Wang, L.-S., and Warnow, T. (2001). Estimating true evolutionary distances between genomes. *In* ''Proceedings 33rd Ann. ACM Symp. Theory of Comput. (STOC'01),'' pp. 637–646. ACM Press, New York.

Wang, L.-S., and Warnow, T. (2004). Distance-based genome rearrangement phylogeny. *In* ''Mathematics of Evolution and Phylogeny'' (O. Gascuel, ed.). Oxford University Press.

[36] Analytical Methods for Detecting Paralogy in Molecular Datasets

Au_C36_1

By JAMES A. COTTON

Abstract

Paralogy (common ancestry through gene duplication rather than speciation) is widely recognized as an important problem for molecular systematists. This chapter introduces the concepts of paralogy and orthology and explains why paralogy can complicate both systematic work and other studies of molecular evolution. The definition of paralogy is explicitly phylogenetic, and phylogenetic methods are crucial in elucidating the pattern of paralogy. In particular, knowledge of the species phylogeny is key. I introduce the theory behind methods for detecting paralogy and briefly discuss two particular software implementations of phylogenetic methods to detect paralogy from molecular data. I also introduce a statistical method for detecting paralogy and some future directions for work on paralogy detection.

Introduction

What is Paralogy?

Since Darwin, for most biologists (or at least, for evolutionary biologists) *homology* has come to mean something like ''similarity due to common descent,'' to distinguish it from similarity due to convergent evolution. An accurate understanding of the relationships between living things depends on correctly identifying homologous characteristics of an organism from other similarity. A classic example would be the wings of bats and birds, which do not share a common evolutionary origin as wings. These wings and the limbs of other terrestrial vertebrates look very different but are truly homologous. Similarly, the genes of an organism can be homologues; genes share a common ancestor as features of an organism, and all mammal hemoglobins are descended from a hemoglobin gene

 0076-6879/05 $35.00

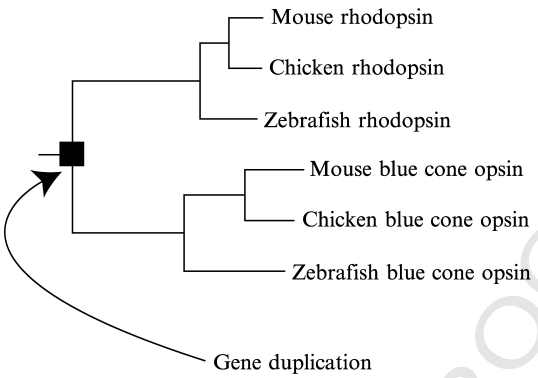

FIG. 1. Orthology and paralogy. The three rhodopsin genes are all orthologous to each other, as are the three blue cone opsin genes. The rhodopsin genes are paralogous to any of the cone opsins, and vice versa.

present in the ancestor of mammals, just as mammal limbs are descended from the limbs of this ancestor.

In genetics, however, homology can be a more complex phenomenon, because genes can be homologous in at least two distinct ways (Fitch, 2000). As well as descending from an ancestral species, genes also share a common ancestor as genes, in that related genes have arisen by duplication and gradual mutation. For example, all globin genes descend from a common ancestral globin gene. Fitch (1970) proposed new terms for these two classes of homology among genes. If the most recent common ancestor of two genes is a gene duplication event, the genes are *paralogous*, otherwise they are *orthologous* (Fig. 1). To use Fitch's original example, α and β hemoglobin are *paralogs*, whereas α hemoglobin in humans and mice are *orthologs*.

Why Does Paralogy Matter?

Fitch's introduction of the two terms makes it clear why the distinction between paralogs and orthologs is important: Only for orthologous genes does the ''history of the gene reflect the history of the species.'' An organismal phylogeny based on a mixture of paralogous genes would be ''biological nonsense'' (Fitch, 1970, p.113). This realization that inference of species relationships should be based on orthologous genes alone dates back to the earliest days of molecular systematics (Fitch and Margoliash, 1967). If, for example, we (unknowingly) sampled just the chicken and zebrafish rhodopsin genes and the mouse cone opsin gene from Fig. 1, we would mistakenly conclude that chickens and zebrafish were more closely

related to each other than either is to mice. We can think of the gene tree (a phylogeny constructed from some particular molecular data) and the species tree (the phylogeny representing the evolution of the organisms the genes have been sampled from) as distinct trees. The process of gene duplication is one of a number of reasons gene trees may not match the correct species tree (Maddison, 1997; Martin and Burg, 2002; Page, 1994).

This *problem of paralogy* is widely recognized by systematists (Sanderson and Shaffer, 2002), and many discussions of suitable genes for molecular phylogenetics suggest that the ideal molecular marker should be ''single copy'' (Cruickshank, 2002). The fear of paralogy has been one of the major reasons for the popularity of organelle genes (which are, perhaps wrongly, generally assumed to be single copy) and of ribosomal RNA genes (which are largely homogenized by gene conversion). This advice may, however, restrict systematists to relatively few loci, because most nuclear genes seem to be parts of families of related genes (Henikoff *et al*., 1997; Kunin *et al*., 2003; Slowinski and Page, 1999). Restricting work to these loci alone would mean rejecting the great possibilities opened up by genomic-level data becoming available for a widening range of organisms (Rokas *et al*., 2003). In any case, even when a gene is single copy in known genomes, it cannot be certain that the gene is single copy in all organisms and has been single copy throughout evolutionary history. Standard molecular systematic studies involving just polymerase chain reaction (PCR) amplification of a locus and sequencing of the product are not readily capable of detecting multiple copies of a gene in a sample. Unfortunately, relatively few concrete suggestions for dealing with the problem have been put forward; if potentially multicopy genes must be used, most authors suggest only a vague hope that paralogy might be recognized by differences in molecular architecture. These might include differences in intron structure or size, as well as changes in codon usage or base composition. Although such molecular approaches may help in recognizing paralogy in specific cases, it seems by no means inevitable that paralogous copies will show such differences.

Molecular biologists may have other reasons for wanting to detect paralogous genes. Gene duplication is probably the most important mechanism by which genes evolve new functions (Long and Thornton, 2001; Ohno, 1970), so that genes that are *orthologs* are more likely to share a common function than *paralogs*. Functional characterization of a gene is, thus, best made by comparison with orthologous sequences. Gene duplication may be the only common mechanism; it is hard to imagine how an arbitrary sequence can evolve a useful function (although see Hayashi *et al*., 2003), so most genes probably acquired their function by gradual evolution from a gene doing a related job. The role of gene duplication in this process

is easy to see: If a gene is performing an essential role in the cell, it is only when a duplicate copy exists to maintain this role that the gene is free to mutate away from the original function and evolve a new one. Of course, most mutations will reduce the gene's usefulness and many mutations will be silencing. A number of authors have envisaged this process as a ''race'' between a gene copy acquiring a new function and being silenced and eventually deleted from the genome (Walsh, 1995). Both empirical (Nadeau and Sankoff, 1997) and theoretical (Walsh, 1995) studies support the importance of gene duplication in the evolution of new gene functions.

Detecting paralogy may also be important in studying the pattern (and so the process) of gene duplication (Cotton, 2003; Page and Cotton, 2002). Empirical interest in the pattern of gene duplication has largely focused on testing the importance of polyploidy or genome duplication in evolution, looking at both phylogenetic and map-based data (Skrabanek and Wolfe, 1998), although a few papers have looked more widely at patterns of gene duplication (Lynch and Conery, 2000; Semple and Wolfe, 1999). To highlight the research interest in gene duplication, at least two major journals have published ''thematic issues'' focusing on evolution by gene duplication within a year (see introductions by Long [2003] and Meyer and Van de Peer [2003]).

Finally, recognizing paralogy is also important in molecular clock dating. If the estimated divergence date of two species is based on paralogous genes, then the event being dated is actually a gene duplication, rather than the speciation event, and the date estimate will be too old (Fig. 2). This could be a significant overestimate, depending on the rates of gene duplication and gene loss, and may be important in explaining at least some of the well-known discrepancies between molecular clock-based date estimates and dates estimated from the fossil record (Benton and Ayala, 2003), although other problems with both clock-based dates and the ways in which the fossil record has been used have also been described (Rodriguez-Trelles *et al*., 2002; Shaul and Graur, 2002).

Similarity (Homology) is not Orthology

The oldest way of identifying related genes is to identify genes that have similar sequences. Although systematists appear to be aware of the dangers of using sequence similarity alone to select genes for phylogenetic analysis (Bapteste *et al*., 2002), similarity has been used as a selection criterion in molecular clock-dating studies (Kumar and Hedges, 1998) and to identify genes that have similar functions (Eisen, 1998; Zmasek and Eddy, 2002). This similarity is most often detected by using BLAST or FASTA searches of sequence databases. This is particularly common in functional

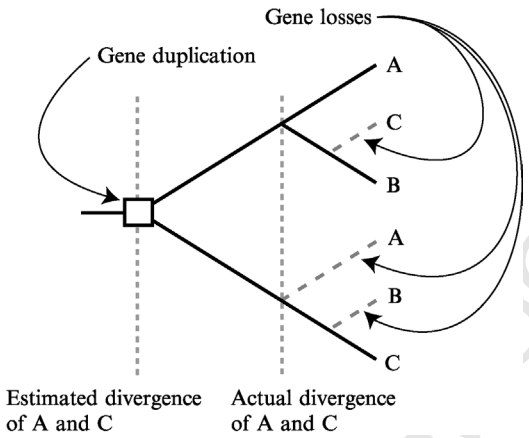

FIG. 2. How paralogy can alter estimates of species divergence dates. Gene duplications and subsequent gene loss will affect molecular estimates of divergence dates if the date of the gene duplication event rather than the actual speciation event is estimated. This will lead to overestimates of divergence dates.

annotation of genes and genomes, where sequence similarity methods are the standard approach. The use of sequence similarity in systematics seems set to become more important as large-scale phylogenetic analyses become possible by combining data from a range of sequencing projects (Rokas *et al*., 2003).

Against this background, it is important to point out that similarity is not the same as orthology, in the sense that orthologs of a particular gene may not be the most similar genes in a database. One major reason for this is the well-known fact that sequence similarity may not accurately reflect phylogenetic relationships, whether because of failure to correct for multiple substitutions at particular sites or because rates of evolution are unequal (Felsenstein, 2004, p.175) (Fig. 3). A different problem is that the database may not contain orthologs of the sequence, because of either being incomplete or because gene loss has led to the disappearance of these orthologs from extant genomes. This first problem is avoided if we build accurate phylogenies for the sequences involved.

Detecting Paralogy on Phylogenies

As is obvious from the definition discussed earlier, gene duplication events are key to understanding paralogy and orthology. A gene duplication event is one in which a piece of DNA is physically duplicated, forming

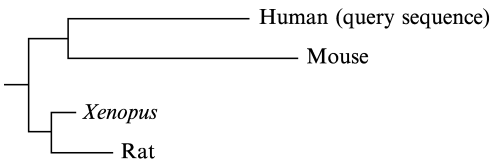

FIG. 3. Sequence similarity can be misleading. Unequal rates of evolution have led to the human query sequence being most similar to the *Xenopus* sequence rather than its ortholog (the mouse sequence). Methods based on sequence similarity alone will suggest that the human sequence is related to the subfamily containing the rat and *Xenopus* gene rather than to the mouse gene (modified, with permission, from Zmasek and Eddy [2002]).

a second copy of the genetic material. This can occur at a range of scales, from a few bases to the entire genome, representing a range of mechanisms from unequal crossing over and slippage during DNA replication to chromosomal non-disjunction and the production of unreduced gametes (Li and Graur, 1991, p.137). For our purposes, gene duplication must have occurred on a sufficiently large scale to have affected an entire locus that could be used for phylogenetic inference. Just as phylogenetic methods make assumptions about the process of nucleotide substitution, we would expect methods for detecting gene duplications to make some assumptions about the process of gene duplication, a point I will return to later.

The internal nodes on a phylogenetic tree represent divergence events, which, in molecular systematics, are usually presumed to be speciation. After a speciation event, the two lineages can no longer interbreed and are free to evolve independently and meet separate evolutionary fates, with the accumulation of mutations leading to them becoming gradually more and more distinct from one another. Gene duplications represent a similar event; after a gene duplication event, the two copies of a gene are free to accumulate independent mutations and diverge (at least in the absence of gene conversion). Gene duplication and speciation are similar splitting events and, in fact, cannot always be distinguished by simple inspection of a molecular phylogeny. If we accept that phylogenetic methods are needed to correctly identify paralogy and orthology, then the problem of detecting paralogy becomes that of identifying which internal nodes of a tree represent gene duplications and which represent speciation events.

Sometimes this can be easy. If two similar genes are present in the same genome, they must be paralogs or at least partial paralogs. Identifying paralogy on larger gene family phylogenies is similarly straightforward if there has been no loss of genes. Multiple copies are descended from a gene duplication, and the most parsimonious placement of the duplication is the least common ancestor (LCA) (the ancestor of both sequences that is

furthest away from the root, also called *most recent common ancestor* [MRCA]) of the duplicate copies. This placement will imply the smallest number of subsequent gene losses or deletions. When we have a phylogeny for the gene family, these nodes can then be discerned simply by inspection, and a number of studies have done exactly this (Katoh and Miyata, 2002). We can see how this is done by looking at Fig. 4D. The marked

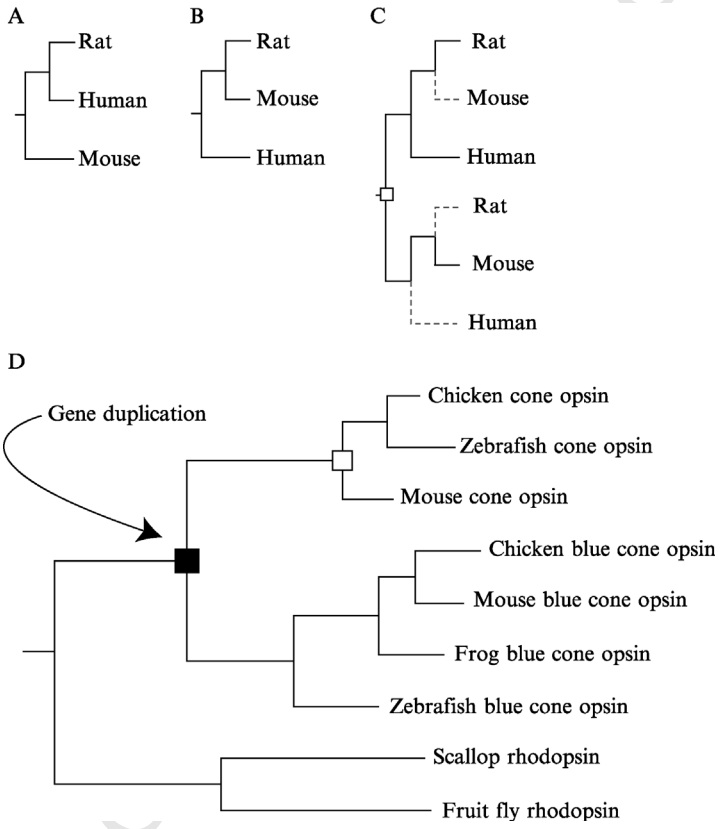

FIG. 4. (A, B, and C) The simplest possible case of gene duplication and gene loss obscuring the pattern of orthology and paralogy. Given a gene tree (A), it might seem that rat and human are more closely related to each other than either is to mouse. The correct species tree is shown (B). The incongruence between (A) and (B) can be explained by postulating a single gene duplication and three gene losses, shown on the reconciled tree (C). (D) A slightly more complex example. One gene duplication is implied by the presence of multiple sequences from mouse, chicken, and zebrafish (shaded box). A second gene duplication (open box) is implied only by the fact that the phylogeny for the top clade of cone opsins does not match the correct species phylogeny. The gene losses are not shown.

duplication event is the LCA of the two chicken opsins (and of the two mouse opsins and the two zebrafish opsins). For large, complex gene families (such as in Fig. 5), it is preferable to have some kind of computerized method for doing this, and LCAs can be found in linear time

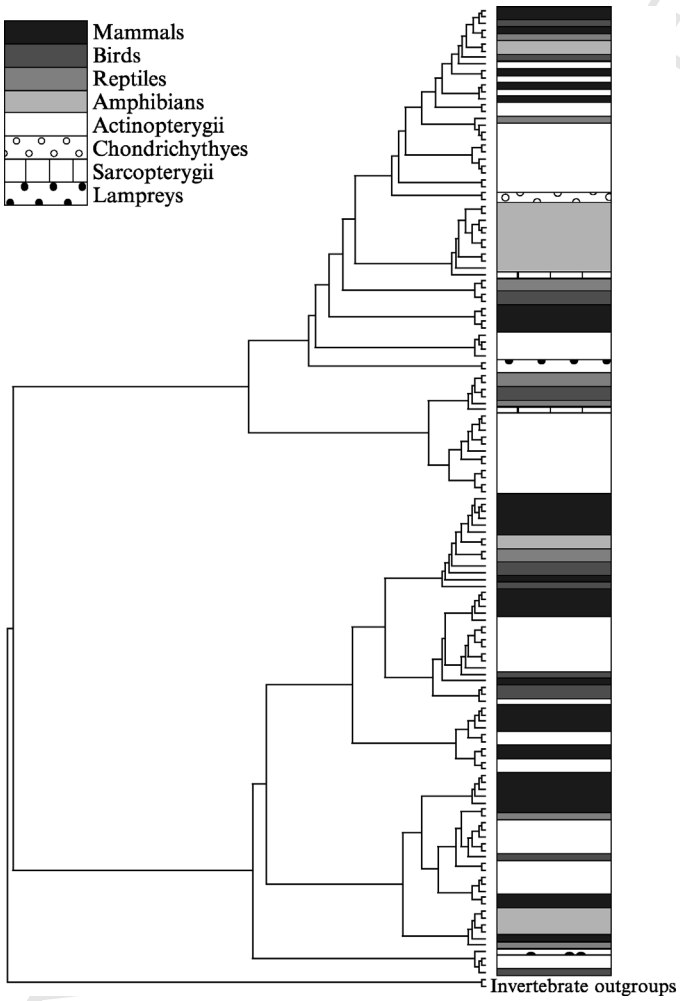

FIG. 5. Gene families can show complex orthology and paralogy relationships. This is the phylogeny for selected vertebrate opsin genes from Hovergen family HBG031788 (Duret *et al.*, 1994), color-coded to reflect the taxonomy of the species included. It is clear that there are a number of clades of related opsin genes and that the phylogeny within these clades does not always reflect the organism phylogeny. The pattern of orthology in this gene family is obscured by both gene loss and failure to sample; even some genomes that are fully sequenced lack certain orthologs.

(Harel and Tarjan, 1984), which means that the worst-case speed of finding the LCA scales as a linear function of the size of the input tree.

Importance of Knowing the Species Tree

The previous discussion ignores one crucial complicating factor: In the presence of gene loss or the absence of some gene copies from a phylogeny, recognizing the pattern of gene duplication (and so of orthology and paralogy) can be much more difficult. To understand this, a conceptually easy case to imagine is if one set of a duplicated pair of genes is lost, there will be no descendants to suggest that the gene duplication occurred at all. Of course, in this case, the survivors are all orthologous and the duplication will have had no effect on the phylogeny of the gene involved. This will not hold in other cases. If one of the two copies from a gene duplication is lost in each descendant lineage, there will be no multiple copies to suggest that a gene duplication occurred, but the two genes will be paralogous.

The most obvious symptom of these kinds of duplication is incongruence between a species tree and the gene tree, and it is this incongruence that allows gene duplication to be correctly inferred in the presence of gene loss, a realization that dates back at least to a seminal paper by Goodman *et al*. (1979). In the absence of any molecular events that introduce differences, we would expect the correct phylogeny for a set of gene sequences to exactly match the phylogeny for the species the genes have been sampled from. By fitting the observed gene tree into the known phylogeny for the species the genes have been sampled from, it is possible to infer evolutionary events, such as gene duplication and gene loss, which have introduced the differences between the two phylogenies.

Figure 4A–C shows this situation. Examining the tree in Fig. 4A by eye does not reveal any evidence of multiple gene lineages produced by a gene duplication event. Most biologists, however, will recognize that mice and rats are more closely related to each other than either is to humans, and the tree in Fig. 4A thus looks wrong. If we assume that the tree is, in fact, a correct estimate of the phylogeny for the gene, we can explain the difference between this tree and what we know to be the correct phylogeny for the three taxa in Fig. 4B as being due to a single gene duplication, followed by three gene losses (Fig. 4C). Even in more complex cases (Figs. 5 and 6), knowing the correct phylogeny for the species involved allows us to infer a scenario of duplications and losses that can explain incongruence between a gene and species tree. This idea of fitting a gene tree into a species tree has become known as *tree reconciliation*.

Au_C36_2

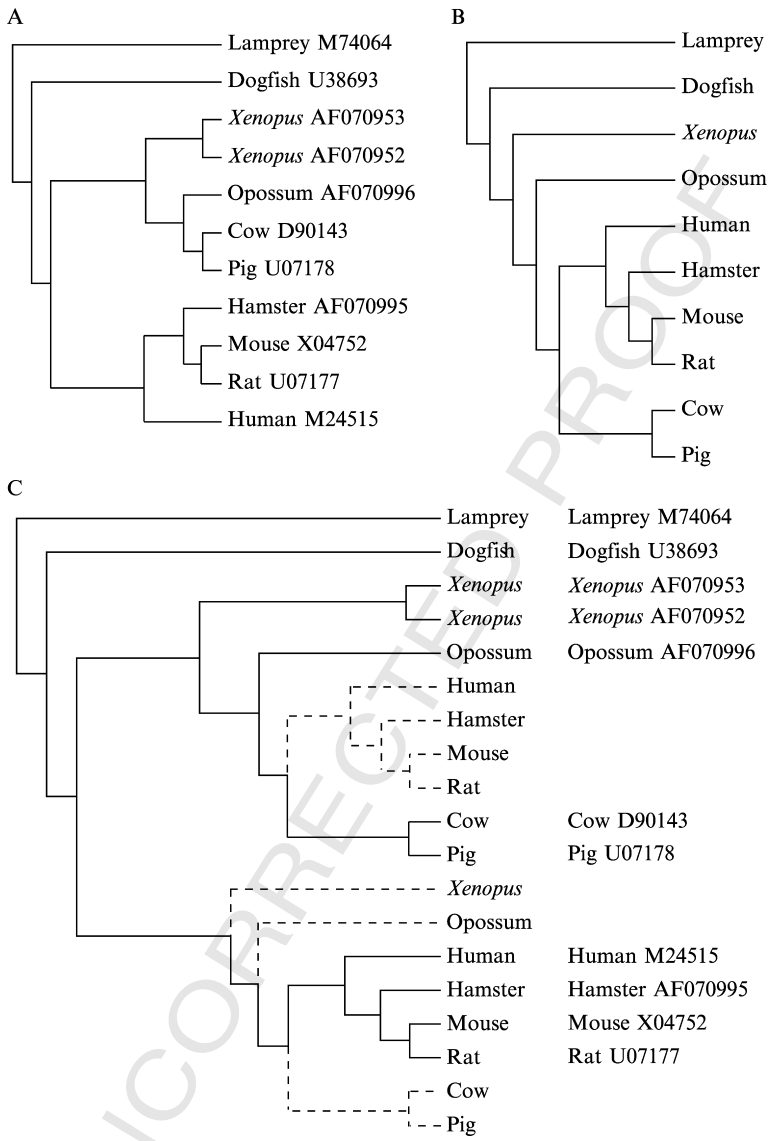

FIG. 6. (A) A molecular phylogeny for some vertebrate lactate dehydrogenase (LDH) genes. GenBank accession numbers are shown for each sequence. (B) The species tree for the organisms included on the tree. (C) A reconciled tree from GeneTree showing the evolutionary history of these genes. The reconciled tree identifies two gene duplications: one within *Xenopus* that is implied by the multiple copies in that species (open box) and another implied by the incongruence between the gene tree and species tree (shaded box). This second duplication correctly separates genes encoding the LDH-A muscle-specific and LDH-C testis-specific isozymes of the LDH enzyme.

Reconciled Trees: A Parsimony Method

Parsimony Mapping and Co-phylogeny

The problem of understanding the difference between two associated trees is a general one and has led to the idea of tree mapping or co-phylogeny. The associated trees can be any two trees that one would expect to be identical in the absence of some specific evolutionary event. If, for example, a parasitic organism has always speciated in response to host speciation, then they will have identical phylogenies, and host–parasite systems are probably the best-known example of associated phylogenies. Such systems can be studied by creating a map between the two trees to elucidate what events could have occurred to introduce any observed differences. For hosts and parasites, these events are things like host switching, in which a parasite population becomes established on a different host species, independent speciation of the parasite without a host speciation, and parasite extinction. In the gene tree–species tree system, the relevant events are lateral gene transfer (LGT), gene duplication, and gene loss. These events have similar phylogenetic effects to the host–parasite events; LGT is equivalent to host switching, gene duplication to independent parasite speciation, and gene loss to parasite extinction (Page and Charleston, 1997). The other commonly discussed system is the biogeographical system of an organism phylogeny and a hierarchy relating the areas the organisms inhabit (Page and Charleston, 1998).

The original concept of Goodman *et al*., that of producing a map between two associated trees in order to explain differences between them, has since been formalized by Page (1994), who presented the first algorithm for reconciling two trees. The algorithm is very simple, involving constructing a map between each node in the gene tree and each node in the species tree. The map is constructed traveling down the tree from leaves to the root. First, each leaf in the gene tree is mapped onto the corresponding leaf in the species tree. Any nonleaf node N in the gene tree is mapped onto the LCA of the species tree nodes onto which the descendants of N are mapped. When this map is completed, a gene duplication event is inferred wherever a gene tree node is mapped onto the same node as its immediate descendant. The number of gene losses can then be computed by another pass through the gene tree. In fact, a number of ways of speeding up this algorithm have been suggested, leading to two linear-time algorithms for reconciling two trees (Eulenstein, 1997; Zhang, 1997) and a simpler algorithm that has inferior worst-case running time but is claimed to be faster on most biological data (Zmasek and Eddy, 2001). The Eulenstein (1997) and Zmasek and Eddy (2001) algorithms are implemented in GeneTree

and RIO, respectively (discussed later in this chapter). In addition to counting gene duplications and gene losses, this map can be used to produce a reconciled tree that represents the evolution of the gene within the species phylogeny (Fig. 6C), allowing ready identification of paralogs and orthologs across the gene tree.

Gene Tree Is not Known without Error

Gene trees inferred by phylogenetic methods from amino acid or nucleotide sequence data are estimates of the true tree. They are unlikely to be estimates without error, whether from sampling error caused by the finite length of sequences used or because of the well-known biases in some phylogenetic methods (Felsenstein, 2004). The reconciled tree methods discussed earlier explicitly assume that the gene tree is known without error, as any incongruence between the gene tree and the species tree is explained in terms of gene duplication and gene loss. Clearly, if some of this incongruence is due to error in the gene tree, some of the implied duplications (and losses) will also be in error. A robust method for detecting paralogy will need to take some account of gene tree error.

A number of ways of dealing with this error have been proposed. One possibility is that an alternative gene tree, less parsimonious, less likely, or less probable than the optimal tree, is, in fact, the correct tree. A number of authors have suggested using the fit between the gene tree and species tree as a criterion for choosing between alternative gene tree topologies. Goodman *et al*. (1979) assigned each hemoglobin gene tree a score based on both the length of the tree in terms of nucleotide substitutions and the number of gene duplications and losses it implied on a species tree. They thus preferred less parsimonious hemoglobin trees that matched the expected species tree more closely. Fitch (1979) criticized this approach as requiring an arbitrary choice between the ''cost'' of a substitution event versus a duplication/loss event, although assigning these costs could be explored experimentally in specific circumstances (Ronquist, 2003).

One way to avoid this dilemma is to use some kind of statistical confidence interval around each gene tree, to contain all the gene trees that cannot be rejected by the sequence data (Martin, 2000). This could be a set of credible trees in the bayesian sense or all the trees inferred from bootstrapping the original sequence data, which would form a (rather conservative) pseudo–confidence interval for each gene tree estimate (Page, 1996; Sanderson, 1989). This kind of bootstrapping procedure has been suggested a number of times (Page and Cotton, 2000; Ronquist, 2003) and has been implemented in the programs GeneTree (Page, 1998), OrthoStrapper (Storm and Sonnhammer, 2002), and RIO (Zmasek and Eddy, 2002)

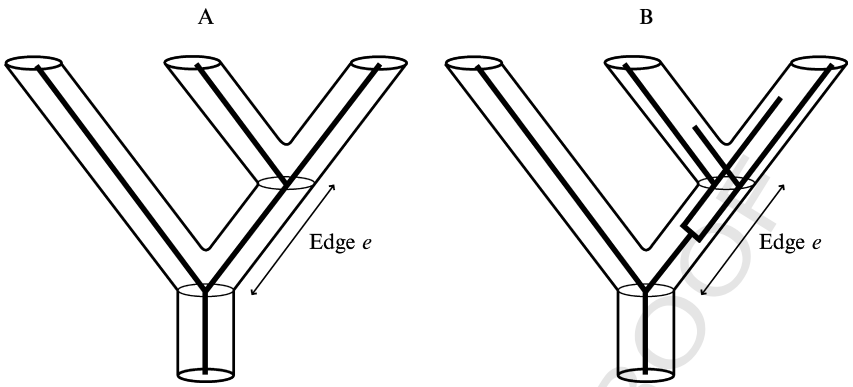

FIG. 7. A difference between parsimony and likelihood tree mapping. The two diagrams show a gene tree evolving within a species tree, where the species tree and gene tree match. In the parsimony case (A), no gene duplications will be inferred, whereas the likelihood method (B) takes into account cases in which gene duplications have occurred, followed by gene losses that reproduce the correct tree. The likelihood method should integrate across any number of duplications, from zero to infinity, along edge *e,* in calculating the probability that the two descendant lineages are orthologous.

(discussed later in this chapter). An obvious alternative to this bootstrapping approach is to use some kind of likelihood function that incorporates a model of sequence evolution and a model of gene duplication and loss. This has also been suggested previously (Page and Cotton, 2000), initially in the related context of allele coalescence (Maddison, 1997), and has led to the development of statistical methods. Finally, at least one other possibility has been explored—using local rearrangements (nearest-neighbor interchanges) (Waterman and Smith, 1978) around poorly supported nodes to make the gene tree better fit the species tree (Chen *et al.*, 2000; Page, 2000).

Detecting Paralogy (and Inferring a Species Tree) with GeneTree

A popular implementation of reconciled tree methods, and perhaps the easiest to use, is GeneTree (Page, 1998). Given a species tree and one or more gene trees, GeneTree will find the reconciled tree that represents the evolution of each gene tree, counting gene duplications and gene losses. It can also graphically show a reconciled tree for each gene family, allowing the user to see which genes are paralogs and orthologs (Fig. 6). GeneTree is a C++ program with a full graphical user interface (GUI), available for Mac OS and Microsoft Windows from

http://taxonomy.zoology.gla.ac.uk/rod/GeneTree/GeneTree.html. A cross-platform command line version is in development.

As the astute reader has probably noticed, detecting paralogy is of particular concern for molecular systematists. Understanding paralogy depends on knowing something about the species tree, so studies intended to elucidate the species tree for a little-known group will have no means of understanding paralogy in the molecular markers used. This is a potentially vicious circle; to get an estimate of a species phylogeny, we need to use orthologous sequences, but to accurately determine orthology, we need to know the species phylogeny accurately! The most common method of breaking this circle is simply to use markers that are considered to be free of paralogy, but other tactics may be available and may even be preferable.

GeneTree implements one approach; it can find the species tree that requires the minimum number of gene duplications to fit it onto the gene trees given. If gene duplications are thought to be sufficiently rare, the species tree minimizing the number of gene duplications (or the total number of gene duplications and gene losses) could be preferred as the best estimate of the species tree. Where gene trees are available from multiple independently evolving gene families, this approach may be particularly powerful and has been advocated as a general approach to molecular systematics (Slowinski and Page, 1999). Note that other approaches to inferring species phylogenies in the presence of paralogy have been proposed (Simmons *et al*., 2000), which may or may not be preferable to reconciled tree-based methods (Cotton and Page, 2003; Simmons and Freudenstein, 2002). In fact, this idea of searching for a tree that minimizes some cost, or distance, from a set of source trees, is one characterization of supertree methods (Thorley and Wilkinson, 2003), and the use of reconciled trees to infer a species tree (which has become known as *gene tree parsimony*) (Slowinski and Page, 1999) can be usefully compared with other supertree methods (Cotton and Page, in press).

Au_C36_3

Identifying Orthologous Genes Using RIO

Although GeneTree is aimed at molecular systematists, RIO is aimed at molecular biologists wanting to identify the functions of newly sequenced genes and, appropriately, takes a rather different approach. RIO (Zmasek and Eddy, 2002) is a suite of C and Java programs connected by a perl pipeline, specifically designed for the inference of orthology and paralogy from a set of sequence data. These programs together automate the entire process of ortholog and paralog identification.

RIO begins by identifying similar sequences in the Pfam protein family database and aligning these sequences using a hidden Markov model

approach, using the HMMER package. This alignment is then bootstrap resampled, and a phylogenetic tree constructed by neighbor joining on ML distances inferred under an empirical amino acid substitution matrix. Each of the bootstrap trees are then rooted to give a minimum number of duplications and then compared with a single species tree based on a number of large, published phylogenies to infer gene duplications and losses. These inferences can then be converted into percentage probabilities for orthology and paralogy between the query sequence and each related sequence identified in Pfam (see later discussion and Zmasek and Eddy [2002] for details of each step and references).

Zmasek and Eddy (2002) have recognized that if gene duplications are responsible for much of the origin of new gene functions, simple paralogy versus orthology may not be the only distinction of importance in functional annotation, leading them to introduce some new terminology. They define *superorthologs* as genes where not only is the LCA of two genes a speciation event rather than a gene duplication, but all the nodes on the shortest path connecting the two genes represent speciation events (this path connects the LCA to the two leaves). If gene duplications can lead to the evolution of new function, then superorthologs (which have undergone no gene duplication since their divergence) are most likely to share a common function. Zmasek and Eddy also introduce *ultra-paralogs,* which are genes for which the smallest subtree containing both genes contains only nodes that represent gene duplications. Such subtrees, which will contain sequences from a single species, represent lineage-specific expansion of a particular gene family. Lineage-specific duplication has been reported in a number of cases (e.g., in the lineage separating humans from the great apes [Nahon, 2003]) and seems to represent the selected growth of a functionally important gene family. Despite the large number of gene duplications, these genes share closely related functions because the newly formed gene copies appear to have partitioned the original function of the parental gene, rather than evolving completely new functions (so they are evolving by subfunctionalization rather than neofunctionalization) (Force *et al*., 1999). Finally, Zmasek and Eddy also introduced the term *subtree neighbors* to define gene copies that are present on the same clade of a certain size, presumably because more closely related genes may sometimes share the same function, whether they are paralogs or orthologs.

Perhaps the greatest strength of RIO is that it automates the entire analysis, performing a number of steps that the user of a program like GeneTree is required to do manually. RIO is designed explicitly for the molecular biologist interested in identifying the orthologs and paralogs of a particular query sequence. Because of its relative ease of use, an already available species tree and its attempt to further dissect paralogy and

orthology to make functional annotation more accurate, RIO is likely to be the first choice for this particular application. RIO is also available as a web service (from http://www.rio.wustl.edu), so users will not even need to install a local copy of the software (although RIO is available to download from http://www.genetics.wustl.edu/eddy/forester/). GeneTree may be of more interest to some other users, both because of its ability to infer an optimal species tree and because the reconciled tree allows the user to interpret paralogy and orthology across an entire gene family tree, rather than with respect to a particular sequence.

Other Implementations

Reconciled trees have also been implemented in a few other software packages, which I mention here for completeness. TreeMap (http://taxonomy.zoology.gla.ac.uk/∼mac/treemap/index.html) implements reconciled tree methods in the context of host–parasite evolution and can deal with LGT (host switching) and gene duplication and gene loss. DupLoss is a JAVA applet that implements an efficient fixed-parameter tractable algorithm for finding duplication and loss histories (Hallett and Lagergren, 2000) (available from http://www.sable.mcgill.ca/∼dbelan2/duploss/applet/duploss.html). The ATV tree editor (Zmasek and Eddy, 2001) also includes a simple algorithm for locating gene duplications and producing reconciled trees. Finally, OrthoParaMap (Cannon and Young, 2003) integrates both phylogenetic and genetic map data in an attempt to identify duplicated genes and divide them into regional and tandem duplications. This is interesting additional information, but the authors themselves admit that ''GeneTree and RIO generally appear to do a better job of identifying probable gene duplications and speciations'' than OrthoParaMap. In the example they give, OrthoParaMap worryingly appears to miss duplications that must have occurred, given the presences of multiple descendent copies in the same genome (Cannon and Young, 2003) (Fig. 4).

A Statistical Approach

All the parsimony methods for tree reconciliation share a number of problems. One problem is that a single parsimony mapping assumes that both the gene tree and the species tree are known without error, the problem that the bootstrapped analyses of RIO, OrthoStrapper, and GeneTree are designed to ameliorate. A closely related problem is that parsimony mapping is deterministic, so the reconciled tree algorithms produce only a single mapping and a single inference of the evolutionary history of a gene family. Again, using a bootstrap profile of input trees

avoids this problem, by providing a bootstrap-based confidence interval of possible reconstructions. A third problem with parsimony methods is that they make assumptions about the processes of gene duplication and gene loss; they implicitly assume that gene duplication and gene loss are both rare events, necessary conditions if minimizing these events is to give a realistic reconstruction of evolutionary history. This mirrors a similar assumption about nucleotide substitution made by parsimony methods for molecular sequences (Felsenstein, 1978). Although it seems likely that gene duplications and gene losses probably are rather rare, this assumption of parsimony leads to well-known undesirable properties of parsimony methods in phylogenetic reconstruction, not least statistical inconsistency in the phenomenon known as *long-branch attraction* (Felsenstein, 2004, pp. 113–122). It seems likely that similar problems will beset parsimony-based tree reconciliation.

A solution to these problems is to use a probabilistic model of gene duplication and gene loss. This is conceptually rather simple if we assume that both of these processes occur at a constant rate over time and independently of each other, a possibly dubious simplifying assumption that makes the models mathematically tractable. Such constant-rate Markov process models have been quite widely used to study the process of gene duplication and gene loss, as well as the related processes of speciation and extinction (Kubo and Isawa, 1995; Lynch and Conery, 2000; Nee *et al*., 1992). Under these models, it is relatively straightforward to calculate a probability for any pattern of gene duplications and gene losses in a single lineage. More complexity ensues when the gene lineage is evolving within a tree, as is needed when dealing with a gene family evolving inside a species phylogeny, but these calculations are certainly feasible (Arvestad *et al*., 2003). If we can calculate the probability of a particular reconstruction, given a gene tree and a species tree, this opens the possibility of both ML and bayesian methods for reconciliation.

At present, only one implementation of bayesian tree reconciliation has been reported (Arvestad *et al*., 2003), and the software for performing this reconciliation is not yet freely available. However, probabilistic methods will clearly be the preferred method for detecting paralogy. The work of Arvestad *et al*. takes trees produced by a bayesian phylogenetic method (such as MrBayes *et al*., 2001) and uses a Markov chain Monte Carlo (MCMC) method to estimate the posterior probability that each node represents a gene duplication (or conversely, that it represents a speciation) under the constant-rate model of duplication and loss. Arvestad *et al*. present a case in which this bayesian method clearly gives a more sensible result than the bootstrap approach as implemented in OrthoStrapper or RIO. This highlights the important difference between parsimony and

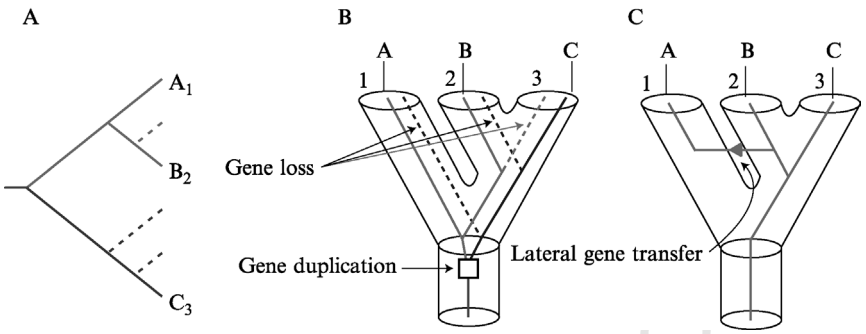

FIG. 8. Lateral gene transfer (LGT) (C) and gene duplication and subsequent loss (B) can have the same phylogenetic effect, introducing incongruence between the gene tree and species tree as shown on the reconciled tree (A). Genes 1, 2, and 3 are evolving inside species A, B, and C respectively.

probabilistic approach to reconciliation, even when uncertainty in the tree is incorporated in the parsimony framework.

The difference between parsimony-based and probabilistic paralogy detection is easily explained (Fig. 8). A parsimony-based method will assume the minimum possible number of gene duplications (as it assumes that gene duplication and gene loss are both rare) and so the minimum amount of paralogy. In the example in Fig. 8, parsimony methods will report no paralogy in gene tree ([A1,B1],C1), as it matches species tree ([A,B],C) exactly (Fig. 8A). In fact, there is a non-zero probability that gene duplications have occurred along any particular edge of non-zero length (such as the edge *e* in the figure), followed by subsequent gene loss in the descendant lineages leaving only a single copy, in such a way that has not affected the phylogeny of the gene (Fig. 8B). Similarly, a non-zero probability is attached to any number of possible duplications along edge *e*, followed by an appropriate number of later gene deletions. This translates into a non-zero probability that genes A_1 and B_1 are in fact paralogs. Note that bootstrapping will not deal with this problem adequately; the bootstrap probability that A_1 and B_1 are orthologous could still be 100% if the gene tree robustly supports their sister–group relationship.

This bayesian framework could be extended in a number of ways. Given that the species tree is rarely known without error, it should be possible to use MCMC to integrate across a probability distribution of species trees rather than a single estimate; such a distribution could, for example, come from analysis of some other gene or combination of genes that is thought to be more reliable than the gene family under investigation. Another possibility is a bayesian method for inference of a species tree

from a set of gene families in the presence of gene duplication and gene loss. This would be a bayesian analogue of gene tree parsimony. Such an approach is certainly technically feasible, although it would require assuming that different gene families are statistically independent estimates of the species phylogeny, which may not be the case for linked genes,

Concluding Remarks and Future Prospects

Lateral Gene Transfer

The reconciled tree methods discussed here are designed to deal correctly with gene duplication and gene loss, but they do not distinguish another form of homology, where genes have undergone LGT. This form of nonorthology has become known as *xenology* (Fitch, 2000; Gray and Fitch, 1983). LGT is certainly common among prokaryotes. Bacterial genomes are increasingly seen as dynamic mosaics of genes (Martin, 1999), with LGT considered to have ''had an extraordinary effect on bacterial genomes'' (Ochman, 2001). Although LGT is of great research interest in its own right, it is directly relevant to studying paralogy, as understanding the pattern of LGT is crucial in understanding the pattern of gene duplication and gene loss. The differences between a gene tree and a species tree introduced by LGT can be identical to those introduced by gene duplication and gene loss (Fig. 8). In many taxa, inferring the pattern of gene duplication and gene loss will thus depend on distinguishing these events from LGT.

Au_C36_4

The idea of reconciled trees has since been generalized to include potential events such as LGT and the equivalent host switching in the host–parasite setting (Page, 2003). Dealing correctly with this kind of event can become rather complex (Charleston, 1998) and makes the problem of correctly weighting different kinds of event even more difficult (see Ronquist [2003] for the most complete available discussion of this problem). At least two parsimony-based algorithms have been proposed to deal with host switching or LGT in a co-phylogenetic context, but there are problems with each. The Jungles algorithm (Charleston, 1998) is computationally intensive and thus too slow and memory hungry for many realistic problems, while the algorithm implemented in TreeFitter (Ronquist, 2003) does not seem to provide explicit reconstructions of the history of the lineage, making it useless for detecting paralogy. A bayesian method that correctly deals with one particular model of host switching has been proposed (Huelsenbeck *et al*., 1997), but this model assumes that only a single lineage is present in a species at any time, making it inapplicable in the context of gene family evolution within a species phylogeny. A

more promising algorithmic approach has been described (Hallett and Lagergren, 2001).

Just as methods for detecting duplication and loss need to take into account the confounding effect of LGT, so methods for studying LGT need to take gene duplication and gene loss into account. Existing methods for detecting LGT are widely seen as unsatisfactory (Eisen, 1998; Sicheritz-Pontén and Andersson, 2001) and the increasing amount of genome sequence data is particularly rich for microbes, where LGT is likely to be important. Developing co-phylogenetic methods and software, and particularly statistical methods, that deal with LGT, gene duplication and gene loss together is clearly an important avenue of research for the future.

Independence of Gene Duplications

One concern is that all of the methods described here assume that gene duplications are independent, both within and between gene families, but this is by no means sure to be the case. As gene duplication events can affect an any size piece of DNA from a few bases to the entire genome, a single event can introduce duplications on a number of gene families simultaneously and can introduce multiple duplications on a particular family. At the extreme, in a whole-genome duplication, all the extant members of the family will be duplicated. Such whole-genome duplications or polyploidization events may be rather common; they are certainly very widespread in flowering plants (Otto and Whitton, 2000) and have been recorded in many other lineages (Skrabanek and Wolfe, 1998), including vertebrates (Furlong and Holland, 2002; Page and Cotton, 2002; Taylor *et al.*, 2001). Methods that can deal with large-scale gene duplications may be more reliable in inferring paralogy and could help study these complex patterns of gene duplications. Reconstructing the pattern of large-scale gene duplications from phylogenetic data alone is computationally complex (Guigó *et al.*, 1996; Page and Cotton, 2002). It seems likely that methods integrating phylogenetic information with genetic map data, such as has been attempted with OrthoParaMap, will be needed to infer some events. Statistical models that relax the assumption of independence of duplications should also be possible to formulate, and inference under these models should be possible using MCMC. These complications might be modeled adequately by allowing the rate of gene duplication to vary somehow over the tree, but other complications might serve to make the probability model of duplication and loss more realistic. There is, for example, reason to believe that rates of duplication and loss may not be independent, as duplicate genes may be more likely to die earlier in their life than later (Force *et al.*, 1999; Walsh, 1995).

Conclusion

I have discussed two main reasons for wanting to detect paralogy: for molecular systematists to ensure they are correctly sampling the species tree, and for molecular biologists to improve the assignment of function. Paralogy detection might also be important for some of the growing number of molecular evolution studies that are being carried out at the scale of whole genomes (Wolfe and Li, 2003) and so involve loci beyond the few well-known gene families. Of course, recognizing paralogy is an essential part of understanding the pattern of gene duplication from phylogenetic data (Page and Cotton, 2002). If we are to make wider use of the enormous amount of phylogenetic information contained in nuclear gene families, methods for dealing with paralogy in molecular systematics will become widely needed. A species tree could be estimated from a large sample of loci without assuming orthology of all gene copies by using methods that explicitly deal with paralogy (Page and Cotton, 2000; Slowinski and Page, 1999), by using the potentially riskier strategy of hoping that the weight of evidence will overwhelm any error from paralogous sequences (Brower *et al*., 1996), or by using a method that is somewhere intermediate (Cotton and Page, 2003; Simmons *et al*., 2000). Which of these will be most popular and most successful remains unclear. In any event, better understanding of the processes of gene duplication, and particularly of gene loss, and in particular better quantitative data, together with statistical approaches to studying these processes, seems likely to have a considerable impact on methods for detecting paralogy.

Acknowledgments

Thanks to Trevor Cotton, Claire Pickthall, and Mark Wilkinson for constructive comments on the manuscript. I also thank Elizabeth Zimmer and Eric Roalson for the invitation to write this chapter. The author is supported by BBSRC grant no. 40/G18385.

References

Arvestad, L., Berglund, A.-C., Lagergren, J., and Sennblad, B. (2003). Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* **19,** 7–15.

Bapteste, E., Brinkmann, H., Lee, J. A., Moore, D. V., Sensen, C. W., Gordon, P., Durufle, L., Gaasterland, T., Lopez, P., Muller, M., and Philippe, H. (2002). The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium, Entamoeba,* and *Mastigamoeba. Proc. Natl. Acad. Sci. USA* **99,** 1414–1419.

Benton, M. J., and Ayala, F. J. (2003). Dating the tree of life. *Science* **300**.

Brower, A. V. Z., DeSalle, R., and Vogler, A. (1996). Gene trees, species trees and systematics: A cladistic perspective. *Annu. Rev. Ecol. Syst.* **27,** 423–450.

Cannon, S. B., and Young, N. D. (2003). OrthoParaMap: Distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies. *BMC Bioinformatics* **4,** 35.

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

Chen, K., Durand, D., and Farach-Colton, M. (2000). NOTUNG: A program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.* **7,** 429–447.

Cotton, J. A. (2003). ''Vertebrate phylogenomics and gene family evolution.'' Ph.D. Thesis, University of Glasgow, Glasgow, U.K. Available at: http://taxonomy.zoology.gla.ac.uk/~jcotton/thesis.htm.

Cotton, J. A., and Page, R. D. M. (2003). Gene tree parsimony vs. uninode coding for phylogenetic reconstruction. *Mol. Phylog. Evol.* **29,** 298–308.

Au_C36_5 Cotton, J. A., and Page, R. D. M. Tangled trees from molecular markers: Reconciling conflict between phylogenies to build molecular supertrees. ''Phylogenetic Supertrees: *In* Combining Information to Reveal the Tree of Life'' (O. R. P. Bininda-Emonds, ed.). Kluwer Academic, Dordrecht, The Netherlands. (in press).

Cruickshank, R. H. (2002). Molecular markers for the phylogenetics of mites and ticks. *Syst. Appl. Acarol.* **7,** 3–14.

Duret, L., Mouchiroud, D., and Gouy, M. (1994). HOVERGEN: A database of homologous vertebrate genes. *Nucleic Acids Res.* **22,** 2360–2365.

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

Eulenstein, O. (1997). ''A Linear Time Algorithm for Tree Mapping.'' St Augustine, Germany.

Felsenstein, J. (1978). A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biol. J. Linn. Soc.* **16,** 183–196.

Felsenstein, J. (2004). ''Inferring Phylogenies.'' Sinauer, Sunderland, MA.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.* **19,** 99–113.

Fitch, W. M. (1979). Cautionary remarks on using gene expression events in parsimony procedures. *Syst. Zool.* **28,** 375–379.

Fitch, W. M. (2000). Homology: A personal view on some of the problems. *Trends Genet.* **16,** 227–231.

Fitch, W. M., and Margoliash, E. (1967). Construction of phylogenetic trees. *Science* **155,** 279–284.

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.

Furlong, R. F., and Holland, P. W. H. (2002). Were vertebrates octoploid? *Philos. Trans. R. Soc. Lond. Series B* **357,** 531–544.

Goodman, M., Czelusniak, J., William-Moore, G., 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. *Syst. Zool.* **28,** 132–168.

Gray, G. S., and Fitch, W. M. (1983). Evolution of antibiotic resistance genes: The DNA sequence of a kanamycin resistance gene from *Staphylococcus aureus*. *Mol. Biol. Evol.* **1,** 57–66.

Guigó, R., Muchnik, I., and Smith, T. F. (1996). Reconstruction of ancient molecular phylogeny. *Mol. Phylog. Evol.* **6,** 189–213.

Hallett, M. T., and Lagergren, J. (2000). New algorithms for the duplication-loss model. *In* ''RECOMB '00, the Fourth Annual International Conference on Computational

Molecular Biology'' (R. Shamir, S. Miyano, S. Istrail, P. Pevzner, and M. Waterman, eds.), pp. 138–146. Association for Computing Machinery, New York.

Hallett, M. T., and Lagergren, J. (2001). Efficient algorithms for lateral gene transfer problems. *In* ''RECOMB '01, Proceedings of the Fifth Annual International Conference on Computational Molecular Biology'' (T. Lengauer, ed.), pp. 149–156. Association for Computing Machinery, New York.

Harel, D., and Tarjan, R. E. (1984). Fast algorithms for finding nearest common ancestors. *SIAM J. Comput.* **13,** 338–355.

Hayashi, Y., Sakata, H., Makino, Y., Urabe, I., and Yomo, T. (2003). Can an arbitrary sequence evolve towards acquiring a biological function? *J. Mol. Evol.* **56,** 162–168.

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

Huelsenbeck, J. P., Rannala, B., and Yang, Z. H. (1997). Statistical tests of host–parasite cospeciation. *Evolution* **51,** 410–419.

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

Katoh, K., and Miyata, T. (2002). Cyclostome hemoglobins are possibly paralogous to gnathostome hemoglobins. *J. Mol. Evol.* **55,** 246–249.

Kubo, T., and Isawa, 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., Lorenzo, V. D., and Ouzounis, C. A. (2003). Myriads of protein families, and still counting. *Genome Biol.* **4,** 401.

Li, W.-H., and Graur, D. (1991). ''Fundamentals of Molecular Evolution.'' Sinauer, Sunderland, MA.

Long, M. (2003). Preface. *Genetica* **118,** 97.

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

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

Maddison, W. P. (1997). Gene trees in species trees. *Syst. Biol.* **46,** 523–536.

Martin, A. P. (2000). Choosing among alternative trees of multi-gene families. *Mol. Phylog. Evol.* **16,** 430–439.

Martin, A. P., and Burg, T. M. (2002). Perils of paralogy: using HSP70 genes for inferring organismal phylogenies. *Syst. Biol.* **51,** 570–587.

Martin, W. (1999). Mosaic bacterial chromosomes: A challenge en route to a tree of genomes. *BioEssays* **21,** 99–104.

Meyer, A., and Van de Peer, Y. (2003). 'Natural selection merely modified while redundancy created'—Susumu Ohno's idea of the evolutionary importance of gene and genome duplications. *J. Struct. Funct. Genomics* **3,** vii–ix.

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.

Nahon, J.-L. (2003). Birth of ''human-specific'' genes during primate evolution. *Genetica* **118,** 193–208.

Nee, S., Mooers, A. Ø., and Harvey, P. H. (1992). Tempo and modes of evolution revealed from molecular phylogenies. *Proc. Natl. Acad. Sci. USA* **89,** 8322–8366.

Ochman, H. (2001). Lateral gene transfer and the nature of bacterial innovation. *Curr. Opin. Genet. Dev.* **11,** 616–619.

Ohno, S. (1970). ''Evolution by Gene Duplication.'' Springer-Verlag, Berlin.

Otto, S. P., and Whitton, J. (2000). Polyploid incidence and evolution. *Annu. Rev. Genet.* **34,** 401–437.

Page, R. D. M. (1994). Maps between trees and cladistic analysis of historical associations among genes, organisms and areas. *Syst. Biol.* **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 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. *Mol. Phylog. Evol.* **14,** 89–106.

Page, R. D. M. (2003). Introduction. *In* ''Tangled Trees: Phylogeny, Cospeciation and Coevolution'' (R. D. M. Page, ed.), pp. 1–21. University of Chicago Press.

Page, R. D. M., and Charleston, M. A. (1997). Reconciled trees and incongruent gene and species trees. *In* ''Mathematical Hierarchies in Biology'' (B. Mirkin, F. R. McMorris, F. S. Roberts, and A. Rzhetsky, eds.), pp. 57–71. American Mathematical Society, Providence, RI.

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

Page, R. D. M., and Cotton, J. A. (2000). GeneTree: A tool for exploring gene family evolution. *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.), pp. 525–536. Kluwer Academic Publishers, Dordrecht.

Page, R. D. M., and Cotton, J. A. (2002). Vertebrate phylogenomics: Reconciled trees and gene duplications. *In* ''Proceedings of the Pacific Symposium on Biocomputing 2002'' (R. B. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, eds.), pp. 536–547. World Scientific Publishing, Singapore.

Rodriguez-Trelles, F., Tarrio, R., and Ayala, F. J. (2002). A methodological bias toward overestimation of molecular evolutionary time scales. *Proc. Natl. Acad. Sci. USA* **99,** 8112–8115.

Rokas, A., Williams, B. L., King, N., and Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **42,** 798–804.

Ronquist, F. (2003). Parsimony analysis of coevolving species associations. *In* ''Tangled Trees: Phylogeny, Cospeciation and Coevolution'' (R. D. M. Page, ed.), pp. 22–64. University of Chicago Press, Chicago.

Sanderson, M. J. (1989). Confidence limits on phylogenies: the bootstrap revisited. *Cladistics* **5,** 113–129.

Sanderson, M. J., and Shaffer, H. B. (2002). Troubleshooting molecular phylogenetic analyses. *Annu. Rev. Ecol. Syst.* **33,** 49–72.

Semple, C., and Wolfe, K. H. (1999). Gene duplication and gene conversion in the *C. elegans* genome. *J. Mol. Evol.* **48,** 555–564.

Shaul, S., and Graur, D. (2002). Playing chicken (*Gallus gallus*): Methodological inconsistencies of molecular divergence date estimates due to secondary calibration points. *Gene* **300,** 59–61.

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

Simmons, M. P., Bailey, C. D., and Nixon, K. C. (2000). Phylogeny reconstruction using duplicate genes. *Mol. Biol. Evol.* **17,** 469–473.

Simmons, M. P., and Freudenstein, J. V. (2002). Uninode coding vs gene tree parsimony for phylogenetic reconstruction using duplicate genes. *Mol. Phylog. Evol.* **23**.

Skrabanek, L., and Wolfe, K. H. (1998). Eukaryotic genome duplication—where's the evidence? *Curr. Opin. Genet. Dev.* **8,** 694–700.

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

Storm, C. E. V., and Sonnhammer, E. L. L. (2002). Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* **18,** 92–99.

Taylor, J. S., Van de Peer, Y., Braasch, I., and Meyer, A. (2001). Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philos. Trans. R. Soc. Lond.* **356,** 1661–1679.

Thorley, J. L., and Wilkinson, M. (2003). A view of supertree methods. *In* ''Bioconsensus'' (M. F. Janowitz, F. J. Lapoine, F. R. McMorris, and F. S. Roberts, eds.), pp. 185–194. American Mathematical Society, Providence, RI.

Walsh, J. B. (1995). How often do duplicated genes evolve new functions? *Genetics* **139,** 421–428.

Waterman, M. S., and Smith, T. F. (1978). On the similarity of dendrograms. *J. Theor. Biol.* **73,** 789–800.

Wolfe, K. H., and Li, W.-H. (2003). Molecular evolution meets the genomics revolution. *Nat. Genet.* **33**(suppl), 255–265.

Zhang, L. (1997). On a Mirchkin-Muchnik-Smith conjecture for comparing molecular phylogenies. *J. Comput. Biol.* **4,** 177–187.

Zmasek, C., and Eddy, S. R. (2002). RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* **3,** 14.

Zmasek, C. M., and Eddy, S. R. (2001). ATV: Display and manipulation of annotated phylogenetic trees. *Bioinformatics* **17,** 383–384.

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.

[37] Analytical Methods for Studying the Evolution of Paralogs Using Duplicate Gene Datasets

By SARAH MATHEWS

Abstract

Gene duplication is widely viewed as an important source of raw material for functional innovation in proteins because at least some duplicate copies will evolve new or slightly modified functions. The study of the molecular processes by which functional innovation occurs interests both evolutionary biologists and protein chemists, and the development of methods to investigate these processes has led to a productive meeting of disciplines and an availability of complementary approaches for exploring datasets. This has resulted in insights into past events, prediction of current function, and prediction of future change. The methods fall broadly into two categories: those that rely on detection of shifts in selective constraints and those that rely on detection of correlations between molecular changes and functional shifts. Strengths and limitations of the methods

0076-6879/05 $35.00