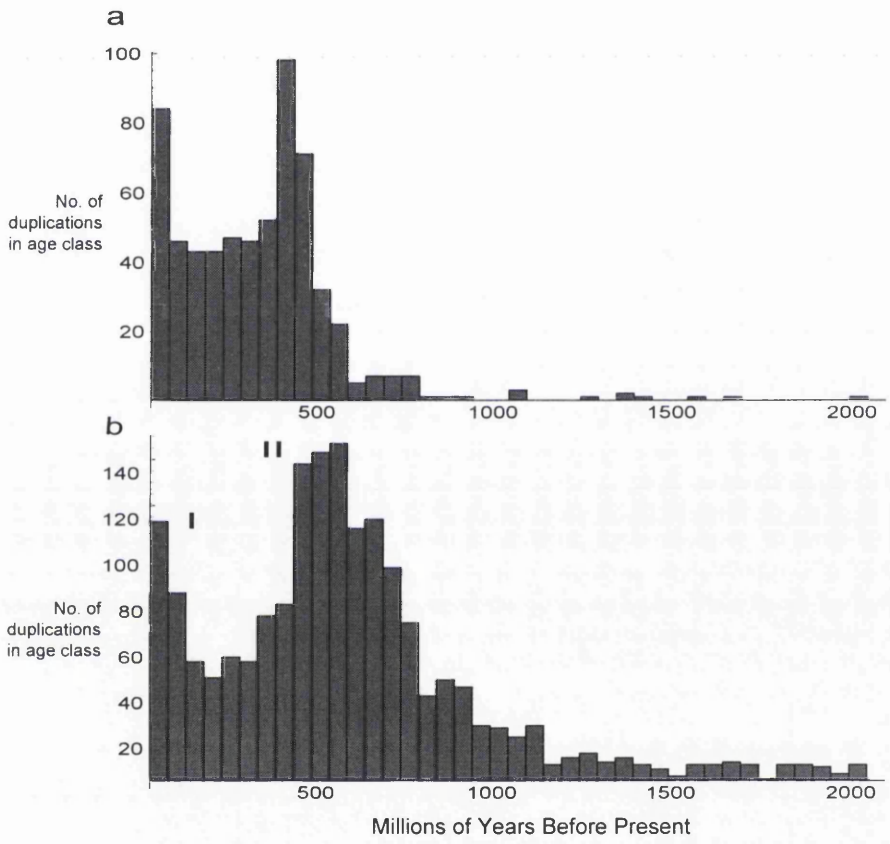

Figure 7.1: Comparison of the results of (a) our data and (b) data from Gu et al. (2002). Figures are histograms showing the numbers of human-lineage gene duplications dated to occur at different times in vertebrate evolution in the two datasets. Roman numerals on figure (b) locate the two episodes of gene duplication identified by Gu et al..

a very large dataset for inferring the dates of divergences within the vertebrates. There is an additional complication, as the location of gene duplications is not exactly determined and a single gene duplication may be part of a larger gene duplication event. This means that duplications on gene family trees may be clustered together to infer larger gene duplications (Page and Cotton, 2002). The distribution of these episodes may not reflect the distribution of individual duplications.

A second concern is with Gu et al.'s interpretation of their results. As shown in figure 7.2, Gu et al. identify two episodes of an increased rate of gene duplication – one putative genome duplication occurring around 500 million years ago, and a second, recent, increase in the rate of duplication, which Gu et al. interpret as representing 'a recent gene family expansion by tandem or segmental duplications', as previously suggested by Eichler (2001).

We disagree with this conclusion. There has been considerable interest in using phylogenies to study the rate of speciation and extinction – processes which are exact analogues of gene duplication and loss. The mathematical models produced to study the processes of speciation and extinction as birth-death processes (Nee et al., 1992) are equally applicable to studying gene duplication and loss, and the results of these models suggests a rather different interpretation of Gu et al.'s results. Birth-death models show a particular characteristic distribution on a graph showing the number of extant lineages against time (known as a lineage-through-time plot – Nee et al., 1992). Where extinction is zero and gene duplication rates per gene copy are constant, these plots show an exponential curve, as the number of lineages present increases exponentially through time and all the lineage persist to the present day. If extinction rates are non-zero, these curves show a characteristic 'hollowed-out exponential' shape, increasing rapidly towards the present, as fewer older lineages persist to the present day and so are observable on phylogenies of recent lineages (Harvey et al., 1994). This appears similar to the pattern shown by the recent episode of gene duplications claimed by Gu et al. (2002).

Birth-death models also allow us to estimate duplication rates per lineage, which allows comparison with previous estimates from other methods (Lynch and Conery, 2000), and also allows us to estimate the rate of gene loss per lineage (Nee et al., 1994) among other parameters of evolutionary interest (Pybus et al., 2003). To test the reality of the proposed episode of gene duplication in recent human evolution, we fit a birth-death model of Kubo and Iwasa (1995) to the data of Gu

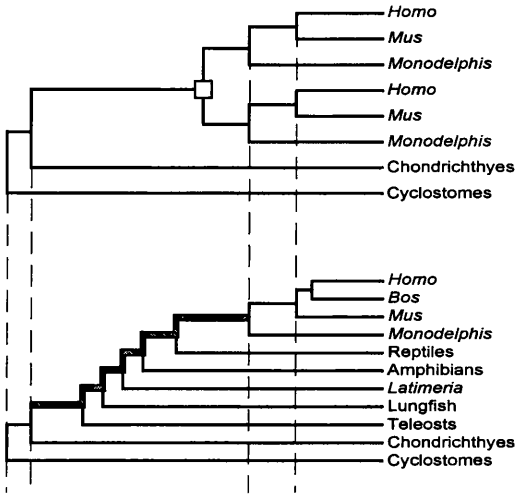

Figure 7.2: Duplications are constrained by neighbouring speciation nodes. The duplication shown here (open rectangle) occurred before the divergence of *Monodelphis* and the placental mammals *Mus* and *Homo*, but after the divergence of the Chondrichthyes and the teleosts. This duplication could thus have occurred anywhere along the highlighted branch of the species tree.

et al. (2002). The same model also lets us make better estimates of rates of gene duplication, and provides the first estimate of the rate of gene loss.

7.2 Methods

7.2.1 Reconstructing gene duplications

Gene families sampling at least five vertebrate classes were selected from the HOVERGEN database (Duret et al., 1994), and gene sequences extracted for a set of 24 genera representing the diversity of extant vertebrates, as described previously (Cotton and Page, 2002, chapter 4). This wide taxonomic coverage increases the chance of sampling ancient gene duplications from early in vertebrate evolution. Outgroups for each gene family were found using sequence similarity searches against a number of sequence databases to identify related genes – either invertebrate orthologues or vertebrate paralogues. Due to the size of the dataset, amino acid sequences were aligned in ClustalW (Thompson et al., 1994) using default

parameters. Alignments were also examined by eye to ensure they were reasonably sensible, and so small sequence fragments that might reduce alignment quality and be difficult to place phylogenetically were removed. Several gene families were excluded at this stage, and some large gene families split into subsets. A maximum-likelihood estimate of distances between each taxa was found using TREE-PUZZLE v.5.0 (Schmidt et al., 2002), using the model selected by the program, with amino-acid frequencies estimated from the data and using an 8-category approximation to a gamma distribution to model rate heterogeneity between sites. These distances were then used to produce a neighbour-joining tree in PAUP 4b10. This produced phylogenies for 118 gene families.

Ultrametric trees were produced from these phylogenies by using the non-parametric rate smoothing method (Sanderson, 1997) implemented in the r8s software package, v1.50, with calibration based on a date of 310 mya for the divergence of mammals and reptiles. This calibration date was used by Kumar and Hedges 1998 and is well-supported from fossil data. All nodes representing the relevant speciation event for this calibration point were constrained to the same age, so there were multiple calibration points in a number of gene families. Similarly, some gene families had no nodes mapping to that particular speciation, and so were not available for estimating dates based on that calibration point. These ultrametric trees were then analysed in a special version of the GENETREE program, where dates were output separately for each node on the species tree, and for duplications mapped onto each branch on the species tree, showing the pattern of gene duplication events through evolutionary time. Dates representing gene duplications along the path from the root of the species tree to humans – the evolutionary lineage of humans – were combined to produce the estimated pattern of gene duplication.

7.2.2 Clustering gene duplications

As gene duplication events can occur at a range of different scales, duplications on different gene family trees may be the result of the same multiple gene duplication event. To investigate this, we clustered gene duplications from individual gene families into the minimum number of sets that may represent these larger gene duplication episodes. To do this, a minimum set cover algorithm was used to find the

smallest set of species nodes that could accommodate all the duplications required by the 118 gene families, identifying which gene duplications took place at each node in the species tree. This clustering algorithm is fully described in Page and Cotton (2002). This clustering can be thought of as the distribution of duplication events if we assume that duplications of any size occur with similar frequency. To examine the history of gene duplications without clustering them into large episodes of duplication, we also reconstructed the most probable distribution of duplication events under the assumption that duplications occurred independently. For each branch of the species tree, the most probable number of duplications actually occurring at that location was found by summing the number of duplications reconstructed as occurring on that branch weighted by the uncertainty in the duplication's position. A duplication that is reconstructed as unambiguously occurring at a particular location added 1 to the number of duplications occurring at that location, while a duplication that could have occurred on any of three different branches added $\frac{1}{3}$ to the estimate for each of the three branches. To scale these distributions, the number of gene duplication episodes from the clustering analysis and the ungrouped distribution of duplications were plotted as histograms, with a bar for each branch on the species tree, and with the x-axis scaled to represent the length of each branch using date estimates from Kumar and Hedges (1998).

7.2.3 Birth-death models

The models of the birth-death process used here are those of Kubo and Iwasa (1995). These models are expressed in terms of numbers of lineages rather than numbers of duplications, so data needs to be transformed. This transformation is simple – we start with 749 lineages, and add one lineage for each gene duplication event. A graph of this data is known as a lineage-through-time plot. The birth-death model with constant birth and death relates N_T (the number of extant lineages) and N_t (the number of lineages at time t), where b is the branching rate and c is the extinction rate, by the equation:

$$\frac{N_t}{N_T} = \frac{b-c}{be^{(b-c)(T-t)} - c} \tag{7.1}$$

Fitting this equation to the lineage-through-time plot allows estimates of b and c, under the assumption that b and c remain constant. The extant number of lineages

(N_T) is 2488, as Gu et al.'s data starts with 749 gene families and includes 1739 duplications on these lineages.

7.3 Results and discussion

7.3.1 Another view of the data

This lumped distribution is shown in figure 7.3. These distributions appear very different from the distributions shown in figure 7.1, but they are actually very similar. The deepest divergence shown on figure 7.3 is dated to about 565 mya by Kumar and Hedges (1998), so the increased rate of duplication shown at the left-most edge of both distributions represents the possible '2R' event identified as episode II by Gu et al.. The lower figure represents the unclustered distribution of gene duplications, and we would expect this to most closely match the molecular-clock based distributions, as these show each lineage distribution as an independent event. These data seem to confirm that the pattern of duplications shown by Gu et al. (2002) and mirrored in the distribution from our data is not simply an artefact of the molecular clock assumption, but is a genuine evolutionary phenomenon needing explanation.

7.3.2 Birth-death models

We have converted the data of Gu et al into a lineage-through-time plot (figure 7.4) and have fitted a birth-death model to this data to estimate rates of gene duplication and gene loss. It is clear that Gu et al. (2002)'s data follow the expected 'hollowed-out exponential' shape.

We can see that there has very clearly been a large increase in duplication rates during the period of around 500 million years ago. In contrast, the more recent sharp increase in numbers of duplications observed in both the Gu et al. data and our own follows exactly the pattern that would be expected if the rate of duplication and extinction per lineage had stayed constant throughout the period, and merely reflects the fact that a greater proportion of extant lineages from recent times are still observable (Harvey et al., 1994). Fitting a constant-rate birth-death curve to this data estimates a duplication rate of 0.000961 per million years per lineage, and an extinction rate of 0.000462 per million years per lineage. To our knowledge, this

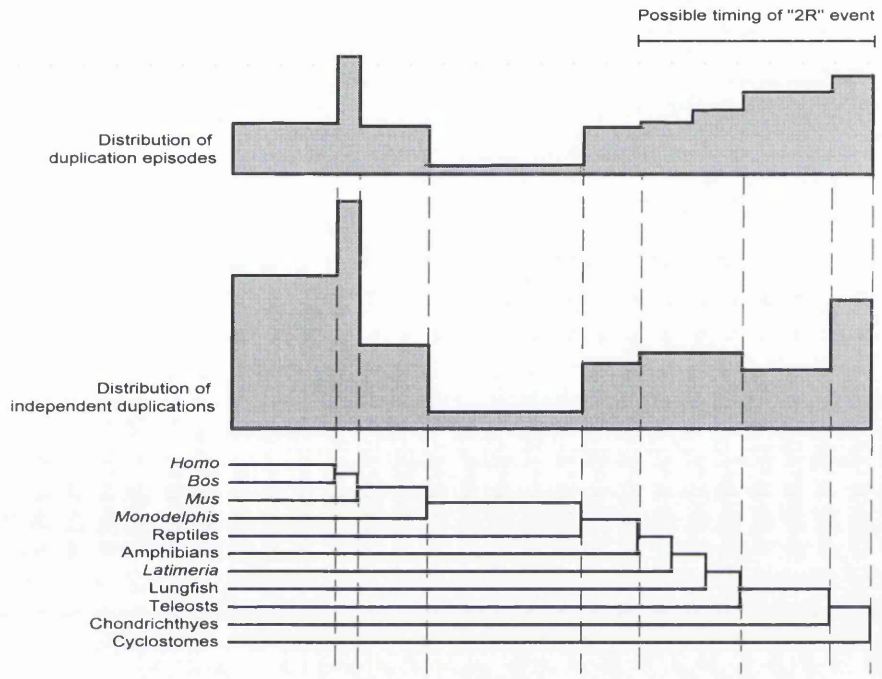

Figure 7.3: A picture of the distribution of gene duplications through human evolution independent of the molecular clock. The locations of human-lineage duplications on our 118 vertebrate gene families were either left unclustered, but with the ambiguity in their positions taken into account (lower figure), or were clustered using the algorithm of Page and Cotton (2002), and the distribution of duplication episodes is shown (top figure). The distributions were scaled so that branch lengths in the species tree reflected dates of cladogenesis events from Kumar and Hedges (1998). Dates were interpolated for events not included in Kumar and Hedges's study.

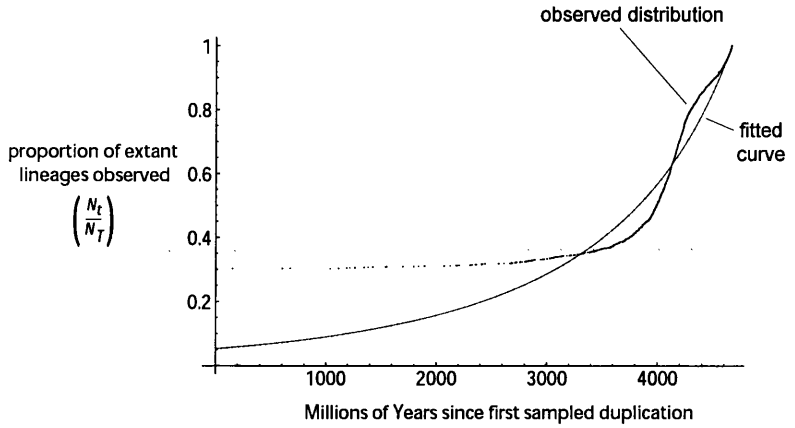

Figure 7.4: The constant rate birth-death model of Kubo and Iwasa (1995) fitted using the least-squares method to the lineages-through-time plot derived from duplication ages from Gu et al. (2002).

is the first estimate of the rate of gene deletion – a parameter difficult to estimate by other approaches.

7.3.3 Discussion

Estimates of divergence dates by this method are comparable with estimates obtained by other methods, as they are per-lineage rates. Our estimate of the rate of gene duplication is rather lower than previous estimates – Lynch and Conery (2000) suggest that *Drosophila* duplicates at a rate of 0.0023 per gene per million years, the yeast *Saccharomyces* at a rate of 0.0083 per gene per million years and *Caenorhabditis* at a rate of 0.0208 per gene per million years, while Lynch and Conery (2001) estimate a duplication rate of 0.0071 per gene per million years for human. The rate of gene loss is important too – for example, it would be an important parameter in determining how common paralogy will be in molecular phylogenies. Using birth-death models to estimate these parameters relies on the accuracy of the molecular clock, which is at best only approximate and may be quite misleading, so its useful to be able to reconstruct the observed pattern of evolution independently of molecular-clock assumptions.

The birth-death models assume that duplications and losses in each lineage are independent, and that the rates of duplication and loss stay constant throughout the

tree, although the effects of varying these rates has been investigated (Kubo and Iwasa, 1995). For example, if purifying selection means that duplicate copies are more likely to go extinct soon after the duplication event that gave rise to them, this will violate the assumptions of the birth-death model and may affect the accuracy of our estimates. It seems clear that there has been variation in the rates of either gene duplication and/or gene loss over the course of vertebrate evolution, most notable in the episode around 500 million years ago identified by Gu et al. (2002). Our estimates can be thought of as long-term average rates of duplication and loss. There are also problems with sampling – as duplicated genes diverge, it will be more and more difficult to detect similarity between them and align the genes properly. This means that any analysis based on gene family phylogenies will be less thorough in sampling older duplications than more recent events. This does not seem likely to have had a major effect on this work, as there are many more recent duplications, so the fitted model (figure 7.4) is fitted largely to this part of the curve, and is less influenced by the sparse, ancient, data.

Genome duplication will be difficult to observe on lineage-through-time plots if there has been a high rate of subsequent gene loss. Kubo and Iwasa (1995) show that a sudden mass speciation (or, in this context, large-scale gene duplication) event will produce a discontinuity in the lineage through time plot as the number of lineages suddenly increases. In fact, the size of this discontinuity will depend upon the extinction rate. At high extinction rates, the discontinuity may be so small as to be difficult to identify against the noisy background of real data. This can be easily shown by some simple simulations – figure 7.5 shows the results of two simulations where constant rates of gene duplication and loss are superimposed on a genome duplication event. In fact, it is even more difficult to detect ancient events of large-scale gene loss – these will be visible only as a slight 'kink' where the gradient of a lineages-through-time plot changes (Kubo and Iwasa, 1995).

Bearing this in mind, it is clear that correct interpretation of the peak in duplication rate observed by Gu et al. (2002) needs good estimates of the rate of gene loss. Gu et al. interpret their data as representing a pattern $mR + C$, meaning m rounds of whole-genome duplication and a background of continuous, smaller scale duplications (C). They conclude that at least one round of genome duplication is necessary to fit the observed pattern, and that the presence of continuous small-scale duplications make it unnecessary to hypothesise more than two gen-

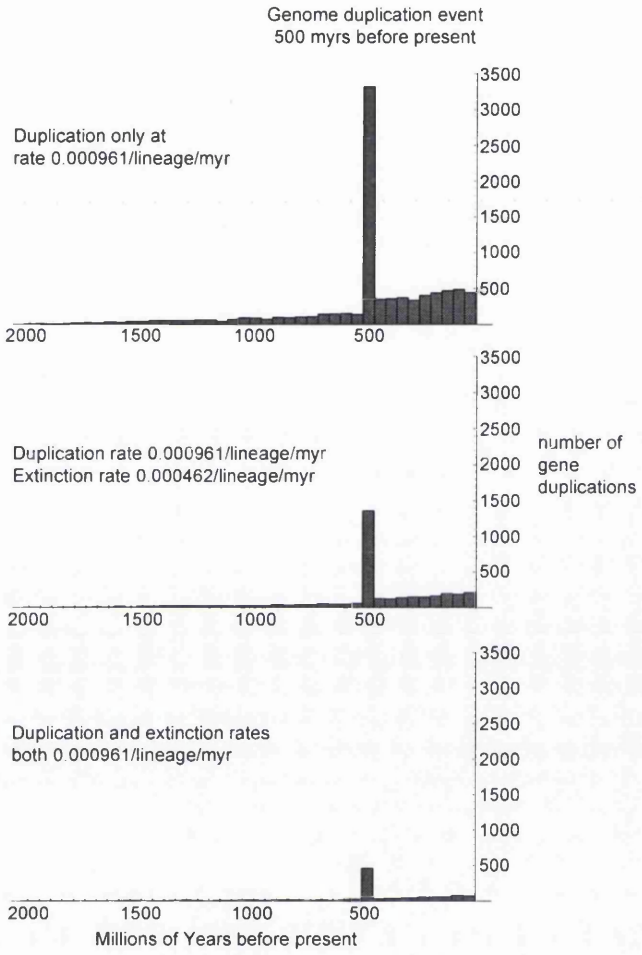

Figure 7.5: The results of simulations showing the effects of gene loss on the signal from an ancient genome duplication event. All three show constant rates of gene duplication and gene loss, with 749 lineages simulated over 2,000 million years. The extinction rate is zero in the top figure, the rate estimated from the birth-death model in the middle figure, and equal to the speciation rate in the top figure. The size of the spike from the genome duplication event 500 million years before present is much less pronounced in the lower figures, as gene loss has erased many of the lineages duplicated in this event. Note that these plots show the number of duplications through time, rather than lineage-through-time plots.

ome duplications, but do not decide between the $2R + C$ hypothesis – a natural extension of the $2R$ idea – and a single round of polyploidisation. We share their caution, but are more sceptical about the suggestion that a probabilistic test of these hypotheses is possible from the kind of data used here. While the scale of ancient gene duplications observed in Gu et al. (2002) is striking, it seems likely that evidence from a number of sources – from the timing of gene duplications, from tree topology, and even from genetic map information – will be needed to finally uncover the history of vertebrate genome evolution.

7.4 Conclusion

Reconstructing the pattern of gene duplications independently of molecular-clock assumptions or reconstructed branch length confirms the pattern of gene duplications through time shown by Gu et al. (2002) and by our data. Using branch-length information, we can use quantitative models of the birth-death process of gene family evolution to estimate rates of gene duplication and gene loss. Gene duplication rates in vertebrate evolution appear to be significantly lower than estimates from other methods (Lynch and Conery, 2000). We also present the first estimate available of the rate of gene loss, suggesting that it is around half the rate of gene duplication. An estimate of the rate of gene loss is crucial in interpreting the pattern of ancient large-scale gene duplication episodes.

Phylogenies used in this work are available from `http://darwin.zoology.gla.ac.uk/~jcotton/vertebrate_data`

Chapter 8

Future Directions

The work in this thesis has shown that reconciled trees can be a powerful tool in studying gene duplication and loss, and for inferring species phylogenies in the presence of these processes. Further progress, however, will probably depend upon new methods and new data for investigating these processes. Here, I briefly introduce several promising avenues for future research.

8.1 Inferring species phylogenies

An ongoing theme of this thesis has been incongruence between different estimates of a species phylogeny. Most of the discussion has focused on how reconciled trees can help understand one such sources of incongruence – paralogy introduced by gene duplication and subsequent loss. This is in marked contrast to usual ways of investigating incongruence, which focus on estimation error as the source of this incongruence. Of course, both processes contribute to incorrect estimates of trees, and I have explored methods for incorporating a measure of estimation error into a reconciled tree framework, using bootstrap profiles of trees for each gene family. This method is somewhat unsatisfactory, not least because of the overly conservative nature of the bootstrap profile (Efron et al., 1996; Hillis and Bull, 1993; Zharkikh and Li, 1992a,b).

One obvious solution to this problem is to combine both gene tree inference and inference of duplications and losses into a single statistical framework. A probabilistic model of the processes of gene duplication and gene loss has been

developed (Lindsey Dubb, pers. comm.) which allows estimates of duplication and loss rates given a gene tree and species tree, and enables us to perform statistical tests about rates of duplication and loss in a likelihood setting (Huelsenbeck and Crandall, 1997). The likelihood of Dubb's model depends upon the rates of duplication and loss (d and l), and involves summation across all possible histories (H) of duplication and loss that could produce the observed gene tree (G):

$$L = \sum_{H} Prob(H|d,l)Prob(G|H) \tag{8.1}$$

This gives a likelihood function of the form $p(gene\ tree|duplication, loss)$. In fact, the observed data is sequence data, rather than the gene tree, so to incorporate tree inference into the model, we can use a standard substitution model to give the probability of the data given the gene tree, and sum across possible gene trees, using MCMC, to give a likelihood function of the form $p(data|duplication, loss)$, which allows inference of duplication and loss rates from the sequence alignment:

$$L = \sum_{H} \sum_{G} Prob(H|d,l)Prob(G|H)Prob(data|G) \tag{8.2}$$

Finally, we can incorporate inference of a species tree into the same framework – the species tree is assumed to be known in Dubb's model, and is included in calculating the probability of particular duplication and loss histories. If we include the species tree as a parameter, S, we can potentially estimate duplication and loss rates by summing across all species trees, or estimate a species tree by summing across duplication and loss rates. These calculations will be extremely computationally intensive. In particular, the sum over all duplication-and-loss histories involves summation over the number of "hidden" or "doomed" lineages at each internal node in the tree – lineages that do not have any extant descendants, and is, in principle, a sum to infinity (figure 8.1).

There is an important distinction between paralogy in general and "hidden paralogy" (Martin and Burg, 2002). Paralogy occurs when a sample of gene sequences includes genes related by gene duplication rather than speciation. Hidden paralogy is a special case of paralogy in which the only gene copies extant for a set of species are paralogous. Paralogy can sometimes be detected by examining the molecular structure of a locus (Sanderson and Shaffer, 2002; Small and Wendel,

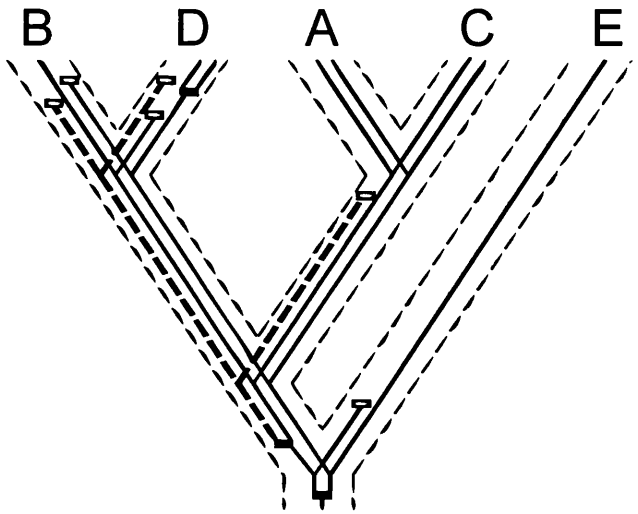

Figure 8.1: Calculating the likelihood of a gene tree given the rates of duplication and loss requires summing over the number of "doomed" lineages at each internal node – lineages that leave no extant descendants. This example shows a gene family evolving in species tree A-E. One unobserved gene duplication leads to a ghost lineage at nodes (B,D,A,C) and (B,D). Taken from J. Felsenstein – lecture for Genetics/MBT 541.

2000), or by sequencing multiple copies of a gene. Furthermore, as sequence data accumulates, paralogy should become easier to detect, at least for well-known species. Hidden paralogy is more worrying and fundamental. One basic question for molecular systematists is "how likely is my data to be affected by hidden paralogy". Answering this question will depend upon knowledge of rates of gene duplication and gene loss in a range of taxa.

The distinction between paralogy and hidden paralogy underlines the potential importance of sampling. If only a limited number of loci have been sequenced from an organism, it is impossible to tell if the absence of a particular gene copy is due to gene deletion or due to that locus simply not having been sequenced. The same issue, of course, is even more vital in the context of supertrees – the entire motivation behind these methods is to combine trees to give a single estimate of phylogeny for species that have no sampling of phylogenetic markers in common. I have advocated using duplication-only measures to avoid the problem of grouping taxa together by sampling alone in reconciled tree methods. Similarly, the probabilistic model described above assumes that the gene family is fully sampled. It seems likely that some kind of model of sampling effort will be be needed if probabilistic models of gene family evolution are to be used beyond model organisms.

8.2 Understanding gene duplication

It is striking that, despite almost 20 years of intense research interest, basic questions about evolution by gene duplication remain unanswered. The debate over the 2R hypothesis still rages. Recent evidence suggests that at least one episode of polyploidisation occurred (Gu et al., 2002; McLysaght et al., 2002), and most agree that an episode of unusually high duplication rate took place (or at least an episode of high maintenance of duplicated copies – Friedman and Hughes, 2003). Despite this, it seems unlikely that debate on the issue is waning.

Part of the reason for this difficulty is that even more basic questions remain unanswered – we can only begin to frame answers to questions like : What is the rate of gene duplication? What is the rate of gene loss? How do these rates vary between taxa, and over time? How large are duplicated segments? How common is polyploidy in organisms other than plants? How rapidly are genes moved around the genome? Are different sorts of genes maintained more frequently than others?

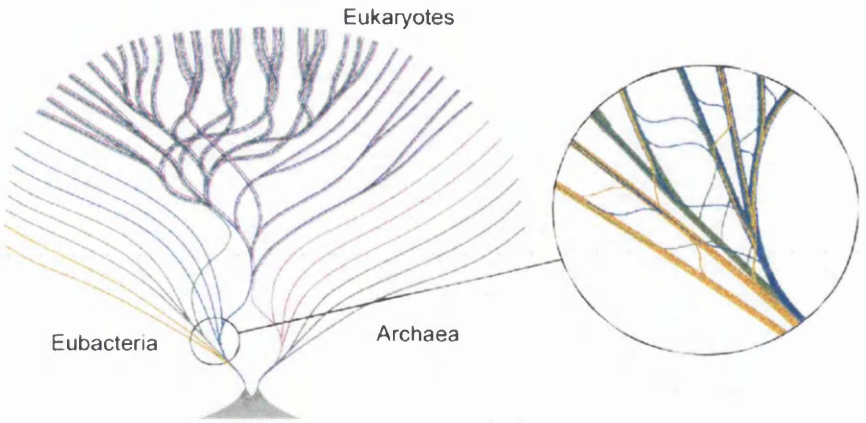

Figure 8.2: Frequent LGT makes bacterial phylogeny a network rather than a tree, while eukaryotes are an evolutionary mosaic of nucleus and organelles. From Martin (1999b).

The methods outlined in the six main chapters of this thesis should help answer some of these questions. Existing techniques need to be applied to existing data beyond the vertebrate examples generally examined. Whole-genome data is vital in separating the roles of gene loss and sampling failure. The great amount of genomic data available for prokaryotes, in particular, might be invaluable in estimating rates and patterns of gene duplication and gene loss, but this will require methods to take lateral gene transfer into account. No doubt, however, many processes will occur differently between prokaryotes and eukaryotes. Many evolutionary questions can only be answered using comparative methods, and it is likely that many fundamental questions about genome evolution in eukaryotes will begin to be answered as more fully-sequenced genomes become available and understood.

8.3 Understanding lateral gene transfer

The reconciled tree methods used in this thesis deal correctly with gene duplication and gene loss, but exclude the possibility of lateral gene transfer (LGT). This is probably reasonable in vertebrates – we would expect that LGT was at least very rare. The same will not be true in other taxa, especially in prokaryotes. Bacterial genomes are increasingly seen as very dynamic, with gene transfers regularly mov-

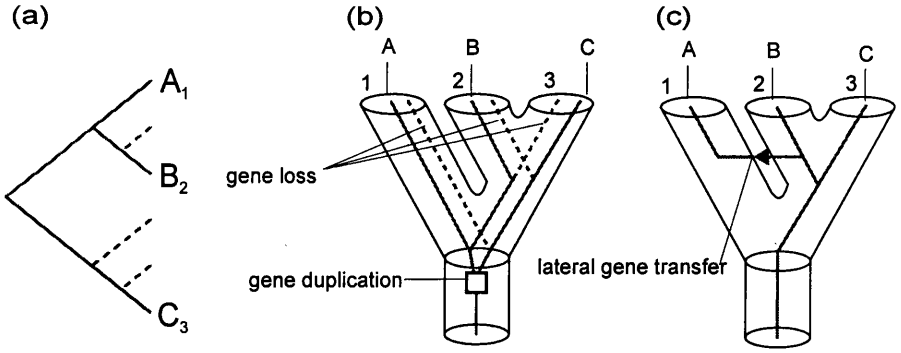

Figure 8.3: LGT (c) and gene duplication followed by loss (b) can have the same phylogenetic effect, introducing incongruence between the gene tree (a) and species tree. Genes 1,2 and 3 are evolving within the species A, B and C respectively.

ing genes between mosaic genomes (Martin, 1999b, – figure 8.2). Most authors agree that LGT "has had an extraordinary affect on bacterial genomes" (Ochman, 2001). There are also likely to be differences in the rate of LGT between different bacterial species (Feil et al., 2001), and over time – several authors have suggested that a large burst of LGT occurred early in prokaryotic evolution, while others prefer a steady-state model of continuing genetic transfers.

The pattern of LGT is thus of great research interest in its own right, but understanding LGT is also crucial in understanding the pattern of gene duplication and gene loss, as the differences between a gene tree and species tree introduced by LGT can be identical to those introduced by gene duplication and gene loss (figure 8.3). Inferring species trees correctly using the methods described here in groups where lateral gene transfer is common will depend upon dealing with host-switching events. One recent attempt at studying LGT in a phylogenetic context developed a novel pattern-based method (Mirkin et al., 2003), but this suffers from the same difficulties of interpretation as previous pattern-based cophylogenetic methods (Page, 1993a; Ronquist and Nylin, 1990). At least two algorithms are available to deal with host-switching in a co-phylogenetic framework (Charleston, 1998; Ronquist, 2003), but both have drawbacks – the *Jungles* algorithm is slow and computationally intensive, while the algorithms implemented in *TreeFitter* do not provide explicit reconstructions of co-phylogenetic history. A proposed Bayesian method is restricted to only a single associate lineage per host, making it

inappropriate for dealing with gene family evolution (Huelsenbeck et al., 2000a). Fortunately, a much faster algorithm has recently been proposed (Hallett and Lagergren, 2001).

Despite these difficulties, incorporating LGT into our models of gene family evolution is a natural progression. Existing methods of detecting LGT are widely seen as unsatisfactory (Eisen, 1998; Sicheritz-Pontén and Andersson, 2001), and correct estimates of rates of gene duplication and gene loss in many taxa will depend upon correctly accounting for LGT.

8.4 Conclusion

In general, future progress in this field, as in the wider field of phylogenetics in general, will depend upon ongoing collaboration between mathematics, computer science and biology. As new data prompts biologists to ask new, more ambitious, questions, they will inevitably need new tools to investigate answers. It is the close interface between these disciplines that has made cophylogenetics and studies of gene duplication and loss interesting and dynamic fields, and exciting ones for the future.

Bibliography

Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P., and Inoko, H. (2002). Evidence of *en bloc* duplication in vertebrate genomes. *Nature Genetics*, 31:100–105.

Agapow, P.-M. and Purvis, A. (2003). Power of eight tree shape statistics to detect nonrandom diversification: A comparison by simulation of two models of cladogenesis. *Systematic Biology*, 51:866–872.

Ahn, S. and Tanksley, S. D. (1993). Comparative linkage maps of the rice and maize genomes. *Proceedings of the National Academy of Sciences, USA*, 90:7980–7984.

Aho, A. V., Sagiv, Y., Szymanski, T. G., and Ullman, J. (1981). Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal of Computing*, 10:405–421.

Allendorf, F. W. and Thorgaard, G. (1984). Tetraploidy and the evolution of salmonid fishes. In Turner, B., editor, *Evolutionary genetics of fishes*, pages 1–46. Plenum, New York.

Archibald, J. M. and Roger, A. J. (2002). Gene duplication and gene conversion shape the evolution of archaeal chaperonins. *Journal of Molecular Biology*, 316:1041–1050.

Ayala, F. J. (1999). Molecular clock mirages. *BioEssays*, 21:71–75.

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W., and Eichler, E. E. (2002). Recent segmental duplications in the human genome. *Science*, 297:1003–1007.

Bailey, W. J., Kim, J., Wagner, G. P., and Ruddle, F. H. (1997). Phylogenetic reconstruction of vertebrate hox cluster duplications. *Molecular Biology and Evolution*, 14:843–853.

Barrett, M., Donoghue, M. J., and Sober, E. (1991). Against consensus. *Systematic Zoology*, 40:486–493.

Batson, C. J. and Ohta, T. (1992). Simulation study of a multigene family, with special reference to the evolution of compensatory mutations. *Genetics*, 13:247–252.

Baum, B. R. (1992). Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon*, 41:3–10.

Becerra, J. X. and Venable, D. L. (1999). Macroevolution of insect-plant associations: The relevance of host biogeography to host affiliation. *Proceedings of the National Academy of Science, USA*, 96:12626–12631.

Bennett, M. D. and Leitch, I. J. (1995). Nuclear DNA amounts in angiosperms. *Annals of Botany*, 76:113–176.

Bininda-Emonds, O. and Bryant, H. N. (1998). Properties of matrix representation with parsimony analyses. *Systematic Biology*, 47:497–508.

Bininda-Emonds, O. R. P. and Sanderson, M. J. (2001). Assesment of the accuracy of matrix representation with parsimony analysis supertree construction. *Systematic Biology*, 50:565–579.

Bishop, M. J. and Friday, A. E. (1988). Estimating the interrelationships of tetrapod groups on the basis of molecular sequence data. In Benton, M. J., editor, *The phylogeny and classification of the tetrapods*, volume 1. Clarendon Press, Oxford.

Brenner, S. E., Hubbard, T., Murzin, A., and Clothia, C. (1995). Gene duplications in *H. influenzae*. *Nature*, 378:140–143.

Britten, R. J. (2002). Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proceedings of the National Academy of Sciences, USA*, 99:13633–13635.

Brooks, D. R. (1981). Hennig's parasitological method: a proposed solution. *Systematic Zoology*, 30:229–249.

Brooks, D. R. (1990). Parsimony analysis in historical biogeography and coevolution: methodological and theoretical update. *Systematic Zoology*, 39:14–30.

Broughton, R. E., Naylor, G. J. P., and Dowling, T. E. (1998). Conflicting phylogenetic patterns caused by molecular mechanisms in mitochondrial DNA sequences. *Systematic Biology*, 47:696–701.

Brower, A. V. Z., DeSalle, R., and Vogler, A. (1996). Gene trees, species trees and systematics: A cladistic perspective. *Annual Review of Ecology and Systematics*, 27:423–50.

Brown, E. K. and Day, W. H. E. (1984). A computationally efficient approximation to the nearest neighbour interchange metric. *Journal of Classification*, 1:93–124.

Brown, J. R. (1996). Preparing for the flood: evolutionary biology in the age of genomics. *Trends in Ecology and Evolution*, 11:510–513.

Bull, J. J., Huelsenbeck, J. P., Cunningham, C. W., Swofford, D. L., and Waddell, P. J. (1993). Partitioning and combining data in phylogenetic analysis. *Systematic Biology*, 42:384–497.

Charleston, M. A. (1998). Jungles: A new solution to the host/parasite phylogeny reconciliation problem. *Mathematical Biosciences*, 149:191–223.

Charleston, M. A. and Perkins, S. L. (2003). Lizards, malaria, and jungles in the Caribbean. In *Tangled Trees: phylogeny, cospeciation and coevolution*, pages 65–92. University of Chicago Press.

Charleston, M. A. and Robertson, D. L. (2002). Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. *Systematic Biology*, 51:528–535.

Chen, D., Diao, L., Eulenstein, O., Fernández-Baca, D., and Sanderson, M. J. (2003). Flipping: a supertree construction method. In *DIMACS Series in Discrete Mathematics and Theoretical Computer Sciences, In Press*. DIMACS-AMS.

Chen, F.-C., Vallender, E. J., Wang, H., Tzeng, C.-S., and Li, W.-H. (2001). Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. *Journal of Heredity*, 92:481–489.

Chor, B., Hendy, M. D., Holland, B. R., and Penny, D. (2000). Multiple maxima of likelihood in phylogenetic trees: An analytic approach. *Molecular Biology and Evolution*, 17:1529–1541.

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature*, 357:543–544.

Cognato, A. I. and Vogler, A. P. (2001). Exploring data interaction and nucleotide alignment in a multiple gene analysis of *Ips* (Coleoptera: Scolytinae). *Systematic Biology*, 50:758–781.

Colless, D. H. (1982). Review of phylogenetics: the theory and practice of phylogenetic systematics. *Systematic Zoology*, 31:100–104.

Cotton, J. A. and Page, R. D. M. (2002). Going nuclear: vertebrate phylogeny and gene family evolution reconciled. *Proceedings of the Royal Society of London Series B*, 269:1555–1561.

Cowlishaw, G. and Clutton-Brock, T. (2001). Primates. In MacDonald, D., editor, *The New Encyclopedia of Mammals*, pages 290–301. Oxford University Press.

Craw, R. (1992). Margins of cladistics: Identity, difference and place in the emergence of phylogenetic systematics 1864-1975. In Griffiths, P., editor, *Trees of Life*. Kluwer Academic Publishers, Dordrecht.

Curole, A. P. and Kocher, T. D. (1999). Mitogenomics: digging deeper with complete mitochondrial genomes. *Trends in Ecology and Evolution*, 14:394–398.

DasGupta, B., He, X., Jiang, T., Li, M., Tromp, J., and Zhang, L. (1997). On distances between phylogenetic trees. *ACM-SIAM Symposium on Discrete Algorithms, New Orleans*, pages 427–436.

Dayhoff, M. O. (1978). *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, D.C.

de Queiroz, A., Donoghue, M. J., and Kim, J. (1995). Seperate versus combined analysis of phylogenetic evidence. *Annual Review of Ecology and Systematics*, 26:657–681.

Delarbre, C., Escriva, H., Gallut, C., Barriel, V., Kourilsky, P., Janvier, P., Laudet, V., and Gachelin, G. (2000). The complete nucleotide sequence of the mitochondrial DNA of the agnathan *Lampetra fluviatilis*: bearings on the phylogeny of cyclostomes. *Molecular Biology and Evolution*, 17:519–529.

Delarbre, C., Gallut, C., Barriel, V., Janvier, P., and Gachelin, G. (2002). Complete mitochondrial DNA of the hagfish, *Eptatretus burgeri*: the comparative analysis of mitochondrial DNA sequences strongly supports the cyclostome monophyly. *Molecular Phylogenetics and Evolution*, 22:184–92.

DeLong, E. F. and Pace, N. R. (2001). Environmental diversity of Bacteria and Archaea. *Systematic Biology*, 50:470–478.

Doyle, J. J. (1992). Gene trees and species trees: molecular systematics as one-character taxonomy. *Systematic Botany*, 17:144–163.

Duret, L., Mouchiroud, D., and Gouy, M. (1994). HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Research*, 22:2360–5.

Edwards, A. W. F. (1972). *Likelihood.* Cambridge University Press.

Edwards, A. W. F. (1996). The origin and early development of the method of minimum evolution for the reconstruction of phylogenetic trees. *Systematic Biology*, 45:79–91.

Efron, B., Halloran, E., and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences, USA*, 93:1342934.

Eichler, E. E. (2001). Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends in Genetics*, 17:661–669.

Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research*, 8:163–167.

El-Mabrouk, N. (2000). Recovery of ancestral tetraploids. In *Comparative Genomics*, pages 465–477. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Escriva, H., Manzon, L., Youson, J., and Laudet, V. (2002). Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. *Molecular Biology and Evolution*, 19:1440–1450.

Eulenstein, O. (1997). A linear time algorithm for tree mapping. *Arbeitspapiere der GMD*, No. 1046.

Eyre-Walker, A. and Keightley, P. D. (1999). High genomic deleterious mutation rates in hominids. *Nature*, 397:344–347.

Eyre-Walker, A., Smith, N. H., and Smith, J. M. (1999). Reply to Macaulay et al. (1999): mitochondrial DNA recombination - reasons to panic. *Proceedings of the Royal Society of London Series B*, 266:2041–2042.

Feil, E. J., Holmes, E. C., Bessen, D. E., Chan, M. S., Day, N. P., Enright, M. C., Goldstein, R., Hood, D. W., Kalia, A., Moore, C. E., Zhou, J., and Spratt, B. G. (2001). Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proceedings of the National Academy of Sciences, USA*, 98:182–187.

Fellows, M., Hallett, M., and Stege, U. (1998). On the multiple gene duplication problem. In *Proceedings of the 9th International Symposium on Algorithms and Computation (ISAAC'98), Taejon, Korea*, volume 1533 of *Lecture Notes in Computer Science*, pages 347–356.

Felsenstein, J. (1973). On the use of the parsimony criterion for inferring evolutionary trees. *Systematic Zoology*, 22:250=256.

Felsenstein, J. (1978a). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27:401–411.

Felsenstein, J. (1978b). A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biological Journal of the Linnean Society*, 16:183–196.

Felsenstein, J. (1981). Evolutionary trees from DNA-sequences – a maximum-likelihood approach. *Journal of Molecular Evolution*, 17:368–376.

Felsenstein, J. (1985). Confidence-limits on phylogenies – an approach using the bootstrap. *Evolution*, 39:783–791.

Fitch, W. M. (1979). Cautionary remarks on using gene expression events in parsimony procedures. *Systematic Zoology*, 28:375–9.

Flajnik, M. F. and Kasahara, M. (2001). Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. *Immunity*, 15:351–362.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151:1531–1545.

Forey, P. and Janvier, P. (1993). Agnathans and the origin of jawed vertebrates. *Nature*, 361:129–134.

Forey, P. L. (1988). Golden jubilee for the coelacanth *Latimeria chalumnae*. *Nature*, 336:727–732.

Friedman, R. and Hughes, A. L. (2001). Pattern and timing of gene duplication in animal genomes. *Genome Research*, 11:1842–1847.

Friedman, R. and Hughes, A. L. (2003). The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Molecular Biology and Evolution*, 20:154–161.

Furlong, R. F. and Holland, P. W. H. (2002). Were vertebrates octoploid? *Philosophical Transactions of the Royal Society of London Series B*, 357:531–544.

Gale, M. D. and Devos, K. M. (1998). Plant comparative genetics after 10 years. *Science*, 282:656–658.

Gallardo, M. H., Bickham, J. W., Honeycutt, R. L., Ojed, R. A., and Kohler, N. (1999). Discovery of tetraploidy in a mammal. *Nature*, 401:341.

Garcia-Fernandez, J. and Holland, P. W. (1994). Archetypal organization of the *Amphioxus* hox gene cluster. *Nature*, 370:563–566.

Gascuel, O., Hendy, M. D., Jean-Marie, A., and McLachlan, R. (2003). The combinatorics of tandem duplication trees. *Systematic Biology*, 52:110–118.

Gatesy, J., O'Grady, P., and Baker, R. H. (1999). Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. *Cladistics*, 15:271–313.

Gaut, B. S. and Doebley, J. F. (1997). DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences, USA*, 94:6809–6814.

Gibson, T. J. and Spring, J. (2000). Evidence in favour of ancient octaploidy in the vertebrate genome. *Biochemical Society Transactions*, 28:259–264.

Goldman, N. (1990). Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process of DNA substitution and to parsimony analyses. *Systematic Zoology*, 39:345–361.

Goodman, M. (1999). The genomic record of humankind's evolutionary roots. *American Journal of Human Genetics*, 64:31–39.

Goodman, M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E., and Matsuda, G. (1979). Fitting the gene lineage into its species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, 28:132–168.

Goodman, M., Porter, C. A., Czelusniak, J., Page, S. L., Schneider, H., Shoshani, J., Gunnell, G., and Groves, C. P. (1998). Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. *Molecular Phylogenetics and Evolution*, 9:585–598.

Gordon, A. D. (1986). Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. *Journal of Classification*, 3:335–348.

Graham, R. L. and Foulds, L. R. (1982). Unlikelihood that minimal phyogenies for a realistic biological study can be constructed in reasonable computation time. *Mathematical Biosciences*, 60:133–142.

Groves, C. P. (1975). Taxonomy and phylogeny of prosimians. In Martin, R. D., Doyle, G. A., and Walker, A. C., editors, *Prosimian Biology*, pages 449–473. Univ of Pittsburgh, Pittsburgh.

Groves, C. P. (1989). *A Theory of Human and Primate Evolution.* Oxford University Press, Oxford, U.K.

Gu, X., Wang, J., and Gu, J. (2002). Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nature Genetics*, 31:205–209.

Gu, Z., Steinmetz, L. M., Gu, X., Sharfe, C., Davis, R. W., and Li, W.-H. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature*, 42:63–66.

Guigó, R., Muchnik, I., and Smith, T. F. (1996). Reconstruction of ancient molecular phylogeny. *Molecular Phylogenetics and Evolution*, 6:189–213.

Gursoy, H. C., Koper, D., and Benecke, B. J. (2000). The vertebrate 7S K RNA separates hagfish (*Myxine glutinosa*) and lamprey (*Lampetra fluviatilis*). *Journal of Molecular Evolution*, 50:456–64.

Guyer, C. and Slowinski, J. B. (1991). Comparison of observed phyogenetic topologies with null expectations among three monophyletic lineages. *Evolution*, 45:340–350.

Haldane, J. B. S. (1933). The part played by recurrent mutation in evolution. *American Naturalist*, 67:5–19.

Hallett, M. T. and Lagergren, J. (2000). New algorithms for the duplication-loss problem. In Shamir, R., Miyano, S., Istrail, S., Pevzner, P., and Waterman, M., editors, *RECOMB '00, Proceedings of the fourth annual international conference on computational molecular biology.* Association for Computing Machinery.

Hallett, M. T. and Lagergren, J. (2001). Efficient algorithms for lateral gene transfer problems. In Lengauer, T., editor, *RECOMB '01, Proceedings of the fifth annual international conference on computational molecular biology.* Association for Computing Machinery.

Harcourt-Brown, K. G. (2002). *Phylogenetic tree shape with special reference to the Cretaceous globotruncaid foraminifera.* PhD thesis, University of Bristol.

Harcourt-Brown, K. G., Pearson, P. N., and Wilkinson, M. (2001). The imbalance of palaeontological trees. *Paleobiology*, 27:188–204.

Harvey, P. H., May, R. M., and Nee, S. (1994). Phylogenies without fossils. *Evolution*, 48:523–529.

Heard, S. B. (1992). Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. *Evolution*, 46:1818–1826.

Heard, S. B. (1996). Patterns in phylogenetic tree shape with variable and evolving speciation rates. *Evolution*, 50:2141–8.

Heard, S. B. and Mooers, A. Ø. (1996). Imperfect information and the balance of cladograms and phenograms. *Systematic Biology*, 45:115–118.

Heckman, D. S., Geiser, D. M., Eidell, B. R., Stauffer, R. L., Kardos, N. L., and Hedges, S. B. (2001). Molecular evidence for the early colonization of land by fungi and plants. *Science*, 293:1129–1133.

Hedges, S. B. and Poling, L. L. (1999). A molecular phylogeny of reptiles. *Science*, 283:998–1001.

Helentjaris, T., Weber, D., and Wright, S. (1988). Identification of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polymorphisms. *Genetics*, 118:353–363.

Henikoff, S., Greene, E., Petrokovski, S., Bork, P., Attwood, T., and Hood, L. (1997). Gene families: the taxonomy of protein paralogs and chimaeras. *Science*, 278:609–614.

Hillis, D. M. and Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, 42:18292.

Hillis, D. M., Bull, J. J., White, M. E., Badgett, M. R., and Molineux, I. J. (1992). Experimental phylogenetics – generation of a known phylogeny. *Science*, 255:589–592.

Holland, P. W., Garcia-Fernandez, J., Williams, N. A., and Sidow, A. (1994). Gene duplications and the origins of vertebrate development. *Development*, Supplement:125–133.

Holland, P. W. H. (1999). Gene duplication: past, present and future. *Seminars in Cell and Developmental Biology*, 10:541–547.

Holmes, S. (2003). Statistics for phylogenetic trees. *Theoretical Population Biology*, 63:17–32.

Huelsenbeck, J. P. and Bull, J. J. (1996). A likelihood ratio test for detection of conflicting phylogenetic signal. *Systematic Biology*, 45:92–98.

Huelsenbeck, J. P. and Crandall, K. A. (1997). Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics*, 28:437–466.

Huelsenbeck, J. P. and Hillis, D. M. (1993). Success of phylogenetic methods in the four-taxon case. *Systematic Biology*, 42:247–265.

Huelsenbeck, J. P. and Kirkpatrick, M. (1996). Do phylogenetic methods produce trees with biased shapes? *Evolution*, 50:1418–1424.

Huelsenbeck, J. P., Larget, B., Miller, R. E., and Ronquist, F. (2002). Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic Biology*, 51:673–689.

Huelsenbeck, J. P., Rannala, B., and Larget, B. (2000a). A bayesian framework for the analysis of cospeciation. *Evolution*, 54:352–364.

Huelsenbeck, J. P., Rannala, B., and Masly, J. P. (2000b). Accomodating phylogenetic uncertainty in evolutionary studies. *Science*, 288:2349–2350.

Hughes, A. L. (1994). The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society of London Series B*, 256:119–124.

Hughes, A. L. (1998). Phylogenetic tests of block duplication of homologous genes on human chromosomes 6, 9, and 1. *Molecular Biology and Evolution*, 15:854–870.

Hughes, A. L. (1999a). *Adaptive Evolution of Genes and Genomes*. Oxford University Press, New York.

Hughes, A. L. (1999b). Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *Journal of Molecular Evolution*, 48:565–576.

Hughes, A. L., da Silva, J., and Friedman, R. (2001). Ancient genome duplications did not structure the human hox-bearing chromosomes. *Genome Research*, 11:771–780.

Janvier, P. (1981). The phylogeny of the craniata, with particular reference to the significance of fossil "agnathans". *Journal of Vertebrate Palaeontology*, 1:121–159.

Janvier, P. (1996). The dawn of the vertebrates: characters versus common ascent in the rise of current vertebrate phylogenies. *Palaeontology*, 39:259–287.

Janvier, P. (1998). A cold look at odd vertebrate phylogenies. *Journal of Molecular Evolution*, 46:375–377.

Johnson, M. E., Viggiano, L., Bailey, J. A., Abdul-Rauf, M., Goodwin, G., Rocchi, M., and Eichler, E. E. (2001). Positive selection of a gene family during the emergence of humans and african apes. *Nature*, 413:514–519.

Jow, H., Hudelot, C., Rattray, M., and Higgs, P. G. (2002). Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. *Molecular Biology and Evolution*, 19:1591–1601.

Kasahara, M., Hayashi, M., Tanaka, K., Inoko, H., Sugaya, K., Ikemura, T., and Ishibashi, T. (1996). Chromosomal localization of the proteasome z subunit gene reveals an ancient chromosomal duplication involving the major histocompatibility complex. *Proceedings of the National Academy of Sciences, USA*, 93:9096–9101.

Kasahara, M., Nakaya, J., Satta, Y., and Takahata, N. (1997). Chromosomal duplication and the emergence of the adaptive immune system. *Trends in Genetics*, 13:90–92.

Kennedy, M. and Page, R. D. M. (2002). Seabird supertrees: Combining partial estimates of procellariiform phylogeny. *Auk*, 119:88–108.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Kirkpatrick, M. and Slatkin, M. (1993). Searching for evolutionary patterns in the shape of phylogenetic trees. *Evolution*, 46:1818–1826.

Kitami, T. and Nadeau, J. H. (2002a). Biochemical networking contributes more to genetic buffering in human and mouse metabolic pathways than does gene duplication. *Nature Genetics*, 32:191–194.

Kitami, T. and Nadeau, J. H. (2002b). Corrigendum: Biochemical networking contributes more to genetic buffering in human and mouse metabolic pathways than does gene duplication. *Nature Genetics*, 32:681.

Kluge, A. (1989). A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). *Systematic Zoology*, 37:315–328.

Koop, B. F., Tagle, D. A., Goodman, M., and Slightom, J. L. (1989). A molecular view of primate phylogeny and important systematic and evolutionary questions. *Molecular Biology and Evolution*, 6:580–612.

Ku, H. M., Vision, T., Liu, J., and Tanksley, S. D. (2000). Comparing sequenced segments of the tomato and arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proceedings of the National Academy of Sciences, USA*, 97:9121–9126.

Kubo, T. and Iwasa, Y. (1995). Inferring the rates of branching and extinction from molecular phylogenies. *Evolution*, 49:694–704.

Kumar, S. and Hedges, S. B. (1998). A molecular timescale for vertebrate evolution. *Nature*, 392:917–920.

Kunin, V., Cases, I., Enright, A. J., de Lorenzo, V., and Ouzounis, C. A. (2003). Myriads of protein families, and still counting. *Genome Biology*, 4:401.

Kuraku, S., Hoshiyama, D., Katoh, K., Suga, H., and Miyata, T. (1999). Monophyly of lampreys and hagfishes supported by nuclear DNA-coded genes. *Journal of Molecular Evolution*, 49:729–735.

Kyrpides, N. (1999). Genomes OnLine Database (GOLD): a monitor of complete and ongoing genome projects world wide. *Bioinformatics*, 15:773–777.

Lander, E. S. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921.

Lapointe, F.-J. and Cucumel, G. (1997). The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. *Systematic Biology*, 46:306–312.

Larget, B. and Simon, D. L. (1999). Markov chain Monte Carlo algorithms for the bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, 16:750–759.

Larhammar, D., Lundin, L. G., and Hallbook, F. (2002). The human hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. *Genome Research*, 12:1910–1920.

Leipoldt, M. (1983). Towards an understanding of the molecular mechanisms regulating gene expression during diploidization in phylogenetically polyploid lower vertebrates. *Human Genetics*, 65:11–18.

Levausser, C. and Lapointe, F.-J. (2001). War and peace in phylogenetics: A rejoinder on total evidence and consensus. *Systematic Biology*, 50:881–892.

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 50:913–926.

Li, M., Thorp, J., and Zhang, L. (1996). On the nearest neighbour interchange distance between evolutionary trees. *Journal of Theoretical Biology*, 182:463–467.

Li, W.-H. (1980). Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. *Genetics*, 95:237–258.

Li, W.-H. (1997). *Molecular Evolution*. Sinauer, Sunderland, Massachusetts.

Li, W.-H., Gu, Z., Wang, H., and Nekrutenko, A. (2001). Evolutionary analyses of the human genome. *Nature*, 409:847–849.

Liao, D. (2000). Gene conversion drives within genic sequences: concerted evolution of ribosomal RNA genes in bacteria and archaea. *Journal of Molecular Evolution*, 51:305–317.

Lin, X. (1999). Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature*, 402:761–768.

Linder, H. P. and Crisp, M. D. (1995). *Nothofagus* and pacific biogeography. *Cladistics*, 11:5–32.

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

Llorente, B., Malpertuy, A., Neuveglise, C., de Montigny, J., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E., Brottier, P., Casaregola, S., Durrens, P., Gaillardin, C., Lepingle, A., Ozier-Kalogeropoulos, O., Potier, S., Saurin, W., Tekaia, F., Toffano-Nioche, C., Wesolowski-Louvel, M., Wincker, P., Weissenbach, J., Souciet, J., and Dujon, B. (2000). Genomic exploration of the hemiascomycetous yeasts: 18. Comparative analysis of chromosome maps and synteny with *Saccharomyces cerevisiae*. *FEBS Letters*, 487:101–112.

Long, M. and Langley, C. H. (1993). Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. *Science*, 260:91–95.

Long, M. and Thornton, K. (2001). Gene duplication and evolution. *Science*, 293:1551.

Losos, J. B. and Adler, F. R. (1995). Stumped by trees? A generalized null model for patterns of organismal diversity. *American Naturalist*, 145:329–342.

Løvtrup, A. (1977). *The phylogeny of the Vertebrata*. Wiley, New York.

Lundin, L. G. (1993). Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics*, 16:1–19.

Lynch, M. and Conery, J. C. (2001). Gene duplication and evolution: Response. *Science*, 293:1551.

Lynch, M. and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, 290:1151–1155.

Ma, B., Li, M., and Zhang, L. (1998). On reconstructing species trees from gene trees in term of duplications and losses. In Istrail, S., Pevzner, P. A., and Waterman, M. S., editors, *Proceedings of the Second Annual International Conference on Computational Biology (RECOMB 98)*, pages 182–191. ACM, New York.

Macaulay, V., Richards, M., and Sykes, B. (1999). Mitochondrial DNA recombination – no need to panic. *Proceedings of the Royal Society of London Series B*, 266:2037–2039.

Maddison, W. P. (1989). Reconstructing character evolution on polytomous cladograms. *Cladistics*, 5:365–377.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46:523–536.

Maisey, J. G. (1984). Higher elasmobranch phylogeny and biostratigraphy. *Zoological Journal of the Linnean Society*, 82:33–54.

Mallatt, J. (1997). Hagfish do not resemble ancestral vertebrates. *Journal of Morphology*, 232:293.

Mallatt, J. and Sullivan, J. (1998). 28S and 18S rDNA sequences support the monophyly of lampreys and hagfish. *Molecular Biology and Evolution*, 15:1706–1718. refd in Kuraku et al (191).

Margush, T. and McMorris, F. R. (1981). Consensus n-trees. *Bulletin of Mathematical Biology*, 43:239–44.

Martin, A. (1999a). Increasing genomic complexity by gene duplication and the origin of the vertebrates. *American Naturalist*, 154:111–128.

Martin, A. P. (2001). Is tetralogy true? Lack of support for the "one-to-four" rule. *Molecular Biology and Evolution*, 18:89–93.

Martin, A. P. and Burg, T. M. (2002). Perils of paralogy: using hsp70 genes for inferring organismal phylogenies. *Systematic Biology*, 51:570–587.

Martin, J., Herniou, E., Cook, J., O'Neill, R. W., and Tristem, M. (1999). Interclass transmission and phyletic host tracking in murine leukemia virus-related retroviruses. *Journal of Virology*, 73:2442–2449.

Martin, W. (1999b). Mosaic bacterial chromosomes: a challenge on route to a tree of genomes. *Bioessays*, 21:99–104.

McLysaght, A. (2001). *Evolution of vertebrate genome organisation.* PhD thesis, Department of Genetics, Trinity College, Dublin.

McLysaght, A., Hokamp, K., and Wolfe, K. H. (2002). Extensive genomic duplication during early chordate evolution. *Nature Genetics*, 31:200–204.

McLysaght, A., Seoighe, C., and Wolfe, K. (2000). High frequency of inversions during eukaryote gene order evolution. In *Comparative Genomics*, pages 47–58. Kluwer Academic Publishers, Dordrecht, The Netherlands.

McMorris, F. R. and Steel, M. (1993). The complexity of the median procedure for binary trees. *4th Conference of the International Federation of Classification Societies, Paris.*

Meyerowitz, E. M. (2001). Prehistory and history of arabidopsis research. *Plant Physiology*, 125:15–19.

Mindell, D. P. and Meyer, A. (2001). Homology evolving. *Trends in Ecology and Evolution*, 16:434–440.

Mirkin, B., Muchnik, I., and Smith, T. F. (1996). A biologically consistent model for comparing molecular phylogenies. *Journal of Computational Biology*, 2:493–507.

Mirkin, B. G., Fenner, T. I., Galperin, M. Y., and Koonin, E. V. (2003). Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evolutionary Biology*, 3:2.

Miyamato, M. M. and Fitch, W. M. (1995). Testing species phylogenies and phylogenetic methods with congruence. *Systematic Biology*, 44:64–76.

Mooers, A. Ø. (1995). Tree balance and tree completeness. *Evolution*, 49:379–384.

Mooers, A. Ø. and Heard, S. (1997). Inferring evolutionary process from phylogenetic tree shape. *Quarterly Review of Biology*, 72:31–54.

Moritz, C., Dowling, T. E., and Brown, W. M. (1987). Evolution of animal mitochondrial DNA: Relevance for population biology and systematics. *Annual Review of Ecology and Systematics*, 18:269–292.

Morris, S. C. (1999). Palaeodiversifications: mass extinctions, "clocks", and other worlds. *Geobios*, 32:165–174.

Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A., and O'Brien, S. J. (2001). Molecular phylogenetics and the origins of placental mammals. *Science*, 409:614–8.

Nadeau, J. H. and Sankoff, D. (1997). Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics*, 147:1259–1266.

Naylor, G. J. and Brown, W. M. (1997). Structural biology and phylogenetic estimation. *Nature*, 388:527–8.

Nee, S., Holmes, E. C., May, R. M., and Harvey, P. H. (1994). Extinction rates can be estimated from molecular phylogenies. *Philosophical Transactions of the Royal Society of London Series B*, 344:77–82.

Nee, S., Holmes, E. C., Rambaut, A., and Harvey, P. H. (1995). Inferring population history from molecular phylogenies. *Philosophical Transactions of the Royal Society of London Series B*, 349:25–31.

Nee, S., Mooers, A. Ø., and Harvey, P. H. (1992). Tempo and modes of evolution revealed from molecular phylogenies. *Proceedings of the National Academy of Sciences, USA*, 89:8322–8366.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Evolution*, 48:443–453.

Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nei, M., Rogozin, I. B., and Piontkivska, H. (2000). Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proceedings of the National Academy of Sciences, USA*, 97:10866–10871.

Nelson, J. S. (1994). *Fishes of the World*. John Wiley & Sons, New York, NY, 3rd edition.

Ng, M. P. and Wormald, N. C. (1996). Reconstruction of rooted trees from subtrees. *Discrete Applied Mathematics*, 69:19–31.

Nixon, K. and Carpenter, J. (1996). On simultaneous analysis. *Cladistics*, 12:221–241.

Ochman, H. (2001). Lateral gene transfer and the nature of bacterial innovation. *Current Opinion in Genetics and Development*, 11:616–619.

Ohno, S. (1970). *Evolution by Gene Duplication*. George Allen & Unwin, London.

Ohno, S. (1985). Dispensable genes. *Trends in Genetics*, 1:160–164.

Ohno, S. (1998). The notion of the Cambrian pananimalia genome and a genomic difference that separated vertebrates from invertebrates. In *Molecular Evolution: Evidence for monophyly of Metazoa*. Springer-Verlag, New York.

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

Ohta, T. (1988a). Further simulation studies on evolution by gene duplication. *Evolution*, 42:375–386.

Ohta, T. (1988b). Time for acquiring a new gene by duplication. *Proceedings of the National Academy of Sciences, USA*, 85:3509–3512.

Ohta, T. (1989). Role of gene duplication in evolution. *Genome*, 31:304–10.

Ono, K., Suga, H., Iwabe, N., Kuma, K., and Miyata, T. (1999). Multiple protein tyrosine phosphates in sponges and explosive gene duplication in the early evolution of animals before the parazoan-eumetazoan split. *Journal of Molecular Evolution*, 48:654–662.

Pace, N. R., Stahl, D. A., Lane, D. J., and Olsen, G. J. (1986). The analysis of microbial populations by ribosomal RNA sequences. *Advances in Microbial Ecology*, 9:1–55.

Page, R. D. M. (1988). Quantitative cladistic biogeography: constructing and comparing area cladograms. *Systematic Zoology*, 37:254–270.

Page, R. D. M. (1993a). Component analysis: a valiant failure? *Cladistics*, 6:119–136.

Page, R. D. M. (1993b). *COMPONENT, Tree comparison software for Microsoft Windows*. The Natural History Museum, London.

Page, R. D. M. (1994a). Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology*, 43:58–77.

Page, R. D. M. (1994b). Parallel phylogenies: reconstructing the history of host-parasite assemblages. *Cladistics*, 10:155–73.

Page, R. D. M. (1996). On consensus, confidence, and 'total' evidence. *Cladistics*, 12:83–92.

Page, R. D. M. (1998). GENETREE: comparing gene and species phylogenies using reconciled trees. *Bioinformatics*, 14:819–820.

Page, R. D. M. (2000). Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. *Molecular Phylogenetics and Evolution*, 14:89–106.

Page, R. D. M. (2002a). Modified mincut supertrees. In Guigó, R. and Gusfield, D., editors, *Proceedings of WABI 2002*, volume 2452 of *Lecture Notes in Computer Science*, pages 537–551. Springer-Verlag.

Page, R. D. M. (2002b). Treemap versus bpa (again): A response to dowling. Technical Report 02-02, Taxonomy, Systematics, and Bioinformatics Group, University of Glasgow.

Page, R. D. M. (2003). Introduction. In *Tangled Trees: phylogeny, cospeciation and coevolution*, pages 1–21. University of Chicago Press.

Page, R. D. M. and Charleston, M. A. (1997a). From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. *Molecular Phylogenetics and Evolution*, 7:231–240.

Page, R. D. M. and Charleston, M. A. (1997b). Reconciled trees and incongruent gene and species trees. In Mirkin, B., McMorris, F., Roberts, F., and Rzhetsky, A., editors, *Mathematical Hierarchies in Biology*, volume 37 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 57–70. American Mathematical Society, Providence, Rhode Island.

Page, R. D. M. and Charleston, M. A. (1998). Trees within trees: Phylogeny and historical associations. *Trends in Ecology and Evolution*, 13:356–359.

Page, R. D. M. and Cotton, J. A. (2000). GENETREE: a tool for exploring gene family evolution. In Sankoff, D. and Nadeau, J. H., editors, *Comparative Genomics*, pages 525–536. Kluwer Academic Publishers, Utrecht.

Page, R. D. M. and Cotton, J. A. (2002). Vertebrate phylogenomics: reconciled trees and gene duplications. In Altman, R. B., Dunker, A. K., Hunter, L., Lauderdale, K., and Klein, T. E., editors, *Pacific Symposium on Biocomputing*, pages 536–547. World Scientific Press, Singapore.

Page, R. D. M. and Holmes, E. C. (1998). *Molecular Evolution: A Pylogenetic Approach*. Blackwell Scientific, Oxford.

Page, S. L. and Goodman, M. (2001). Catarrhine phylogeny: noncoding DNA evidence for a diphyletic origin of the mangabeys and for a human-chimpanzee clade. *Molecular Phylogenetics and Evolution*, 18:14–25.

Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution*, 5:568–83.

Patterson, C., Williams, D. M., and Humphries, C. J. (1993). Congruence between molecular and morphological phylogenies. *Annual Review of Ecology and Systematics*, 24:153–188.

Pebusque, M. J., Coulier, F., Birnbaum, D., and Pontarotti, P. (1998). Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Molecular Biology and Evolution*, 15:1145–1159.

Penny, D., Murray-McIntosh, R. P., and Hendy, M. D. (1998). Estimating times of divergence with a change of rate: the orangutan-african ape divergence. *Molecular Biology and Evolution*, 15:608–610.

Penny, D., Watson, E. E., and Steel, M. A. (1993). Trees from languages and genes are very similar. *Systematic Biology*, 42:382–384.

Pisani, D., Yates, A. M., Langer, M. C., and Benton, M. J. (2002). A genus-level supertree of the Dinosauria. *Proceedings of the Royal Society of London Series B*, 269:915–921.

Porter, C. A., Czelusniak, J., Schneider, H., Schneider, M. P. C., Sampaio, I., and Goodman, M. (1997). Sequences of the primate epsilon-globin gene: implications for systematics of the marmosets and other new world primates. *Gene*, 205:59–71.

Porter, C. A., Goodman, M., and Stanhope, M. J. (1996). Evidence on mammalian phylogeny from sequences of exon 28 of the von Willebrand factor gene. *Molecular Phylogenetics and Evolution*, 5:89–101.

Prince, V. E. and Pickett, F. B. (2002). Splitting pairs: the diverging fates of duplicated genes. *Nature Review Genetics*, 3:827–837.

Purvis, A. (1995a). A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society of London Series B*, 348:405–421.

Purvis, A. (1995b). A modification to Baum and Ragan's method for combining phylogenetic trees. *Systematic Biology*, 44:251–255.

Pybus, O. G., Rambaut, A., Holmes, E. C., and Harvey, P. H. (2003). New inferences from tree shape: numbers of missing taxa and population growth rates. *Systematic Biology*, 51:881–889.

Ragan, M. A. (1992). Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution*, 1:53–58.

Rambaut, A. and Bromham, L. (1998). Estimating divergence dates from molecular sequences. *Molecular Biology and Evolution*, 15:442–448.

Ramsey, J. and Schemske, D. W. (1998). Pathways, mechanisms and rates of polyploid formation in flowering plants. *Annual Review of Ecology and Systematics*, 29:467–501.

Ramsey, J. and Schemske, D. W. (2002). Neopolyploidy in flowering plants. *Annual Review of Ecology and Systematics*, 33:589–639.

Rasmussen, A. S. and Arnason, U. (1999). Phylogenetic studies of complete mitochondrial DNA molecules place cartilaginous fishes within the tree of bony fishes. *Journal of Molecular Evolution*, 48:118–23.

Rasmussen, A. S., Janke, A., and Arnason, U. (1998). The mitochondrial DNA molecule of the hagfish (*Myxine glutinosa*) and vertebrate phylogeny. *Journal of Molecular Evolution*, 46:382–8.

Rieppel, O. (2000). Turtles as diapsid reptiles. *Zoologica Scripta*, 29:199–212.

Rieppel, O. and deBraga, M. (1996). Turtles as diapsid reptiles. *Nature*, 384:453–455.

Robinson, D. F. (1971). Comparison of labelled trees with valency three. *Journal of Combinatorial Theory B*, 11:105–119.

Rodriguez-Trelles, F., Tarrio, R., and Ayala, F. J. (2002). A methodological bias toward overestimation of molecular evolutionary time scales. *Proceedings of the National Academy of Sciences, USA*, 99:8112–8115.

Rogers, J. S. (1993). Response of Colless's tree imbalance to number of terminal taxa. *Systematic Biology*, 42:102–105.

Rogers, J. S. (1994). Central moments and probability distribution of Colless's coefficient of tree imbalance. *Evolution*, 48:2026–2036.

Rogers, J. S. (1996). Central moments and probability distribution of three measures of phylogenetic tree imbalance. *Systematic Biology*, 45:99–110.

Ronquist, F. (1996). Matrix representation of trees, redundancy, and weighting. *Systematic Biology*, 45:247–253.

Ronquist, F. (2003). Parsimony analysis of coevolving species associations. In *Tangled Trees: phylogeny, cospeciation and coevolution*, pages 22–64. University of Chicago Press.

Ronquist, F. and Nylin, S. (1990). Process and pattern in the evolution of species associations. *Systematic Zoology*, 39:323–344.

Rosen, D. E. (1978). Vicariant patterns and historical explanation in biogeography. *Systematic Zoology*, 27:159–188.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425.

Salamin, N., Hodkinson, T. R., and Savolainen, V. (2002). Building supertrees: an empirical assessment using the grass family. *Systematic Biology*, 51:136–150.

Sanderson, M. J. (1997). A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution*, 14:1218–1231.

Sanderson, M. J., Purvis, A., and Henze, C. (1998). Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology and Evolution*, 13:105–109.

Sanderson, M. J. and Shaffer, H. B. (2002). Troubleshooting molecular phylogenetic analyses. *Annual Review of Ecology and Systematics*, 33:49–72.

Schmidt, H. A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002). Tree-puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18:502–504.

Semple, C. (2003). Reconstructing minimal rooted trees. *Discrete Applied Mathematics*, In Press, available online through http://www.sciencedirect.com.

Semple, C. and Steel, M. (2000). A supertree method for rooted trees. *Discrete Applied Mathematics*, 105:147–158.

Seoighe, C. and Wolfe, K. H. (1998). Extent of genomic rearrangement after genome duplication in yeast. *Proceedings of the National Academy of Sciences, USA*, 95:4447–4452.

Seoighe, C. and Wolfe, K. H. (1999). Updated map of duplicated regions in the yeast genome. *Gene*, 238:253–261.

Sharman, A. C. and Holland, P. W. (1998). Estimation of the hox gene cluster number in lampreys. *International Journal of Developmental Biology*, 42:617–620.

Shoshani, J., Groves, C. P., Simons, E. L., and Gunnell, G. F. (1996). Primate phylogeny: morphological vs molecular results. *Molecular Phylogenetics and Evolution*, 5:102–154.

Sicheritz-Pontén, T. and Andersson, S. G. E. (2001). A phylogenomic approach to microbial evolution. *Nucleic Acids Research*, 29:545–552.

Sidow, A. (1996). Gen(om)e duplications in the evolution of early vertebrates. *Current Opinion in Genetics and Development*, 6:715–722.

Simberloff, D., Heck, K. L., McCoy, E. D., and Connor, E. F. (1981). There have been no statistical tests of cladistic biogeography hypotheses. In *Vicariance Biogeography: A Critique*, pages 40–63. Columbia University Press, New York.

Simmons, M. P., Bailey, C. D., and Nixon, K. C. (2000). Phylogeny reconstruction using duplicate genes. *Molecular Biology and Evolution*, 17:469–473.

Simmons, M. P. and Freudenstein, J. V. (2002). Uninode coding vs gene tree parsimony for phylogenetic reconstruction using duplicate genes. *Molecular Phylogenetics and Evolution*, 23:481–498.

Skrabanek, L. and Wolfe, K. H. (1998). Eukaryotic genome duplication – where's the evidence? *Current Opinion in Genetics and Development*, 8:694–700.

Slowinksi, J. and Page, R. D. M. (1999). How should species phylogenies be inferred from sequence data? *Systematic Biology*, 48:814–825.

Slowinski, J. B., Knight, A., and Rooney, A. P. (1997). Inferring species trees from gene trees: a phylogenetic analysis of the Elapidae (Serpentes) based on the amino acid sequences of venom proteins. *Molecular Phylogenetics and Evolution*, 8:349–362.

Slowinski, J. B. and Page, R. D. M. (1999). How should phylogenies be inferred from sequence data? *Systematic Biology*, 48:814–825.

Small, R. L. and Wendel, J. F. (2000). Copy number lability and evolutionary dynamics of the adh gene family in diploid and tetraploid cotton (*Gossypium*). *Genetics*, 155:1913–1926.

Smith, N. G., Knight, R., and Hurst, L. D. (1999). Vertebrate genome evolution: a slow shuffle or a big bang? *Bioessays*, 21:697–703.

Sokal, R. R. and Rohlf, F. J. (1995). *Biometry, 3rd ed.* W. H. Freeman and Co., New York.

Soltis, D. E. and Soltis, P. S. (1995). The dynamic nature of polyploid genomes. *Proceedings of the National Academy of Sciences, USA*, 92:8095–8091.

Soltis, D. E. and Soltis, P. S. (1999). Polyploidy: recurrent formation and genome evolution. *Trends in Ecology and Evolution*, 14:348–352.

Spring, J. (1997). Vertebrate evolution by interspecifc hybridization – are we polyploid? *FEBS Letters*, 400:2–8.

Spring, J. (2002). Genome duplication strikes back. *Nature Genetics*, 31:128–129.

Stam, E. (2002). Does imbalance in phylogenies reflect only bias? *Evolution*, 56:1292–1295.

Stauffer, R. L., Walker, A., Ryder, O. A., Lyons-Weiler, M., and Hedges, S. B. (2001). Human and ape molecular clocks and constraints on paleontological hypotheses. *J. Heredity*, 92:469–474.

Steel, M. (1992). The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, 9:91–116.

Steel, M., Dress, A. W. M., and Böcker, S. (2000). Simple but fundamental limitations on supertree and consensus tree methods. *Systematic Biology*, 49:363–368.

Steel, M. and McKenzie, A. (2001). Properties of phylogenetic trees generated by Yule-type speciation models. *Mathematical Biosciences*, 170:91–112.

Stock, D. and Whitt, G. (1992). Evidence from 18S ribosomal RNA sequences that lampreys and hagfish form a natural group. *Science*, 257:787–789.

Strimmer, K. and von Haeseler, A. (1996). Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution*, 13:964–969.

Suga, H., Hoshiyama, D., Kuraku, S., Katoh, K., Kubokawa, K., and Miyata, T. (1999). Extensive gene duplication in the early evolution of animals before the parazoan-eumetazoan split demonstrated by g proteins and protein tyrosine kinases from sponge and hydra. *Journal of Molecular Evolution*, 48:646–653.

Sullivan, J. (1996). Combining data with different distributions of among-site rate variation. *Systematic Biology*, 45:375–380.

Suzuki, M., Kubokawa, K., Nagasawa, H., and Urano, A. (1995). Sequence analysis of vasotocin cDNAs of the lamprey *Lampetra japonica*, and the hagfish, *Eptratretus burgeri*: evolution of cyclostome vasotocin precursors. *Journal of Molecular Endocrinology*, 14:67–77.

Swofford, D. L. (1991). When are phylogenetic estimates from molecular and morphological data incongruent? In Miyamato, M. M. and Cracraft, J., editors, *Phylogenetic Analysis of DNA sequences*, pages 295–333. Oxford University Press.

Swofford, D. L. (1998). *PAUP* - Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sinauer Associates, Sunderland, Massachusetts.

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic inference. In Hillis, D. M., Moritz, C., and Mable, B. K., editors, *Molecular Systematics, 2nd Edition*. Sinauer, Sunderland, Massachusetts.

Syvanen, M. (1994). Horizontal gene transfer: evidence and possible consequences. *Annual Review of Genetics*, 28:237–261.

Takezaki, N. and Gojobori, T. (1999). Correct and incorrect vertebrate phylogenies obtained by the entire mitochondrial DNA sequences. *Molecular Biology and Evolution*, 16:590–601.

Takezaki, N., Rzhetsky, A., and Nei, M. (1995). Phylogenetic test of the molecular clock and linearized trees. *Molecular Biology and Evolution*, 12:823–833.

Tavarè, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetic models. *Theoretical Population Biology*, 26:119–164.

Tavarè, S., Marshall, C. R., Will, O., Soligo, C., and Martin, R. D. (2002). Using the fossil record to estimate the age of the last common ancestor of extant primates. *Nature*, 416:726–729.

Terryn, N., Heijnen, L., Keyser, A. D., Asseldonck, M. V., Clercq, R. D., Verbakel, H., Gielen, J., Zabeau, M., Villarroel, R., Jesse, T., Neyt, P., Hogers, R., Daele, H. V. D., Ardiles, W., Schueller, C., Mayer, K., Dehais, P., Rombauts, S., Montagu, M. V., Rouze, P., and Vos, P. (1999). Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analyzing a 400-kb contig at the APETALA2 locus on chromosome 4. *FEBS Letters*, 445:237–245.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the owering plant *Arabidopsis thaliana*. *Nature*, 408:796–815.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–80.

Thorley, J. L. (2000). *Cladistic information, Leaf stability and Supertree construction*. PhD thesis, University of Bristol.

Thorley, J. L. and Wilkinson, M. (1999). Testing the phylogenetic stability of early tetrapods. *Journal of Theoretical Biology*, 200:343–344.

Thorley, J. L. and Wilkinson, M. (2003). Reduced consensus and supertree methods. In Janowitz, M., Lapointe, F.-J., McMorris, F., Mirkin, B., and Roberts, F., editors, *Bioconsensus, In Press*. DIMACS-AMS.

Thorne, J. L., Goldman, N., and Jones, D. T. (1996). Combining protein evolution and secondary structure. *Molecular Biology and Evolution*, 13:666–673.

Thorne, J. L., Kishino, H., and Painter, I. S. (1998). Estimating the rate of evolution of the rate of evolution. *Molecular Biology and Evolution*, 15:1647–1657.

Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L. X., Kirkness, E. F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H. G., Glodek, A., McKenney, K., Fitzegerald, L. M., Lee, N., Adams, M. D., Hickey, E. K., Berg, D. E., Gocayne, J. D., Utterback, T. R., Peterson, J. D., Kelley, J. M., Cotton, M. D., Weldman, J. M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W. S., Weidman, J. M., Borodovsky, M., Karp, P. D., Smith, H. O., Fraser, C. M., and Venter, J. C. (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*, 388:539–547.

Tuffley, C. and Steel, M. (1997). Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology*, 59:581–607.

Venter, J. C. (2001). The sequence of the human genome. *Science*, 291:1304–1351.

Vision, T. J., Brown, D. G., and Tanksley, S. D. (2000). The origins of genomic duplications in arabidopsis. *Science*, 290:2114–2117.

Waddell, P. J., Okada, N., and Hasegawa, M. (1999). Towards resolving the interordinal relationships of placental mammals. *Systematic Biology*, 48:1–5.

Wagner, A. (1998). The fate of duplicated genes: loss or new function? *BioEssays*, 20:785–788.

Wagner, A. (2000). Robustness against mutations in genetic networks of yeast. *Nature Genetics*, 24:355–361.

Walsh, J. B. (1995). How often do duplicated genes evolve new functions? *Genetics*, 139:421–8.

Wang, Y. and Gu, X. (2000). Evolutionary patterns of gene families generated in the early stage of vertebrates. *Journal of Molecular Evolution*, 51:88–96.

Wareham, H. T. (1993). On the computational complexity of inferring evolutionary trees. Technical Report 9301, Department of Computer Science, Memorial University of Newfoundland.

Waterman, M. S. and Smith, T. F. (1978). On the similarity of dendrograms. *Journal of Theoretical Biology*, 73:789–800.

Watterson, G. A. (1983). On the time for gene silencing at duplicate loci. *Genetics*, 105:745–766.

Wendel, J. F. (2000). Genome evolution in polyploids. *Plant Molecular Biology*, 42:225–249.

Wilson, R. J. (1996). *Introduction to Graph Theory, 4th edition.* Longman, Harlow, Essex.

Woese, C. R. (2000). Interpreting the universal phylogenetic tree. *Proceedings of the National Academy of Sciences, USA*, 97:8392–8396.

Woese, C. R. and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences, USA*, 74:5088–5090.

Wolfe, K. H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics*, 2:333–41.

Wolfe, K. H. and Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387:708–713.

Wood, B. (2002). Palaeoanthropology: Hominid revelations from chad. *Nature*, 418:133 – 135.

Wray, G. (2001). Dating branches on the tree of life using DNA. *Genome Biology*, 3:0001.0001–0001.0007.

Yang, Z., Goldman, N., and Friday, A. (1995). Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Systematic Biology*, 44:384–399.

Yang, Z. and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution*, 19:908–917.

Yoder, A. D., Cartmill, M., Ruvolo, M., Smith, K., and Vilgalys, R. (1996). Ancient single origin for Malagasy primates. *Proceedings of the National Academy of Sciences, USA*, 93:5122–5126.

Yoder, A. D. and Yang, Z. (2000). Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution*, 17:1081–1090.

Yule, G. U. (1924). A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London Series A*, 213:21–87.

Zardoya, R. and Meyer, A. (1996). The complete nucleotide sequence of the mitochondrial genome of the lungfish (*Protopterus dolloi*) supports its phylogenetic position as a close relative of land vertebrates. *Genetics*, 142:1249–1263.

Zardoya, R. and Meyer, A. (1997). The complete DNA sequence of the mitochondrial genome of a "living fossil," the coelacanth (*Latimeria chalumnae*). *Genetics*, 146:995–1010.

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

Zardoya, R. and Meyer, A. (2001b). Vertebrate phylogeny: limits of inference of mitochondrial genome and nuclear rDNA sequence data due to an adverse phylogenetic signal/noise ratio. In Ahlberg, P. E., editor, *Major events in early vertebrate evolution*, pages 135–156. Taylor and Francis, London.

Zhang, J. and Nei, M. (1996). Evolution of antennapedia-class homeobox genes. *Genetics*, 142:295–303.

Zhang, L. (1997). On a Mirchkin-Muchnik-Smith conjecture for comparing molecular phylogenies. *Journal of Computational Biology*, 4:177–187.

Zhang, L. (2000). Inferring a species tree from gene trees under the deep coalescence cost. *Poster, RECOMB2000, Tokyo, Japan.*

Zhang, L., Gaut, B. S., and Vision, T. J. (2001). Gene duplication and evolution. *Science*, 293:1551.

Zharkikh, A. and Li, W. H. (1992a). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. i. four taxa with a molecular clock. *Molecular Biology and Evolution*, 9:111947.

Zharkikh, A. and Li, W. H. (1992b). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. ii. four taxa without a molecular clock. *Journal of Molecular Evolution*, 35:356–366.

Zietkiewicz, E., Richer, C., and Labuda, D. (1999). Phylogenetic affinities of tarsier in the context of primate alu repeats. *Molecular Phylogenetics and Evolution*, 11:77–83.

Zmasek, C. M. and Eddy, S. R. (2001). A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics*, 17:821–828.

Zuckerandl, E. and Pauling, L. (1965). Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, 8:357–366.

Appendix A

GENETREE: A tool for exploring gene family evolution[1]

Molecular biologists interested in the evolution of gene families and molecular systematists interested in the evolution of whole organisms are both concerned with the relationship between gene phylogenies and organism phylogenies. We present reconciled trees as a tool for exploring this relationship. In discussing recent developments, we focus on techniques which enable researchers to take account of uncertainty in the underlying gene phylogenies and to locate gene duplications and episodes of gene duplication on the species tree. Implementation of these methods should allow rapid, automated analysis of large sets of gene families and even of whole genomes, producing well supported organism phylogenies and allowing us to quantitatively investigate patterns of gene family evolution.

A.1 Introduction

Evolutionary trees for gene sequences are studied from two complementary, but distinct, perspectives. Molecular biologists seek to understand the evolution of the structure and function of a particular gene, and discover relationships among families of genes. Molecular systematists use gene trees to recover organismal phylogeny. Central to both perspectives is the relationship between gene and organismal phylogeny.

[1]This appendix has been published in *Comparative Genomics*, D. Sankoff and J. H. Nadeau, eds. Kluwer Academic Publishers. It was co-authored with Rod Page.

The key assumption that motivates molecular systematics is that evolutionary trees for genes also contain information about the evolutionary relationships of organisms. Indeed, it is often assumed that gene trees are the same as species trees – hence one can obtain a species tree simply by sequencing the same gene in a range of species, and replacing the names of the genes with the names of the corresponding species. However, two observations contradict this assumption: (1) species may contain more than one copy of the same gene, and (2) different gene trees may imply different species trees. If two or more copies of a gene are sequenced (for example, haemoglobin α and β from *Homo sapiens*) then replacing the genes by the corresponding species will result in the same species occurring more than once in the tree. In this case there is no longer a one-to-one correspondence between the gene and species trees, raising the problem of how to extract the latter from the former. If different gene trees support different species trees (i.e. the gene trees are incongruent) then this raises the question of how to choose among these alternative species trees.

For molecular biologists, the relationship between gene and organismal phylogeny can be crucial in identifying orthologous genes. If only single copies of a gene have been sequenced in a range of taxa, it may not be obvious from the gene tree alone whether the genes are orthologous or paralogous. Comparison of gene and species trees can identify unrecognised instances of paralogy among genes. Once the history of gene duplication and loss events is determined for a set of genes, broader evolutionary questions can be asked, such as rates of gene duplication and loss, and the relative timing of duplications in different gene families.

The analysis of gene family phylogenies represents a considerable challenge for the study of genome evolution, especially when one considers how common gene duplication has clearly been in some taxa. Within vertebrates, paralogy is pervasive (Figure A.1) and a similar picture is found in the Eubacteria and Archaea when data from Hobacgen (Perrière et al., 2000) are examined.

Our goal here is to explore some issues in the analysis of gene family evolution using reconciled trees as implemented in GENETREE (Page, 1998). This software package is freely available for Windows 95/NT and MacOS operating systems from `http://taxonomy.zoology.gla.ac.uk/rod/genetree/genetree.html`. To illustrate specific points we use the L-lactate dehydrogenase (L-LDH) gene family (`http://www.expasy.ch/cgi-bin/nicezyme.`

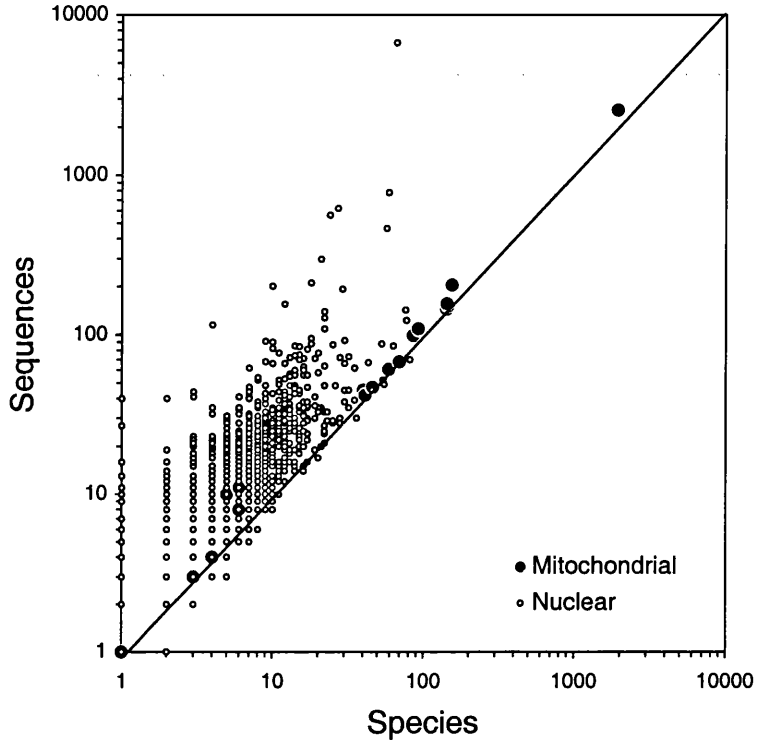

Figure A.1: Number of sequences plotted against number of species for vertebrate gene families in release 29 (March 17, 1998) of the HOVERGEN (Duret et al., 1994) data base. Note that usually each species has a single mitochondrial sequence for a given gene (hence, the mitochondrial genes fall along the 1:1 line), whereas most nuclear genes are present in multiple copies. Due to redundancy in species names (for example, "human" and "*Homo sapiens*" being used to describe the source of different genes in the same family), some gene families appear to have fewer sequences than species. From Slowinksi and Page (1999, fig. 1).

pl?1.1.1.27), which has often served as a model data set for developing ideas about reconciled trees (Martin, 1999a; Page, 1994; Page and Charleston, 1997a) and about gene family evolution more generally (Holmes, 1972; Li et al., 1983).

A.2 Reconciled trees

A reconciled tree is the simplest embedding of a gene tree within a species tree. The technique has its origins in Goodman et al. (1979), a study of haemoglobin gene evolution where there were significant discrepancies between gene and organismal phylogenies. Suppose we have a phylogeny for four species and a phylogeny for four genes sampled from those species, and that the gene and species trees – which we believe to be correct – disagree (Figure A.2a).

The question is, how can the trees both be true, and yet be discordant? One approach is to embed the gene tree in the species tree (Figure A.2b), which requires us to postulate a number of gene duplications and subsequent gene losses (in this instance one duplication and three losses). This embedding can also be represented using a reconciled tree (Figure A.2c), which simply takes the embedded gene tree and "unfolds" it so that it lies flat on the page. The reconciled tree depicts the complete history of the gene if there had been no gene losses. In this example, given the gene duplication we would expect species 2, 3, and 4 to each have two copies of the gene. It is the presence of only one copy of the gene in each of these species that leads us to infer three gene losses. An alternative explanation for these "losses" is that the other copy of the gene is present in these species, but as yet undetected. Given the unevenness of the sampling of different organisms (indicated by the preponderance of a few model organisms in the sequence data banks), this may often be the case. Indeed, the "losses" indicated by the reconciled tree can be viewed as predictions about the existence of undiscovered genes. In the example shown, further sequencing may uncover copy 1 in species 4, and copy 2 in species 2 and 3. The reconciled tree also shows that genes b and c are paralogous to gene d, which is not apparent from the gene phylogeny alone. This highlights the role organismal phylogeny can play in identifying homology relationships among genes. Direct evidence for paralogy is the presence of multiple genes in the same species (e.g., haemoglobin α and β in the same species), but many additional paralogous genes may be identified using reconciled trees.

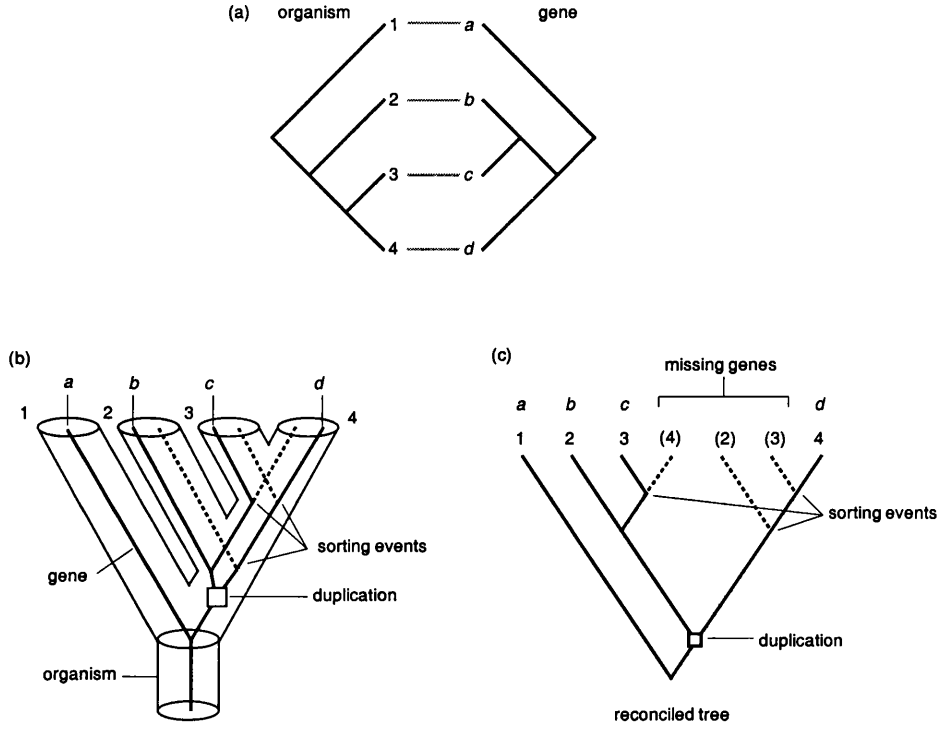

Figure A.2: (a) Incongruent gene and species trees. This incongruence can be explained by hypothesising a gene duplication (h) at the base of the gene tree (b). The presence of only a single gene (a-d) extant in each of the present-day species (1-4) requires postulating three gene losses. (c) The corresponding reconciled tree. After Page (2000).

A.3 Inferring species trees

One basic goal of analysing gene families is to shed light on the evolutionary relationships of the organisms from which those genes were obtained. Given one or more gene trees we can ask what species tree would accommodate those gene trees with the fewest number of duplications and losses (Page and Charleston, 1997a). The problem of finding the optimal species tree is NP-complete (Ma et al., 1998), so we must rely on heuristics for all but the smallest problems.

GENETREE implements a simple "hill-climbing" heuristic, where an initial species tree (either a random tree or one supplied by the user) is rearranged in search of a species tree with a better cost. Random trees provide a useful tool for exploring the tree landscape (Charleston, 1995), but searches that start from a random tree tend to be time consuming. Often it is substantially quicker to start from a species tree based on some other evidence, such as the currently accepted taxonomic classification. However, this may bias the results, especially if a poor rearrangement strategy is used. The importance of effective search strategies is emphasised by Page and Charleston (1997b), who used GENETREE to find substantially more parsimonious species trees than those found by Guigó et al. (1996) using the same set of eukaryote gene trees.

The extreme taxonomic bias of the sequence data bases towards a few model organisms (93% of vertebrate nucleotide sequences in GENBANK come from humans, rats or mice) means it is almost certainly the case that not all genes will have been discovered (or, indeed, looked for) in all the taxa of interest. This can lead to cases where species will be grouped on the absence of genes, rather than on actual evidence of their relationship. This problem is avoided by using the number of duplications alone as the optimality criteria for selecting species trees (Page and Charleston, 1997a), but this could lead to incorrect assumptions of orthology if actual gene loss events are common. Missing sequences also lead to a rapid increase in the number of species trees that are equally parsimonious explanations of the gene trees (Page, 2000). Where some taxa are sampled for only one or few gene families, this poor taxonomic overlap will result in some of these many parsimonious species trees being biologically absurd. One solution to this problem is to use constraint trees (Constantinescu and Sankoff, 1986) to enforce some species groupings that are considered incontrovertible (such as "mammals"), but clearly

this requires us to accept some species relationships *a priori*.

New algorithms for finding optimal species trees are appearing. Stege (1999) presents a fixed-parameter tractable algorithm (Downey and Fellows, 1998) for finding the species tree that minimises the number of duplications for a set of gene trees, parameterised by the number of duplications needed. Hallett and Lagergren (2000) have developed an algorithm minimising both duplications and losses where the parameter is the "width" – the maximum number of gene lineages that coexist in a species at any one time. These algorithms can find the globally optimal species trees in cases where their parameter values are small – generally in fairly simple cases – and the latter has been used to show that the species trees found by Page and Charleston (1997b) were indeed the most parsimonious.

A.4 Uncertain gene trees

Gene trees inferred from sequence data are estimates of the true gene tree. So far we have assumed that the gene tree is obtained without error, but this will rarely be the case. Figure A.3 shows a phylogeny for vertebrate L-LDH sequences. Some of the species relationships implied by this tree (figure A.4b) seem anomalous: the two amphibians are not grouped together, the shark is basal to tetrapods and the relationships between mammalian orders are unconventional. This suggests that the gene tree may not be entirely accurate.

It may be that an alternative gene tree - less parsimonious or less likely than the optimal tree - is the actual gene tree, and the fit between gene and species tree could be used as an additional criterion for selecting among competing gene trees. Goodman et al. (1979) suggested such a strategy in their pioneering work on reconciled trees, in which they preferred less parsimonious haemoglobin gene trees which had better fit to accepted species trees than most parsimonious trees that required more duplications and losses. Their approach assigned each gene tree a total score based on the length of the tree in terms of number of nucleotide substitutions plus the number of gene duplications and losses, where each type of event had the same cost. This drew immediate criticism from Fitch (1979), who argued that there was no obvious way of determining the relative cost of a nucleotide substitution versus a gene duplication. Another approach would be to consider a set of gene trees for each gene, such as those comprising a "confidence interval" around

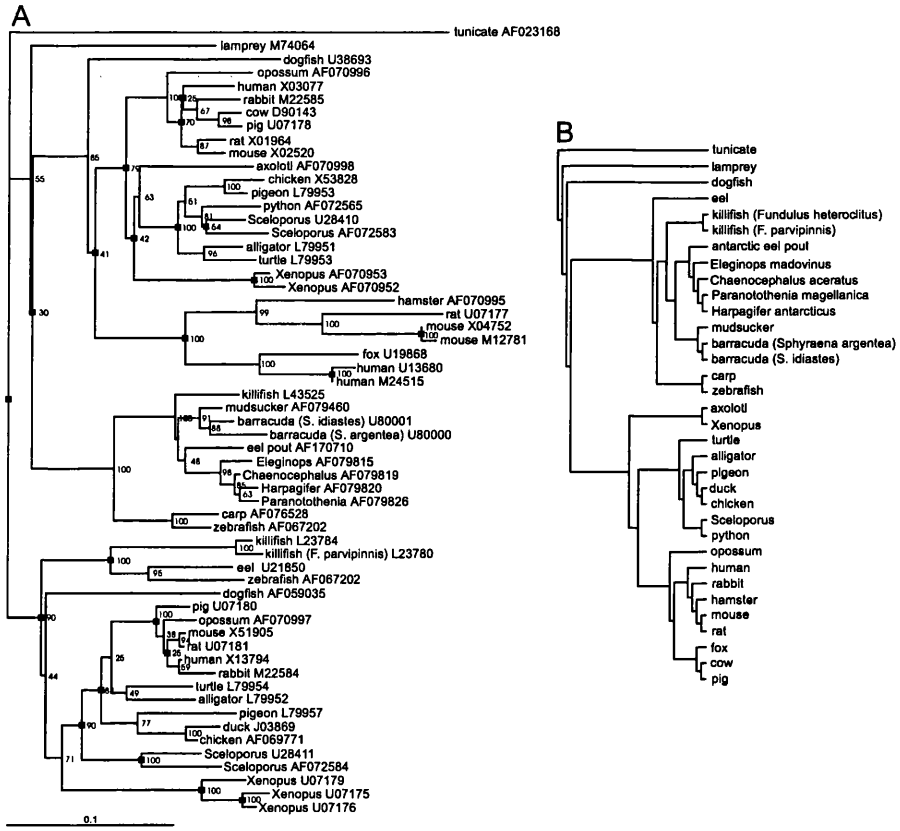

Figure A.3: (a) Neighbour joining tree for vertebrate L-LDH sequences, rooted with a tunicate ("sea squirt") as the outgroup, with GENBANK accession numbers. The numbers on the internal nodes of the tree are bootstrap values, the scale bar represents 0.1 amino acid replacements per site. Gene duplications required by reconciling this tree with currently accepted relationships amongst the species (b) are shown as filled boxes.

the optimal gene tree (Page, 1996; Sanderson, 1989). The best estimate of the gene tree would be that tree within the confidence interval that had the best fit to the species tree. Martin (1999a) chose the L-LDH gene phylogeny with lowest duplication and loss cost that was not significantly worse than the most parsimonious gene tree, effectively giving a greater weight to duplications and losses than to substitution events.

Alternative approaches to the problem of uncertainty in gene trees deserve to be explored. One method would be to rearrange the optimal gene tree to improve its fit to the species tree. This idea has been formalised by Chen et al. (2000), who describe a simple greedy rearrangement algorithm that takes the initial estimate of the gene tree and performs nearest neighbour rearrangements (Waterman and Smith, 1978) around nodes with bootstrap support less than some specified value. This inverts the problem from one of finding the optimal species tree given a gene tree to one of finding the optimal gene tree, within certain constraints, given a species tree. A maximum likelihood framework has been suggested in the context of coalescence models by Maddison (1997). However, while reasonable statistical models of nucleotide substitution exist, there are none yet for gene duplication, and any such model would need to incorporate the extreme sampling bias that exists in the sequence databases (and hence that many gene "losses" are sampling artefacts).

Uncertainty in gene trees also has implications for inferring species trees. Available implementations of reconciled trees do not give any measure of the degree of support for any nodes in the species tree. This makes it difficult to evaluate competing hypotheses, such as the relationships among hagfish and lampreys. Reconciled tree analysis of nine vertebrate gene families supported grouping the lamprey with the rest of the vertebrates, to the exclusion of the hagfish (Page, 2000), whereas analyses of ribosomal genes suggest hagfish and lampreys are sister taxa (Mallatt and J. Sullivan, 1998). One brute force approach to coping with uncertainty in gene trees would be to construct species trees for each tree in the set of bootstrap trees for a gene family and use the majority rule consensus (Margush and McMorris, 1981) of those resulting trees as the best estimate of species relationships. Applying this to the L-LDH sequences, we get the species tree shown in figure A.4a, revealing which relationships are only weakly supported by the L-LDH data.

If one has a set of gene families one could apply resampling methods to those families. This is analogous to Felsenstein's use of the bootstrap on sequence data

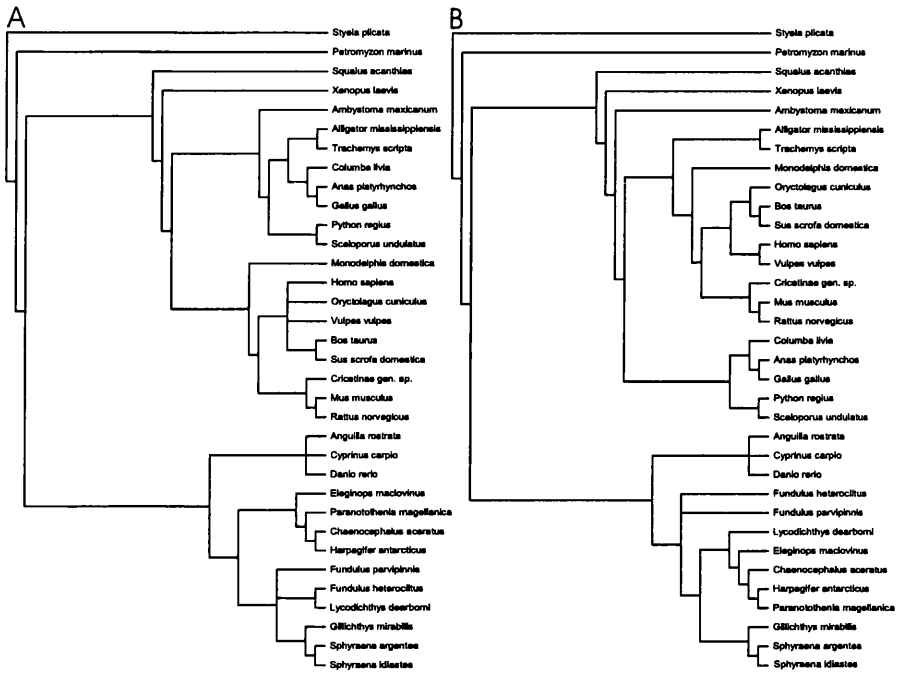

Figure A.4: (a) Majority rule consensus tree for selected vertebrate species based on 100 bootstrap gene trees for L-LDH. (b) Strict consensus of 9 optimal species tree for the L-LDH data, requiring 12 duplications and 32 losses.

(Felsenstein, 1985), however, we would resample the gene families rather than the nucleotide or amino acid sites for each gene family. This amounts to treating each gene family as a single character.

A.5 Locating gene duplications

> Take four, or maybe eight, decks of 52 playing cards. Shuffle them all together and then throw some cards away. Pick 20 cards at random and drop the rest on the floor. Give the 20 cards to some evolutionary biologists and ask them to figure out what you've done.(Skrabanek and Wolfe, 1998, p. 698)

Although the mapping between a gene and species tree is unique (Page and Charleston, 1997b) – and hence each node in the gene tree is mapped onto a single node in the species tree – if the species tree is poorly sampled then there will still be ambiguity in the actual location of a duplication on the species tree. This ambiguity means that many gene duplications may cluster together, indicating DNA duplication events affecting large stretches of sequence, or even whole genomes. Genome duplication has been posited as a major factor in the evolution of complexity in vertebrates, although there is considerable debate as to the number and location of these putative duplications (Figure A.5). Recent analyses (Martin, 1999b) using an earlier implementation of reconciled trees (Page, 1993) suggest that gene duplications within vertebrates have been largely independent.

Guigó et al. (1996) encountered this ambiguity in their study of eukaryote gene families. They reconciled 53 gene family trees with a species tree comprising 16 taxa. Because many of their gene trees were small (comprising 4-5 genes) there was some ambiguity in the placement of some of these duplications. Using a heuristic algorithm to cluster together the duplications, they found that the 46 duplications could be accounted for by five genome duplications at four different points on the species tree.

Currently implemented algorithms for reconciled trees assume that duplications in different gene families are independent, that is, the algorithms seek to minimise the number of gene duplications. Minimising the number of episodes of gene duplication is a significantly harder problem (Fellows et al., 1998).

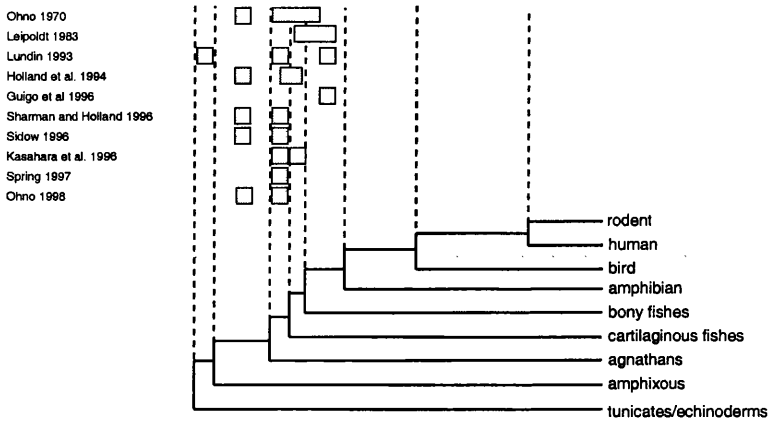

Figure A.5: Alternative hypotheses of genome duplication in vertebrates. The phylogeny is drawn with branch lengths proportional to time. From Martin (1999b, fig. 1).

A.6 Future

As more and more gene trees are assembled, the metaphor of a simple tree of life becomes increasingly strained, leading us to view organism phylogeny as a "cloud" or statistical distribution of gene histories, largely congruent with one another but showing significant variance (Maddison, 1997). Gene duplication and loss may not be the only cause of this variance. Horizontal transfer of genes makes reconstructing the history of a gene much more difficult, but can be addressed with reconciled trees using techniques developed for an analogous situation in the context of host-parasite coevolution (Charleston, 1998). Horizontal transfer seems unlikely to be of any great importance in vertebrate gene families, but would certainly have to be addressed in other cases, e.g. in bacteria (Martin, 1999c).

There is also the inevitable lag between theoretical developments and their implementation in software. The current release of GENETREE has some of these developments, such as a linear time algorithm for tree mapping (Eulenstein, 1997), but has yet to include more recent results.

Another pragmatic issue is how well the software can cope with the ever growing flood of sequence data. GENETREE was originally conceived as a test bed for algorithms for displaying reconciled trees. There is now a need to enable it to handle numerous, large gene families. For example, it would be very useful to be

able to extract gene trees from data bases like HOVERGEN (Duret et al., 1994) and input these directly into GENETREE. It would then be possible to obtain the best estimates of species phylogeny based on simultaneous analysis of thousands of gene families, and to locate episodes of gene duplication in these families. Work on this is currently in progress.

Bibliography

Charleston, M. A. (1995). Towards a characterization of landscapes of combinatorial optimisation problems, with special reference to the phylogeny problem. *Journal of Computational Biology*, 2:439–50.

Charleston, M. A. (1998). Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Mathematical Biosciences*, 149:191–223.

Chen, K., Durand, D., and Farach-Colton, M. (2000). Notung: dating gene duplications using gene family trees. *RECOMB2000*.

Constantinescu, M. and Sankoff, D. (1986). Tree enumeration modulo a consensus. *Journal of Classification*, 3:349–56.

Downey, R. G. and Fellows, M. R. (1998). *Parameterized Complexity*.

Duret, L., Mouchiroud, D., and Gouy, M. (1994). Hovergen: a database of homologous vertebrate genes. *Nucleic Acids Research*, 22:2360–2365.

Eulenstein, O. (1997). A linear time algorithm for tree mapping. *Arbeitspapiere der GMD*, No. 1046.

Fellows, M., Hallett, M., and Stege, U. (1998). On the multiple gene duplication problem. In *Proceedings of the 9th International Symposium on Algorithms and Computation (ISAAC'98), Taejon, Korea*, volume 1533 of *Lecture Notes in Computer Science*, pages 347–356.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39:783–91.

Fitch, W. M. (1979). Cautionary remarks on using gene expression events in parsimony procedures. *Systematic Zoology*, 28:375–9.

Goodman, M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E., and Matsuda, G. (1979). Fitting the gene lineage into its species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, 28:132–168.

Guigó, R., Muchnik, I., and Smith, T. F. (1996). Reconstruction of ancient molecular phylogeny. *Molecular Phylogenetics and Evolution*, 6:189–213.

Hallett, M. T. and Lagergren, J. (2000). New algorithms for the duplication-loss problem. *RECOMB2000*.

Holmes, R. S. (1972). Evolution of lactate dehydrogenase genes. *FEBS Letters*, 28:51–55.

Li, S. S.-L., Fitch, W. M., Pan, Y.-C. E., and Sharief, F. S. (1983). Evolutionary relationships of vertebrate lactate dehydrogenase isozymes A_4 (muscle), B_4 (heart), and C_4 (testis). *The Journal of Biological Chemistry*, 258:7029–7032.

Ma, B., Li, M., and Zhang, L. (1998). On reconstructing species trees from gene trees in term of duplications and losses. *RECOMB98*.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46:523–536.

Mallatt, J. and J. Sullivan, J. (1998). 28s and 18s rdna sequences support the monophyly of lampreys and hagfishes. *Molecular Biology and Evolution*, 15:1706–1718.

Margush, T. and McMorris, F. R. (1981). Consensus n-trees. *Bulletin of Mathematical Biology*, 43:239–44.

Martin, A. (1999a). Choosing among alternative trees of multi-gene families. *Molecular Phylogenetics and Evolution*, (In Press).

Martin, A. (1999b). Increasing genomic complexity by gene duplication and the origin of the vertebrates. *American Naturalist*, 154:111–128.

Martin, W. (1999c). Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *BioEssays*, 21:99–104.

Page, R. D. M. (1993). *COMPONENT, Tree comparison software for Microsoft Windows*. The Natural History Museum, London.

Page, R. D. M. (1994). Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology*, 43:58–77.

Page, R. D. M. (1996). On consensus, confidence, and 'total' evidence. *Cladistics*, 12:83–92.

Page, R. D. M. (1998). Genetree: comparing gene and species trees using reconciled trees. *Bioinformatics*, 14:819–820.

Page, R. D. M. (2000). Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. *Molecular Phylogenetics and Evolution*, 14:89–106.

Page, R. D. M. and Charleston, M. A. (1997a). From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Molecular Phylogenetics and Evolution*, 7:231–240.

Page, R. D. M. and Charleston, M. A. (1997b). Reconciled trees and incongruent gene and species trees. In Mirkin, B., McMorris, F., Roberts, F., and Rzhetsky, A., editors, *Mathematical Hierarchies in Biology*, volume 37 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 57–70. American Mathematical Society, Providence, Rhode Island.

Perrière, G., Duret, L., and Gouy, M. (2000). Hobacgen: Database system for comparative genomics in bacteria. *Genome Research*, 10:379–385.

Sanderson, M. J. (1989). Confidence limits on phylogenies: The bootstrap revisisted. *Cladistics*, 5:113–29.

Skrabanek, L. and Wolfe, K. H. (1998). Eukaryotic genome duplication - where's the evidence? *Current Opinion in Genetics and Development*, 8:694–700.

Slowinksi, J. and Page, R. D. M. (1999). How should species phylogenies be inferred from sequence data? *Systematic Biology*, 48:814–825.

Stege, U. (1999). Gene trees and species trees: The gene-duplication problem is fixed-parameter tractable. Technical Report 319, Department of Computer Science, ETH Zurich.

Waterman, M. S. and Smith, T. F. (1978). On the similarity of dendrograms. *Journal of Theoretical Biology*, 73:789–800.

Appendix B

Vertebrate Phylogenomics: Reconciled Trees and Gene Duplications[1]

Ancient gene duplication events have left many traces in vertebrate genomes. Reconciled trees represent the differences between gene family trees and the species phylogeny those genes are sampled from, allowing us to both infer gene duplication events and estimate a species phylogeny from a sample of gene families. We show that analysis of 118 gene families yields a phylogeny of vertebrates largely in agreement with other data. We formulate the problem of locating episodes of gene duplication as a set cover problem: given a species tree in which each node has a set of gene duplications associated with it, the smallest set of species nodes whose union includes all gene duplications specifies the locations of gene duplication episodes. By generating a unique mapping from this cover set we can determine the minimal number of such episodes at each location. When applied to our data, this method reveals a complex history of gene duplications in vertebrate evolution that does not conform to the "2R" hypothesis.

[1]This chapter has been published in *Pacific Syposium on Biocomputing, 2002*, R. B. Altman, A.K. Dunker, L. Hunter, K. Lauderdale and T. E. Klein, Eds., World Scientific Press (See `http://psb.stanford.edu/`)

B.1 Introduction

Most genes belong to large gene families, so the analysis of the gene family evolution represents a considerable challenge for the study of genome evolution. Within vertebrates, paralogy (the relationship between genes within a family) is pervasive, and gene duplication has clearly been particularly common (Page and Cotton, 2000), but a broadly similar pattern is found in prokaryotes. The timing and frequency of gene duplications is of particular interest, given that gene (and genome) duplication has been posited as a major factor in the evolution of complexity in vertebrates (Ohno, 1999). A popular – and controversial (Hughes, 1999; Skrabanek and Wolfe, 1998) – hypothesis of vertebrate genome evolution postulates two successive genome duplications early in vertebrate evolution (the "2R" hypothesis). Understanding the evolution of vertebrate genomes requires a well supported phylogenetic framework for vertebrates, and methods for locating episodes of gene duplication. In this paper we explore the use of reconciled trees (Goodman et al., 1979; Page, 1994) to address the latter question.

B.1.1 Reconciled trees

Conventional phylogenetic methods use molecular sequences as characters of organisms, which conflates organismal and gene phylogenies. However, gene phylogenies are not species phylogenies - processes such as gene duplication, gene loss, and lineage sorting can introduce important differences between the correct phylogenetic tree for a set of genes and the correct tree for the corresponding species. An alternative is to investigate the relationship between gene trees and species trees using reconciled trees. A reconciled tree (Goodman et al., 1979; Page, 1994) is a map between a gene tree and a given species tree, with gene duplications and losses being postulated to explain any incongruence between the two trees. If the species tree is unknown then the most parsimonious estimate of the species tree is that minimising the number of gene duplications required on a gene tree (Page and Charleston, 1998; Slowinski and Page, 1999). We can extend the method to many genes, so the most parsimonious species tree is that which implies the minimum number of gene duplication (or duplication and loss) events over the set of gene families (Slowinski and Page, 1999, "gene tree parsimony"). The map between a gene tree and a species can be computed in linear time (Zhang, 1997), mak-

ing reconciled trees practicable for very large analyses, and potentially even for genome-wide comparisons.

B.1.2 Vertebrate phylogeny

To test the performance of gene tree parsimony on a real dataset, we constructed a data set of 118 vertebrate gene families[2] based on data from the HOVERGEN database (Duret et al., 1994). The higher-level phlyogeny and ancient evolution of the vertebrate in many ways represents an ideal test-case for these methods, because there has been considerable recent interest in both their phylogeny and in evolution by gene duplication in the group. A fairly robust consensus on the main relationships within the group had emerged, based on morphological evidence from both fossil and extant taxa (Benton, 1988), but analyses of whole mitochondrial genomes have produced unorthodox and controversial phylogenies, provoking new debate (Zardoya and Meyer, 2001).

The species tree we obtained using gene tree parsimony (Fig. B.1) differs little from a conventional view of vertebrate phylogeny (Benton, 1988), in marked contrast to the unorthodox trees obtained from mitochondrial genomes (Zardoya and Meyer, 2001). This result confirms preliminary findings (Page, 2000) that reconciled tree methods can reconstruct phylogeny accurately in the face of gene duplication and loss.

B.1.3 Genome duplications

The timing and location of gene duplications is a key problem in understanding the evolution of gene families and genomes. Existing techniques for mapping gene trees onto species trees can identify gene duplications, but do not necessarily locate them precisely on the species tree. Furthermore, gene duplication events can occur on any scale, from small pieces of DNA carrying fragments of genes right up to polyploidisation events due to hybridisation or incorrect division, so duplications on individual gene trees could be correlated, occurring as a result of the same molecular events. Identifying these events is complicated by the fact that most gene families are known from only some species, so there can be considerable

[2]The GENETREE file and individual alignments and gene trees are available from `http://kimura.zoology.gla.ac.uk/vertebrate_data`.

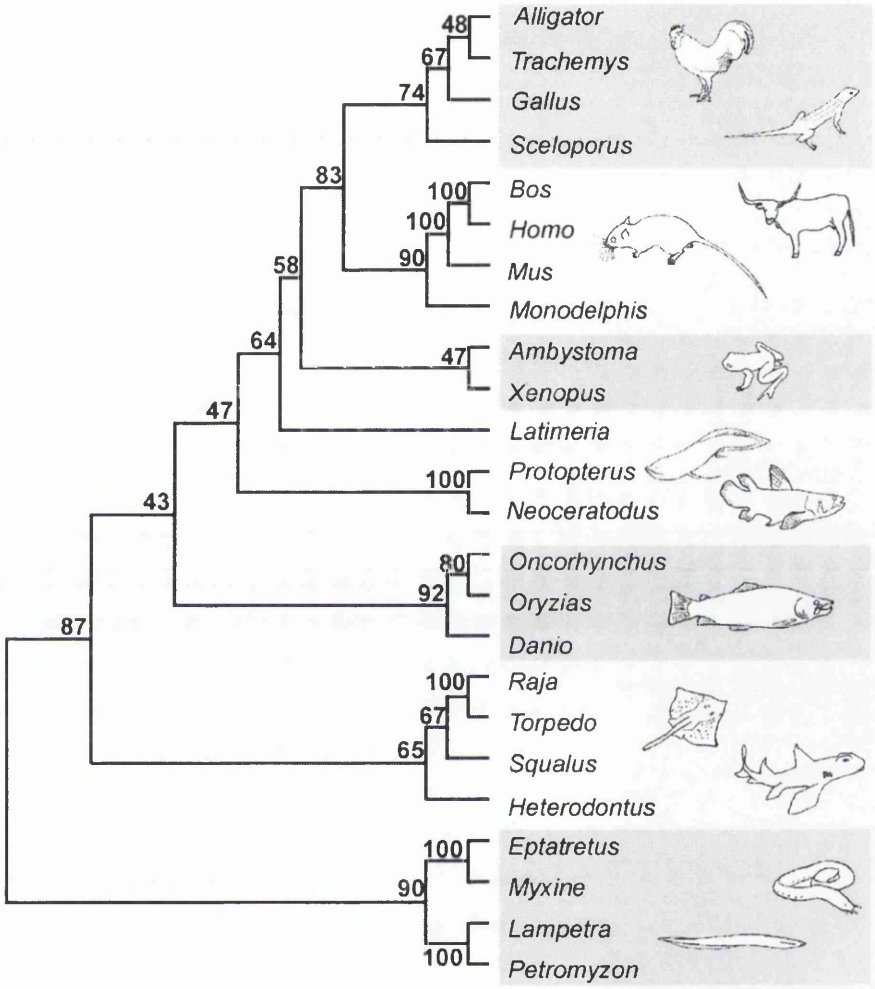

Figure B.1: Phylogeny of vertebrates reconstructed using gene tree parsimony in GENETREE (Page, 1998) on a set of 118 nuclear genes. Bands of shading identify higher taxonomic groups of vertebrates. This is the majority-rule consensus of 100 species trees, generate from 100 bootstrap trees for each gene tree. Figures on nodes are bootstrap percentages.

uncertainty in where particular duplications occurred on the species tree. We need techniques that can identify these "duplication episodes" by clustering individual gene duplications (Fellows et al., 1998; Guigó et al., 1996). We now present a method for achieving this and apply the technique to our vertebrate data set.

B.2 Locating gene duplications

B.2.1 Terminology

We will restrict ourselves to rooted trees. The immediate ancestor of a node in a tree is its *parent*, and the immediate descendants of a node are its *children*. A node with no children is a *leaf*. Let G be a rooted tree for m genes obtained from $n \leq m$ species (a *gene tree*), and S be a rooted tree for the species (a *species tree*). For each node in S the set of nodes that are its descendants form that nodes *cluster*. The cluster of the root is $\{1, \ldots, n\}$, the clusters of the leaves are $\{1\}, \{2\}, \ldots, \{n\}$. Following Margush and McMorris (Margush and McMorris, 1981), we use the shorthand of treating the node and its cluster as synonymous. Hence, for any pair of nodes x and y in S, if $x \subset y$ then x is a descendant of y. For any node $g \in G$, let $\eta(g)$ be the set of species in which occur the extant genes descendant from g (if g is a leaf then $\eta(g)$ is the species from which gene g was obtained). For any $g \in G$, let $M(g)$ be the node in S with the smallest cluster satisfying $\eta(g) \subseteq M(g)$. A map from G into S associates each node $g \in G$ with a node $M(g) \in S$, and can be visualised using a reconciled tree (Page, 1994). Let l and r be the left and right children of a node $g \in G$. If either l or r (or both) map onto $M(g)$ (i.e., $M(l) = M(g)$ and/or $M(r) = M(g)$) then we infer that g is a gene duplication (Goodman et al., 1979).

B.2.2 The problem

The problem of locating gene duplications using reconciled trees was first addressed by Guigó *et al.* (1996), who noted that the map between gene tree and species tree puts bounds on the location of a given duplication, rather than necessarily locating the duplication precisely. Whereas the map between gene and species tree associates each node g in the gene tree with a single node $M(g) = s$ in the species tree, the actual gene duplication may have occurred anywhere along the

path between $M(g)$ and $M(parent(g))$ [3]. Given this ambiguity, our task is to find the optimal placement of the duplications required to reconcile a set of gene trees $G_1, G_2, \ldots, G_k$ with a species tree S. It is important to clearly distinguish between episodes of gene duplication and *genome* duplication. Guigó *et al.* refer to *any* clustering of gene duplications as a "genome duplication," regardless of whether the whole genome or only a part of it duplicated. Here we use the term "episode" as the generic term for two or more duplications in different gene families that can be explained by a single event.

B.2.3 Guigó et al.'s algorithm for placing duplications

Guigó *et al.* partition gene duplications into three categories:

> *free*: if g is the root of G.
> *locked*: if g is not the root of G.
> *absolutely locked*: if g is locked and $M(parent(g)) = parent(M(g))$.

Examples of these three categories can be seen in Figure B.2b. Guigó *et al.* sketched an algorithm to cluster gene duplications into the minimum number of locations on the species tree. First we identify the set of allowed locations A_g in the species tree for a duplication g. If g is the root of the gene tree then $A_g = \{s \in S : M(g) \subseteq s\}$ (the set of all nodes in the species tree from $M(g)$ down to the root). If g is not the root of the gene tree then $A_g = \{s \in S : M(g) \subseteq s \subseteq M(parent(g))\}$ (the set of nodes in the species tree from $M(g)$ down to, but not including, the node into which the parent of g is mapped). Duplications are placed as follows:

> Step 1: Place on the species tree S all absolutely locked duplications (for which $A_g = M(g)$). The set of locations of absolutely locked duplications is $D_{absolute}$.
>
> Step 2: For all locked duplications g_l for which $A_{g_l} \cap D_{absolute} \neq \emptyset$ find the absolutely locked duplication(s) ($g_a : A_{g_l} \cap A_{g_a} \neq \emptyset$). If $|A_{g_l} \cap A_{g_a}| > 1$

[3] Note that moving a duplication down the species tree towards the root will require additional losses to be postulated. However, given that many apparent "losses" in reconciled trees may be due to lack of knowledge (such as poor taxonomic or genomic sampling), rather than actual gene loss, invoking additional losses does not seem unreasonable.

place g_L at the $s \in A_{g_l} \cap A_{g_a}$ that is furthest from the root of S. The set of locations of locked duplications is D_{locked}.

Step 3: For all locked duplications g_l for which $A_{g_l} \cap D_{absolute} = \emptyset$, if $A_{g_l} \cap D_{locked} = \emptyset$ then g_l is placed at the node $M(g)$, otherwise the duplication is placed such that the total number of locations of gene duplications is minimal.

Step 4: Free duplications g_f for which $A_{g_f} \cap D_{locked} \neq \emptyset$ are placed at the node $s \in A_{g_f} \cap D_{locked}$ that is furthest from the root of S, otherwise they are placed at the root of S.

The result of applying these steps is a clustering of gene duplications into episodes, and a final mapping of duplications onto the species tree. Note that although Guigó *et al.* gave hints about how to minimize the number of gene duplications (Step 3) they did not present a formal algorithm for doing this.

B.2.4 An alternative formulation

Fellows *et al.* (Fellows et al., 1998) define the MULTIPLE GENE DUPLICATION problem as being the mapping of a set of gene trees $G_1, G_2, \ldots, G_k$ into a species tree S such that the number of multiple gene duplication events is minimal. They go on to show that this problem is *NP*-hard. Their formulation of the problem is somewhat different from Guigó *et al.*'s – those authors aim to minimise the number of locations in S where gene duplications have occurred, but do not postulate any additional duplications over and above those required to reconcile each gene tree G_i with S. Fellows *et al.*, however, will invoke additional duplications if it reduces the number of multiple gene duplication events. For example, given the two gene trees in Figure B.2a, using the rules of Guigó *et al.* the duplication at node ABCDE in G_1 is absolutely locked and hence cannot be moved. However, Fellows *et al.* move this duplication to the root of the species tree (at the cost of an additional duplication). Similarly, Fellows *et al.* state that "it is not beneficial" to move node ABC in G_2. However, in Guigó *et al.*'s terminology, this duplication is not absolutely locked and could be placed anywhere along the path from ABC to ABCDEF in S. Moving it to node ABCDE in S reduces the number of multiple gene duplications from 4 to 3, the same score as for the Fellows *et al.* reconstruction, but without invoking an extra duplication.

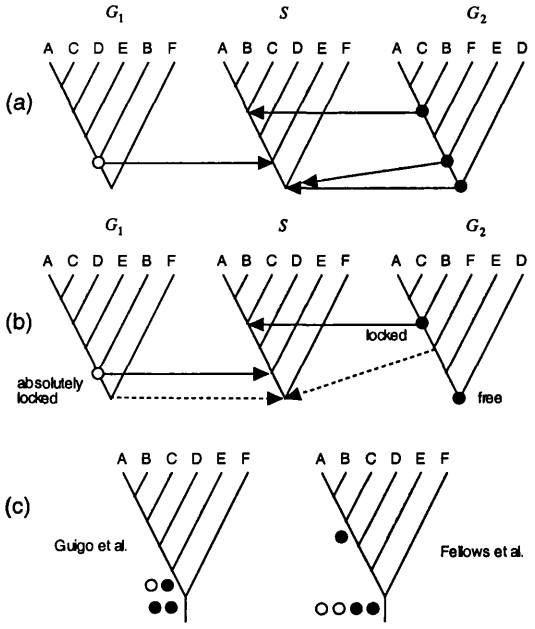

Figure B.2: (a)Two gene trees and their species tree with nodes mapped onto S. (b) Node ABCDE in G_1 is absolutely locked, whereas node ABC in G_2 is locked. (c) Comparison of how Guigó *et al.* (Guigó et al., 1996) and Fellows *et al.* (Fellows et al., 1998) would place the duplications on S to minimise the number of multiple gene duplications.

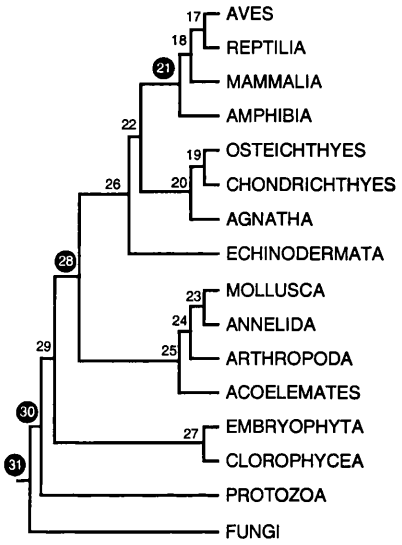

Figure B.3: A species tree for 16 eukaryotes from Guigó *et al.* (Guigó et al., 1996). Internal nodes are labelled $17 - 31$ in postorder. The locations of the "genome" duplications inferred by Guigó *et al.* (Guigó et al., 1996) are highlighted.

B.3 Placing duplications using set cover

We can reformulate Guigó *et al.*'s algorithm as a set cover problem. Let D be the set of all nodes $g \in G_i, i = 1, \ldots, k$ that are gene duplications. Each $s \in S$ has associated with it a set of duplications $D_s = \{d : d \in D, s \in A_d\}$. Finding the smallest number of locations at which gene duplication has taken place corresponds to finding the smallest number of sets such that their union is D. The set cover problem is *NP*-complete, but heuristics are available (Cormen et al., 1990).

We illustrate this approach using Guigó *et al.*'s data set. This has played an important role in developing methods of tree reconciliation. Previous work has shown that they miscount the number of gene losses (Page and Charleston, 1997) and that their species tree is not optimal for the 53 gene trees (Hallett and Lagergren, 2000; Page and Charleston, 1997).

The species tree shown in Figure B.3 requires 46 gene duplications, which are distributed over 7 nodes in the species tree:

$$D_{21} = \{2, 22, 36, 37, 44, 46\}$$

$D_{22} = \{8, 9, 13, 32, 33, 35 - 38, 44\}$

$D_{26} = \{8 - 9, 13, 32, 33, 35 - 38, 44\}$

$D_{28} = \{1, 4, 6, 8, 9, 13 - 17, 19, 20, 25, 26, 29, 32, 33, 35 - 38, 41\}$

$D_{29} = \{1, 6, 8, 9, 13 - 17, 19, 20, 24 - 26, 30, 32 - 38, 41\}$

$D_{30} = \{1, 7 - 9, 13 - 17, 19, 20, 24 - 26, 30, 32 - 38, 40, 42, 43\}$

$D_{31} = \{1, 3, 57 - 18, 21, 23 - 28, 31 - 39, 45\}$

The duplications are arbitrarily numbered $1 - 46$. The minimal set cover for D is $\{D_{21}, D_{28}, D_{30}, D_{31}\}$. These are the same four locations of the "genome" duplications identified by Guigó *et al.* (Figure B.3).

B.3.1 Final mapping

The minimal set cover might not yield an unambiguous mapping between the gene trees and the species tree; for example, duplication 36 is an element of all four sets in the minimal cover. This node occurs at the root of the gene tree for β-Nerve growth factor precursor (NGF) which has the topology (REPTILIA,(MAMMALIA, (AMPHIBIA,AVES))), and hence in Guigó *et al.*'s terminology is "free." Its set of allowable locations comprises vertex S_{21} and all its ancestors in the species tree (Figure B.3). Following Guigó *et al.*, any duplication g which occurs in more than one set in the minimal set cover is mapped onto the node closest to $M(g)$. This can be easily done as follows:

Step 1: Let F be a set of duplications. Initially $F \leftarrow \emptyset$.

Step 2: Process each node in S in postorder. For each node s for which $D_s \neq \emptyset$ go to Step 3.

Step 3: If $F = \emptyset$ then $F \leftarrow D_s$, otherwise $D_s \leftarrow D_s \setminus F$ and $F \leftarrow F \cup D_s$

The result of this procedure is a unique mapping from the gene trees into the species tree, consistent with the minimal set cover. Applying this to Guigó *et al.*'s data we obtain the following mapping, where duplications are labelled by the abbreviated gene family name from Guigó *et al.*'s table 2.

$D_{21} = \{$ACHG, GLUC, NGF, NGF, PAHO, TBB2, TPMA$\}$.

$D_{28} = \{$ACH2, ACT2, ACT3, ACTB, ANFC, COLI, CYLA, CYLA, CYLB, CYLB, G3P, G3P2, H2B, H2B, H4, HBA1, HBA2, PRVA, TBA1$\}$.

D_{30} = {ACT3, H2A3, H4, HMDH, TBA1, TBA1, TBB}.
D_{31} = {ACT, ACT2, ATPB, CATA, CISY, CYLH, G6PI, H2A2, H2B1, H31, H4, RLA2, TOP2}.

This mapping differs from that shown by Guigó *et al.* (their fig. 4), in that those authors assign one duplication in gene NGF to D_{28}, and one duplication of the genes CYLA, CYLB, and TBA1 to D_{30}. However, these placements violate Guigó *et al.*'s own rule that "free duplications are placed at the closest location preceding the node in which the duplication is mapped where a duplication – absolutely locked or locked –, if any, has already been placed" (Step 4 in section B.2.3 above).

B.3.2 Counting the number of episodes of gene duplication

If more than one duplication in a gene tree G is associated with the same node s in the species tree S (i.e., $|G \cap D_s| > 1$) then we may have to postulate multiple episodes of gene duplication occurring at s. For example, given two nodes g_1 and g_2 where g_1 is ancestral to g_2, if both nodes are in D_s then two duplication episodes are needed. However, if neither g_1 nor g_2 is ancestral to the other then both could be explained by the same event. Let the duplication height, $h(g)$, of a node $g \in G$ be the number of nodes along the path between g and the root of G for which are in D_s. Any duplication $g \in D_s$ with the same height can be explained by the same duplication event. Hence, the minimum number of distinct episodes of duplication at node s in gene family G is then $E_{(G,s)} = \mathrm{MAX}(h(g) : g \in G, g \in D_s) + 1$. The minimum number of episodes of duplication at node s across all k gene families is then $\mathrm{MAX}(E_{(G,s)} : G_1, \ldots, G_k)$.

For the Guigó *et al.* example, we require two episodes of gene duplication at D_{21}, D_{28}, and D_{30}, and one at D_{31}. This differs from their finding single duplications at all locations except D_{30}, where they postulate that a double duplication occurred. This difference stems from their misplacing the duplications for genes NCF, CYLA, CYLB, and TBA1 (see Sec. B.3.1).

B.3.3 Duplication patterns in vertebrates

The locations of the 1380 inferred gene duplications in our 118 gene family data set (Sec. B.1.2) were found using the above algorithm (Sec. B.3), showing that

they can be strongly clustered on the species tree (Fig. B.4). Many apparent duplications occur near the tips of the tree in the mouse and human lineages, but the bulk of these "duplications" actually represent multiple alleles at polymorphic loci, rather than gene duplications. Figure B.4 shows that substantial numbers of duplication events have occurred throughout vertebrate evolution, often affecting many gene families simultaneously. The largest single such event (duplicating 58 out of 118 families) occurred after the divergence of sharks and rays and prior to the divergence of teleosts and lobe fin fish. Gene duplication is clearly an important feature of vertebrate evolution, but the pattern shown in figure B.4 is more complex than that expected from the "2R hypothesis". Some gene families have undergone as many as 11 successive episodes of duplication, and at no point in vertebrate phylogeny can we explain all gene duplications that occurred at that time by a single genome-wide event.

B.4 Future directions

Further work on this problem is needed. There are two limitations of our algorithm that we are aware of. Our algorithm for the final mapping (Sec. B.3.1) minimises the number of location in the species tree at which gene duplications occur, but it does not guarantee to minimise the total number of episodes of gene duplication. It is possible to construct examples where spreading gene duplications across more locations will reduce the overall number of episodes of duplication.

Our algorithm uses only the topology of the tree, and hence may make erroneous placements of duplications. For example, Figure B.5 shows a gene tree for vertebrate adrenergic receptor $\alpha 1$ (ADRA1). The descendants of the duplication at node A are all mammalian sequences, hence a reconciled tree would place this duplication at the base of mammals. The set of allowed location for this duplication includes the common ancestor of mammals, and every node ancestral to that node that postdates the split between mammals and fish (equivalent to node B in Figure B.5) [4]. However, if we consider the branch lengths in the tree, node A is deeper

[4] This problem will be more prevalent in those gene families that have poorly sampled taxonomically, or have undergone substantial gene loss. Finding a single fish ADRA1 sequence related to either of the group 1 or group 2 mammal sequences would result in the method described here correctly inferring that node A pre-dates the split between fish and mammals.

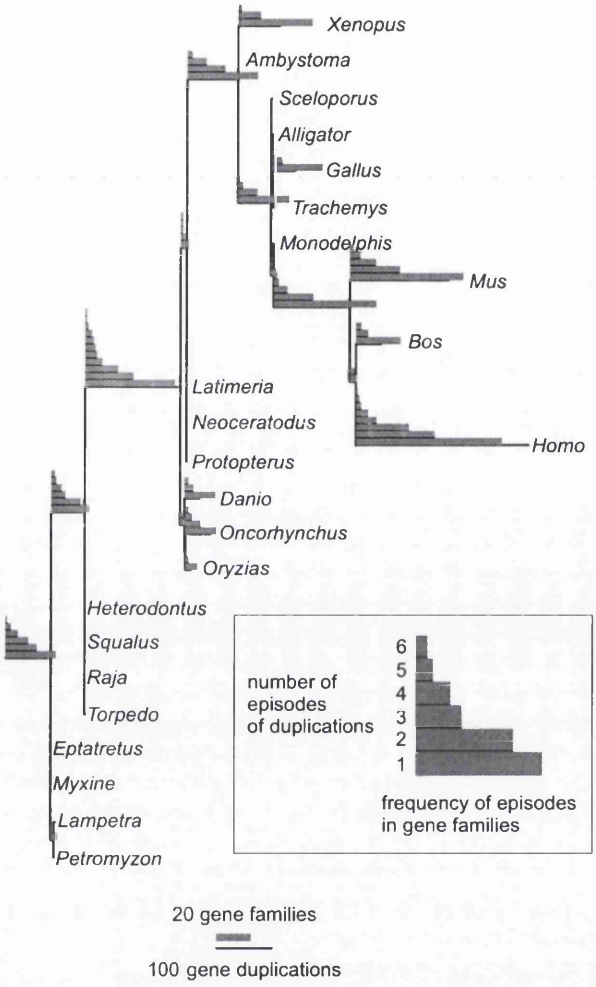

Figure B.4: Distribution of gene duplications during vertebrate evolution. The species tree is one of three most parsimonious trees from a GENETREE (Page, 1998) search. Branch lengths represent the number of separate gene duplications inferred to have occurred along each branch. Stacked bars represent the number of distinct episodes of gene duplication in each of the gene families that have duplicated along the branch. For clarity, bars have been omitted where only a single duplication episode is inferred for each gene family.

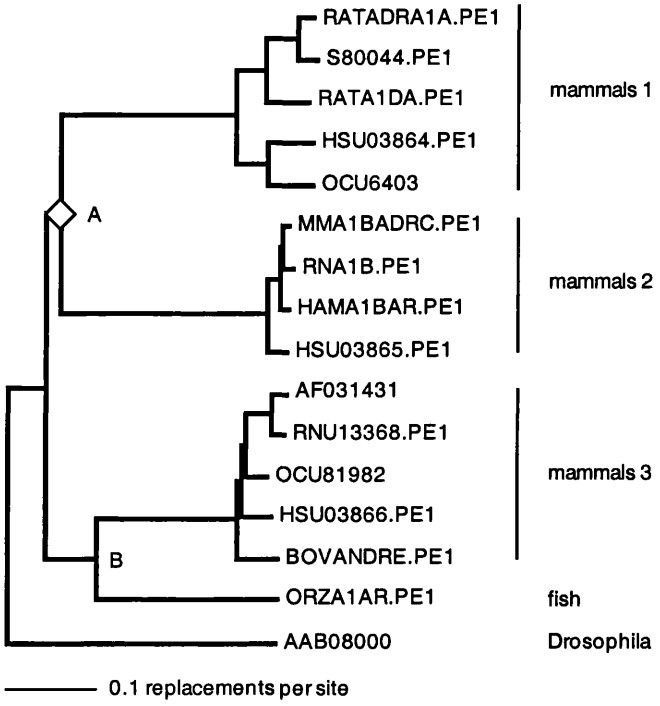

Figure B.5: Phylogeny for vertebrate adrenergic receptor α-1 sequences. The method for locating gene duplications described in this paper would place node A somewhere after the split of fish and mammals, but prior to the last common ancestor of mammals. Based on relative amount of sequence divergence with respect to node B (the split between fish and mammals), node A in fact pre dates the separation of fish from the ancestors of mammals. Data supplied by Xun Gu (Wang and Gu, 2000) . Sequence names are those used in the HOVERGEN database (Duret et al., 1994), in which ADRA1 is family FAM000048.

than node *B* in the tree and hence pre dates the oldest node in its allowed set of locations. One way to address this problem would be to refine the rules for determining sets of allowed location for gene duplications to take into account amounts of molecular sequence divergence (if they are sufficiently clock-like).

Bibliography

Benton, M. J. (1988). *The Phylogeny and Classification of the Tetrapods*. Clarendon Press, Oxford.

Cormen, T. H., Leiserson, C. E., and Rivest, R. L. (1990). *Introduction to algorithms*. MIT Press, Cambridge, MA.

Duret, L., Mouchiroud, D., and Gouy, M. (1994). *Nucleic Acids Res.*, 22:2360.

Fellows, M., Hallett, M., and Stege, U. (1998). In Kyung-Yong, C. and Ibarra, O. H., editors, *Proceedings of the 9th International Symposium on Algorithms and Computation*. Springer, Heidelberg.

Goodman, M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E., and Matsuda, G. (1979). *Syst. Zool.*, 28:132.

Guigó, R., Muchnik, I., and Smith, T. F. (1996). *Mol. Phylogenet. Evol.*, 6:270.

Hallett, M. T. and Lagergren, J. (2000). In *RECOMB '00, Proceedings of the fourth annual international conference on computational molecular biology*. Association for Computing Machinery.

Hughes, A. L. (1999). *J. Mol. Evol.*, 48:565.

Margush, T. and McMorris, F. R. (1981). *Bull. Math. Biol.*, 43:239.

Ohno, S. (1999). *Cell Devel. Biol.*, 10:517.

Page, R. D. M. (1994). *Syst. Biol.*, 48:53.

Page, R. D. M. (1998). *Bioinformatics*, 14:819.

Page, R. D. M. (2000). *Mol. Phylogent. Evol*, 14:89.

Page, R. D. M. and Charleston, M. A. (1997). In Mirkin, B., McMorris, F. R., Roberts, F. S., and Rzhetsky, A., editors, *Mathematical Hierarchies in Biology*. American Mathematical Society, Providence, RI.

Page, R. D. M. and Charleston, M. A. (1998). *Trends Ecol. Evol.*, 13:356.

Page, R. D. M. and Cotton, J. A. (2000). In *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics*. Kluwer Academic Publishers.

Skrabanek, L. and Wolfe, K. H. (1998). *Curr. Opin. Genet. Dev.*, 8:694.

Slowinski, J. B. and Page, R. D. M. (1999). *Syst. Biol.*, 48:81.

Wang, Y. and Gu, X. (2000). *J. Mol. Evol.*, 51:88.

Zardoya, R. and Meyer, A. (2001). In Ahlberg, P., editor, *Major Events in Early Vertebrate Evolution*. Taylor and Francis, London.

Zhang, L. (1997). *J. Comput. Biol.*, 4:177.