Annu. Rev. Genomics Hum. Genet. 2000. 01:41–73

GENE FAMILY EVOLUTION AND HOMOLOGY: Genomics Meets Phylogenetics

Joseph W. Thornton[1] and Rob DeSalle[2]
[1]*Department of Biological Sciences and Center for Environmental Research and Conservation, Columbia University, New York, New York 10027; e-mail: jt121@columbia.edu;*
[2]*Division of Invertebrates, American Museum of Natural History, New York, New York 10024; e-mail: desalle@amnh.org*

Key Words orthology; gene duplication; exon shuffling; lateral gene transfer; concerted evolution; molecular evolution; maximum likelihood; parsimony; evolution of novelty

■ **Abstract** With the advent of high-throughput DNA sequencing and whole-genome analysis, it has become clear that the coding portions of the genome are organized hierarchically in gene families and superfamilies. Because the hierarchy of genes, like that of living organisms, reflects an ancient and continuing process of gene duplication and divergence, many of the conceptual and analytical tools used in phylogenetic systematics can and should be used in comparative genomics. Phylogenetic principles and techniques for assessing homology, inferring relationships among genes, and reconstructing evolutionary events provide a powerful way to interpret the ever increasing body of sequence data. In this review, we outline the application of phylogenetic approaches to comparative genomics, beginning with the inference of phylogeny and the assessment of gene orthology and paralogy. We also show how the phylogenetic approach makes possible novel kinds of comparative analysis, including detection of domain shuffling and lateral gene transfer, reconstruction of the evolutionary diversification of gene families, tracing of evolutionary change in protein function at the amino acid level, and prediction of structure-function relationships. A marriage of the principles of phylogenetic systematics with the copious data generated by genomics promises unprecedented insights into the nature of biological organization and the historical processes that created it.

A PHYLOGENETIC APPROACH TO GENE FAMILIES

Burgeoning DNA sequence data have made clear that the coding portions of the genome are organized hierarchically into families and superfamilies. (Traditionally, a gene family has been defined as a group of genes all of whose members have $>50\%$ pairwise amino acid similarity, and a superfamily as an alignable group of genes with similarity below this threshold (48); in this review, we use

the term "gene family" to encompass both types of groups.) Of the genes in the bacterium *Escherichia coli*, >50% are members of identified gene families (63), and the proportion of gene family members in eukaryotes may be in the same range or even higher (16, 108). The hierarchy of genes, like the nested organization of living organisms, has been produced primarily by processes of lineage splitting (gene duplication) and divergence (48, 89), so the concepts and analytical tools used in phylogenetic systematics are also applicable for reconstructing the evolutionary relationships among genes in genomes. Just as these techniques allow the overwhelming diversity of taxa in nature to be systematized into a concise and historically meaningful conceptual framework, they represent a powerful way to organize and interpret the ever-increasing body of gene sequence data.

Comparative biological analysis can be carried out only in the context of a phylogeny (49, 85). A sound classification of gene family relationships is therefore a prerequisite for virtually all types of inference about the evolution of genes and the proteins for which they code. With a reliable gene phylogeny in hand, we can predict the structure and function of uncharacterized proteins, infer the mechanisms by which new genes appeared and took on novel functions, reconstruct the biochemical pathways and gene complements of ancestral organisms, analyze coevolutionary relationships and dynamics among proteins, and understand links between genomic change and morphological innovation (63, 64).

Despite the power of phylogenetics for comparative analysis, a clear understanding of its principles has been lacking from most studies of genomes and gene families. Our purpose in this review is to present phylogenetic principles and techniques as they can be applied to issues in comparative genomics and to highlight concerns about several widely used approaches that conflict with these principles. We cannot hope to present all of the voluminous literature on gene family evolution and comparative genomics; instead, we cite those works that we believe exemplify the opportunities and hazards of the various modes of inference that are available to researchers in the field. In this section, we review the fundamentals, strengths, and weaknesses of the major approaches to tree building and evolutionary inference.

Parsimony

The goal of the phylogenetic approach to gene families is to recover the nested hierarchy of relationships among genes and test hypotheses about the evolutionary process, based on the hierarchical distribution of amino acid or nucleotide characters in DNA or protein sequences. Although phylogenetic methods—cladistic parsimony in particular—have not been dominant in the field of gene family studies and comparative genomics (but see 2, 17, 47, 92, 93, 104, 114, 124), their advantages vis-à-vis the more popular similarity-based (phenetic) approach are compelling. Since the 1980s, biological systematists have widely accepted the superiority of phylogenetic to phenetic methods, for both theoretical and practical reasons (29, 30, 58, 119); today, virtually no systematist would classify taxa by quantitative measures of pairwise similarity. The new field of comparative genomics should,

in our view, take account of the conceptual foundations, practical experience, and technical tools developed by systematics researchers over the last several decades.

The central assumption of phylogenetic systematics is no less valid for genes in a superfamily than it is for species in a genus: If genes have evolved by duplication and divergence from common ancestors, the genes will exist in a nested hierarchy of relatedness, and these relations will be manifest in a hierarchical distribution of shared derived characters (synapomorphies) in the gene sequences (30, 52). On this theoretical foundation, the most parsimonious gene family tree—the one with the fewest parallel and reverse character changes—is the phylogenetic hypothesis that best explains the distribution of shared character states as the result of common inheritance.

Parsimony methods can be computationally demanding. To find the most parsimonious tree, the number of amino acid or nucleotide changes required by every possible topology must be calculated. As the number of genes or taxa (T) in the analysis increases, the number of possible unrooted trees increases in faster-than-exponential fashion, according to the formula $N(T) = \sum_{i-3}^{T}(2i - 5)$ (33). With just 10 genes, the number of trees is $\sim$2 million, and it exceeds 8 trillion with 15 genes, so exhaustive searches are not possible for most gene families. Very efficient heuristic strategies have been developed, however, to evaluate huge numbers of topologies and explore tree space without becoming trapped in nonoptimal "islands" (88, 118). These algorithms, along with fast computers, have made parsimony analyses of hundreds of genes tractable, with reasonable confidence that the most parsimonious tree has been found (101, 113).

The major concern about parsimony methods is that they can be unreliable when applied to certain combinations of grossly unequal branch lengths. Sequences that have diverged greatly from each other, due to rapid evolutionary rates or very long periods of time, can become saturated with changes, resulting in similarity at some portion of sites by chance alone. When two sequences at the end of such "long branches" are combined with other sequences that are not saturated, shared character states produced by saturation may cause the first two sequences to group together, even if they are not closely related (32, 51). To avoid this problem, care must be taken to avoid anciently diverged sequences or characters that have not been subject to strong selection (such as third positions of protein-coding DNA sequences), to break up long branches with denser taxon sampling, and to use amino acid characters—which are less saturable than nucleotides—when long-branch attraction might be a concern. With the exception of this generally correctable problem, parsimony methods provide a phylogenetic technique that can be applied in a wide variety of circumstances with a minimum of assumptions.

Phenetics

Fundamentally different from the parsimony framework are phenetic approaches, which classify genes or proteins based on a single quantitative measure of pairwise similarity. These metrics represent the observed or corrected fraction of amino

acids or nucleotides that are identical between two aligned sequences. Trees are constructed by assuming that more similar genes shared a common ancestor more recently than less similar genes.

Methods of this type, such as the neighbor-joining, unweighted-pair-group (UPGMA), minimum-evolution, and Fitch-Margoliash techniques, have been dominant in studies of gene family phylogeny to date (examples include 4, 7–9, 11, 16, 18, 70, 106). Reliance on phenetic methods has become particularly acute as whole-genome sequences have become available: numerous computationally sophisticated informatics techniques, all based on phenetic criteria, have been implemented with the stated goal of recovering evolutionary and functional relationships among genes in genomes (16, 54, 64, 83). For example, the influential clusters of orthologous groups (COG) method for establishing gene orthology (121) and several recent proposals to predict protein function and interactions from whole-genome sequences (26, 76, 77, 97) all rely on pairwise similarity scores found in Basic Local Alignment Search Tool (BLAST; National Center for Biotechnology Information, Bethesda, MD) searches or other intergenomic comparisons.

Phenetic comparisons have the advantage of computational efficiency. Most such methods are algorithms for constructing a single tree rather than evaluating a large ensemble of possible topologies, using an optimality criterion, so they can rapidly produce a similarity-based tree from very large numbers of sequences. This is an important advantage when very large numbers of genes are being evaluated, as is often the case in comparative genomics.

But there are conditions under which phenetic approaches to tree building fail to recover evolutionary relationships, and these occur with some frequency in the evolution of gene families. Ohno's model of gene duplication predicts that new genes diverge rapidly after their duplication because of the relaxed selection pressures caused by functional redundancy (assuming that a higher "dose" of the gene product gives no selective advantage); if the copy takes on a new function, evolutionary rates are then expected to slow considerably as new selective constraints are imposed (89). An alternative model—the subfunctionalization hypothesis—proposes that, after duplication of a gene with multiple functions, both of the resulting paralogs diverge in sequence until the capacities of the ancestral gene product are gradually allocated between its descendants, at which time selection constrains further sequence change (39). Both views are consistent with the extreme sequence divergence of nonfunctional pseudogenes, the intermediate degree of divergence among paralogous genes with different functions, and the low divergence among orthologous sequences that have identical functions (74, 125). Whenever either of these processes holds, the following distance-based approaches will be inappropriate for gene family reconstruction:

1. UPGMA- and BLAST-based methods assume that divergence rates are identical in all lineages, which is often not the case in gene family evolution. For example, actin genes have evolved much more quickly in certain sea urchin lineages than in other taxa (61), and divergence rates

among paralogs in the nuclear receptor superfamily vary substantially (69, 124). When rates are variable, these methods will yield inaccurate phylogenies (74).

2. The pairwise similarity scores on which all phenetic techniques rely include not only phylogenetically informative synapomorphies but also shared ancestral characters (symplesiomorphies) and unique derived ones (autapomorphies). As a result, distances between closely related fast-evolving sequences—recently diverged paralogs, for instance—will be inflated by autapomorphies, and these methods will fail to recover these relationships. In turn, they will cluster slowly evolving sequences (such as anciently diverged orthologs with a conserved function) together, even when they are distantly related, because such sequences retain symplesiomorphies that reduce the pairwise distance between them (Figure 1; see color insert).
3. The neighbor-joining technique for tree construction and the minimum-evolution and Fitch-Margoliash methods for tree evaluation are less subject to distortion by unequal rates. They can recover evolutionary relationships, however, only when pairwise distances between genes are additive—that is, when distances between any pair of sequences are equal to the sum of the distances on the branches that connect them to their common ancestor. This assumption is often violated when sequences are subject to multiple changes at the same amino acid or nucleotide site. There are methods to correct for multiple changes, but they are not reliable when the frequency of multiple hits varies among sites or is higher in some lineages than in others (48). In gene families in which paralogs diverge at different rates, multiple hits are more likely in some genes than in others. Multiple hits are also more likely on the deep internal branches of a tree immediately after gene duplication events, when new paralogs explore sequence space more freely than they do after selection constrains their new or allocated functions more narrowly. And in any coding sequence, multiple hits are more likely at sites that are not critical to conserved aspects of function than at those subject to stronger selective constraints.
4. All phenetic methods require that distances between sequences be accurately calculated, a condition that can be difficult to satisfy when the differences between sequences are very large, sequences are short, or divergence rates vary substantially among sites in the sequence (74). All of these problems can occur in gene family reconstruction. Distances are often great, reflecting the fact that paralogous genes in many gene families have diverged considerably since their duplication hundreds of millions of years ago. These paralogs often contain only relatively short regions that are conserved enough to allow multiple alignment, making distance calculations subject to considerable error. And, as noted above, rates often vary considerably among positions in functional proteins.

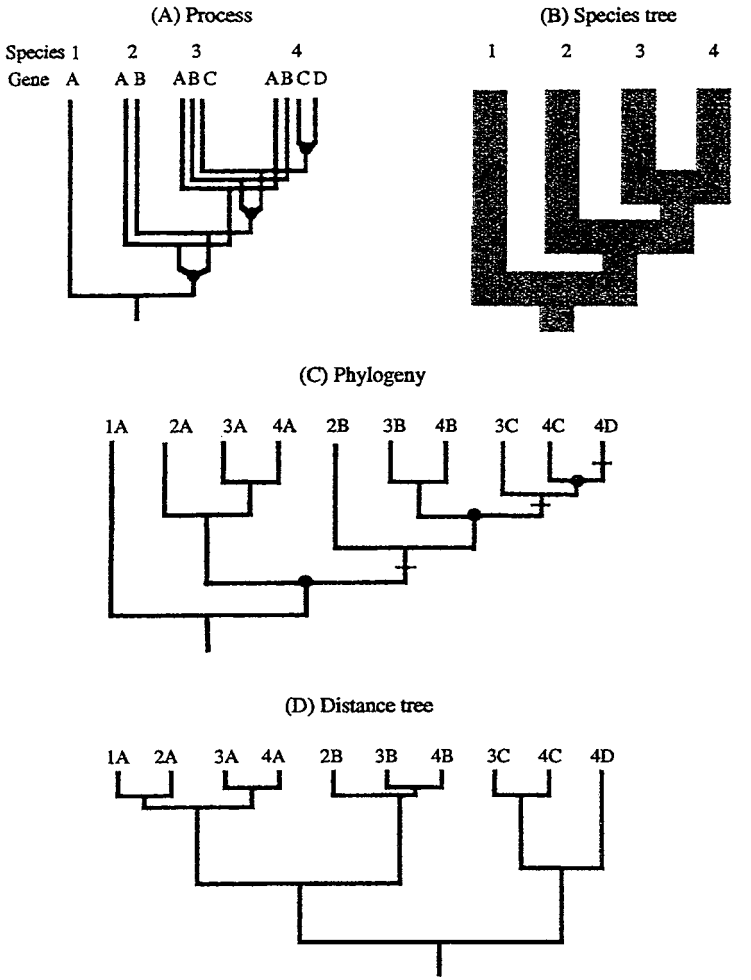


Figure 1 Cladistic and phenetic reconstructions of a gene family phylogeny. *A*. Evolutionary scenario of gene duplications (marked with *dark circles*) and cladogenesis (*unmarked nodes*) that generates a hypothetical gene family. Each group of colored branches leads to a group of similar orthologs. *B*. Species tree for the process in *A*. *C*. Gene phylogeny for the same process, correctly inferred by using the parsimony criterion. *D*. UPGMA tree for the process in *A*, assuming no homoplasy and 10-fold–higher rates of sequence change on branches on which gene duplications lead to new paralogous genes (marked with horizontal bars on the phylogeny in *C*) than on branches leading to conserved orthologs. This tree does not accurately represent evolutionary relationships.

For all of these reasons, pairwise distance methods cannot be relied on to accurately reconstruct evolutionary relationships among gene family members. Even under conditions that do not violate the assumptions of phenetic techniques, parsimony has two additional advantages over phenetics. First, distance methods collapse character information into a single quantitative measure of similarity. By preserving the information in individual amino acid or nucleic acid states, parsimony methods make possible a detailed examination of the processes by which molecular characters evolved and brought about novel aspects of protein function. Second, phenetic methods have a tendency to create a false sense of certainty. If the data in fact offer equal support to a number of topologies, those distance methods that are algorithms for tree construction rather than evaluation (neighbor joining and UPGMA in particular) will present just one phylogeny as the "true" tree. In contrast, the cladistic approach evaluates many trees by using the parsimony criterion, and it allows the degree of support for any phylogenetic hypothesis to be evaluated relative to others. When several topologies are equally supported, all can be presented as most parsimonious trees, avoiding the arbitrary resolution of phenetic methods.

Maximum Likelihood

Maximum likelihood (ML) is a third method for phylogenetic inference, the advantages and disadvantages of which continue to be debated (56, 109). This technique (34; reviewed in 82) selects the tree that, given an explicit model of sequence evolution, is most likely to have generated the sequence data observed. ML is useful because it is not subject to long-branch attraction, and it can take advantage of any generalizable knowledge about the patterns and dynamics of sequence evolution (119). ML algorithms are considerably more computationally demanding than even parsimony analyses, so reasonably thorough heuristic searches for the ML tree may become intractable before the number of orthologs and paralogs necessary for most gene family analyses is reached. Eventually, as computer speeds continue to increase, this limitation is likely to be overcome.

The reliability of ML methods depends on the realism of the evolutionary model; the tree that maximizes the likelihood of the data under an incorrect model of sequence evolution will not necessarily be the ML tree under a different and more accurate set of assumptions. Models of amino acid and codon evolution are not as well developed or validated as those for noncoding nucleotide sequences, and none adequately account for the nonindependence of sites in a protein or the fact that the probability of change from one type of amino acid to another is likely to be different and not necessarily predictable at different sites in the protein (see 82). Indeed, there are fundamental questions about whether any system that models protein evolution as a site-by-site probabilistic process can ever adequately capture the patterns produced when complex and nonlinear selection pressures act on three-dimensional protein conformations in ways that vary among sites and lineages. For example, the transformation frequencies that characterize the probability of

change from one amino acid to another are likely to be different at various sites in the protein, depending on whether side-chain volume, hydrophobicity, electrostatic potential, or ability to form disulfide bonds is the primary selective parameter at that site. Furthermore, the transformation frequencies at any one amino acid site are likely to be different for each paralog in a family, if paralogs bind to different ligands or cofactors, or if they display slight differences in folding that bring different residues into contact with each other (e.g. 120).

The degree to which likelihood methods are robust to these violations of their models' assumptions is unknown. The reliability of ML for reconstruction of relationships among coding sequences—particularly those in gene families—is thus currently in question, and cladistic parsimony remains for now the most useful and theoretically sound approach to inferring gene family phylogenies. As we discuss below, however, once a phylogeny is generated by using parsimony, the statistical nature of ML makes it a useful method for testing specific evolutionary hypotheses, such as those concerning dates of gene duplications or rates of sequence divergence (102).

HOMOLOGY, PARALOGY, AND ORTHOLOGY

Homology vs Similarity

Homology is the central concept in comparative and evolutionary biology (85). Meaningful biological comparisons must contrapose entities that are different versions of the same thing, and it is precisely this form of sameness that the term homology is intended to capture. Since Darwin, whether characteristics of organisms are "versions of the same thing" has been a matter of evolutionary history.

In the classic phylogenetic definition, homology means "derived from an equivalent characteristic of the common ancestor" (78). The vertebrae of mice and teleost fish are homologs because the two structures descended consistently from the vertebrae of their common ancestor >400 million years ago. Homology is the opposite of analogy, which describes the relationship among features that are similar because of convergent or parallel evolution rather than common descent: the wings of birds and of bats are analogous, because their common ancestor had no wings. Characters can therefore be similar without being homologous, and they can be homologous without being identical.

In 1987, an eminent group of biologists pointed out a fundamental difference between homology and similarity (100): sequences can be more or less similar, but homology is a strictly either-or proposition. Thus, if proteins X and Y have identical amino acids at 30 out of 40 aligned sites, we can say that they are 75% similar, but it is meaningless to say that they are 75% homologous. While most journals in systematics and evolutionary biology now attempt to maintain the correct terminology, our survey of three major molecular biology journals—*Cell*,

Development, and the *EMBO Journal*—for the most recent year indicates that this improper conflation of homology and pairwise percent similarity continues to be used in ∼50% of papers in which gene sequences are compared.

Homology says absolutely nothing about similarity of function. Unrelated proteins have been shown to converge to serve identical functions—a phenomenon called nonorthologous gene displacement—demonstrating that functional similarity can be analogous rather than homologous (62). Conversely, phylogenetically homologous proteins can diverge to serve subtly or grossly different purposes in different organisms, as is the case with the FtzF1-alpha gene product that regulates embryonic segmentation in *Drosophila melanogaster*, and its similar ortholog SF-1, which controls the expression of steroidogenic enzymes in vertebrates (124).

Orthologs and Paralogs

For sequence data, there are two major kinds of homology (96). Fitch defined orthologs as genes in different genomes that have been created by the splitting of taxonomic lineages, and paralogs as genes in the same genome created by gene duplication events (36). In the hypothetical case of Figure 1*c*, gene A in taxon 3 and gene A in taxon 4 are orthologs, whereas genes 4C and 4D are paralogs. These categories are analogous to the terms true homology and serial homology in morphological systematics, where the former refers to the same structure in two different organisms and the latter refers to structures within one individual that evolved by repetition of a single feature in an ancestral organism, such as segments, vertebrae, or limbs (96).

Distinguishing orthologous from paralogous genes is central to comparative genomics. It is only orthologs that can be said to be versions of the same gene in two different organisms, and mistaking a paralog for an ortholog is to follow a red herring in the genome. Indeed, the fundamental activity of comparative genomics is to track the presence, structural characteristics, function, and map position of orthologs in multiple genomes. Orthology identification must be accurate for these types of inference to be reliable.

There are fundamental problems with the ways that orthologs are currently identified in comparative genomics, which almost always involve finding the most similar pairs of genes between genomes based on pairwise similarity (16, 62, 63) The COG approach, for example, considers gene X from species 1 and gene Y from species 2 to be orthologs if X has a higher percentage of similarity to Y than to any other gene from species 2 and vice versa. Consider the relationship of 3C and 4C in Figure 1*c*. The COG framework would call these two genes orthologs, based on their close and unsurpassed sequence similarity. But homology is by definition a phylogenetic relationship, not a phenetic one. In a phylogenetic sense, 4C is no more closely related to 3C than 4D is; 4C and 4D are equally orthologous to 3C, as reflected in the common ancestry of 3C with 4C and with 4D at the same node on the tree. This problem affects the orthology not only of recent paralogs

but also of ancestral genes. In the COG framework, gene A in the stem species 1 would be considered an ortholog of the other As in the tree and a paralog of all other members of the gene family. In fact, 1A is equally related phylogenetically to every other member of the gene family in the analysis. This ambiguity remains unresolved no matter how similar the sequence of 1A is to the other As or how different it is from all the Bs, Cs, and Ds on the tree.

Orthology as defined in most comparative genomic frameworks is thus an inappropriately phenetic concept. Choosing the one most similar gene out of several phylogenetic orthologs is not unreasonable to make functional predictions; if gene 4C is very similar in sequence to gene 3C, but its paralog 4D (which is equally orthologous to 3C) has diverged considerably, then it is likely that 4C and 3C share a conserved function (63, 99, 121). But for reasoning about the evolutionary process, orthology based on phenetic similarity will lead to false conclusions, because other unrecognized orthologs may be lurking elsewhere in the genome. Phenetic orthology, for example, has been the foundation of most comparative mapping exercises (15, 84), but there is no reason that the conserved member of a duplicated pair—rather than the more divergent one—must occupy the same map position as the ancestral gene. Similarly, the presence of phenetic orthologs in pairs of distantly related organisms and their absence from more closely related ones has been used to infer lateral gene transfer among taxonomic lineages (6, 86), but the presence of unrecognized orthologs has the potential to explain these patterns without invoking horizontal transfer. In addition, reports that attempt to reconstruct the minimal protein sets of ancestral organisms based on the presence or absence of phenetic orthologs in descendant species (62, 66) will omit true members of that set whenever less similar orthologs are not recognized. Genes that do not form monophyletic groups of orthologs but are closest to each other in a phenetic sense should be called phenologs, not orthologs, and phenology should not be mistaken for true orthology in the evolutionary sense.

Homology as Hypothesis

In a phylogenetic context, a statement that two features or genes are homologous is not an observed fact but a hypothesis about the evolution of characters, which must be evaluated in the context of a phylogenetic tree (1, 96, 100, 122). The process of assessing homology for morphological features has multiple stages (12, 20). First, a hypothesis of homology (a primary homology statement) is formulated based on such criteria as topographical location on the organism and similarity of the character state. Second, information from all available characters is used to infer a phylogeny and evaluate whether the character is analogous or homologous (a secondary homology statement). This second stage requires more taxa (or genes) and characters than the ones being evaluated for homology, because the test of the homology hypothesis is based on the congruence of this feature with a body of other phylogenetically informative characters. At minimum, we require the two taxa or sequences that contain the character being evaluated, an outgroup to polarize the

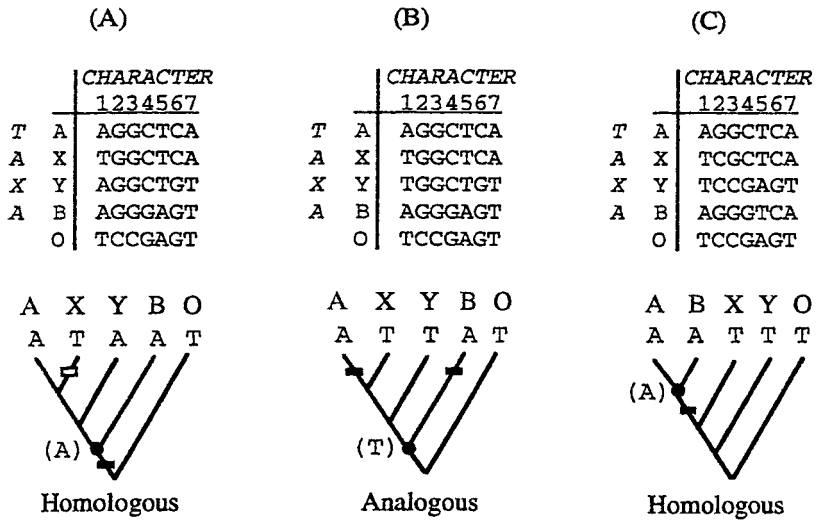


Figure 2 Homology assessment must include more characters and taxa than those being tested. By definition, the hypothesis that state *A* for character number 1 (*red*) is homologous in species *A* and *B* (*blue*) implies that the common ancestor of *A* and *B* (*blue circle*) had state *A*. Testing this hypothesis requires enough taxa to support the reconstruction that the ancestor did not have state *A* under at least some combinations of character states and enough characters to resolve the phylogeny of these taxa. Each tree shows the most parsimonious phylogeny for the data given, with the most parsimonious reconstruction of state changes for character 1. Gain of the state *A* is represented as a *filled box* and losses as *open boxes*. Trees *A* and *B* show that the outcome of homology assessment for character 1 in species *A* and *B* depends on the state of that character in other taxa. Trees *B* and *C* show that the outcome depends on the phylogeny inferred from all available characters.

characters, and two intervening taxa to determine whether the common ancestor had the feature in question (Figure 2, see color insert).

Even if the primary homology statement is corroborated by this test, the secondary homology statement remains a testable and potentially refutable hypothesis, not a fact. Data on the same character in other taxa may later establish that the common ancestor did not have the shared feature. Alternatively, additional character data could revise the topology of the tree, which could have the same effect. The more taxa used in the analysis, the stronger the test. A phylogeny with only birds, bats, frogs, and fish would suggest—very weakly—that the wings of birds and bats are homologous. But adding a larger number of intervening taxa, such as other mammals, crocodiles, and dinosaurs, would strongly refute that hypothesis.

In the context of molecular sequences, the first step in assessing homology is sequence alignment. This procedure establishes the topographical identity of individual nucleotide or amino acid sites. The similarity of nucleotide or amino

acid characters at aligned sites is then examined, and a primary-homology statement can be made. Whether identical character states at any site in a sequence are homologous or analogous is then tested on a phylogeny, which must be inferred using all available sites and sequences (12). As in morphological homology assessment, the more extensive the taxon sampling, the more decisive the test (1).

In comparative genomics, we are interested not only in the homology of individual amino acids or nucleotides within a sequence but also in whether the genes themselves are homologous and, if so, whether they are orthologs or paralogs. That is, we are interested in the genes not as vessels or lineages that contain evolving molecular characters but in the presence or absence of the genes as characters within genomes. Evaluating the orthology and paralogy of genes in this sense requires a slightly different two-stage procedure. The first step again consists of alignment as a prerequisite to phylogenetic inference. Given the vast number of combinations of nucleotides or amino acids in sequences with >10–20 sites, the ability to align sequences at all is a priori evidence that the sequences descended from a common ancestral gene. The second stage—assessment of the type of homology—requires identifying whether two related genes are descended from a gene duplication event or the splitting of taxonomic lineages. As in the tracing of molecular characters, such an assessment can be made only by using a phylogeny that represents the relationship of the two genes in the context of other paralogously and orthologously related sequences in the same family. As we detail below, orthology and paralogy are judged by the congruence of the relationships among genes in a gene family with the expected relationship based on a well-corroborated phylogeny of the taxa that carry them.

Orthology statements based on similarity alone without a phylogenetic analysis—such as COGs—should therefore be seen as primary hypotheses of orthology and no more. They are hypotheses to be explored further, but until they have been phylogenetically tested they remain without empirical support and are likely to yield incomplete and sometimes erroneous representations of homology relationships. As we have seen, orthology statements from COGs may be inaccurate due to the ambiguities of phenetic algorithms; they may also incorrectly assign a gene to an orthology group if the true ortholog in that species has been lost during evolution (121). Both of these problems will be revealed when the hypothesis of orthology is examined in the context of a gene family phylogeny, allowing false or incomplete orthology statements to be refuted or refined.

Several research groups have attempted to develop methods to identify genes as paralogs based on the three-dimensional structure of their products, even when their sequences are too divergent to be aligned (54, 81). But homology can never be established based on a single character; if sequences cannot be aligned, there are no other data on which to infer a phylogeny, and there is no way to test the hypothesis that similarity is due to homology rather than analogy. Assertions of gene orthology or paralogy based on protein structure alone are thus not only untested but thoroughly untestable statements with no possibility of empirical support.

BUILDING GENE FAMILY TREES

Techniques for analyzing gene family relationships with parsimony are almost identical to those used for inferring taxonomic relationships from gene sequences. First, the sequences to be analyzed must be selected. In principle, nucleotide or amino acid sequences—or both (2)—may be used; in practice, amino acids are more often analyzed, because their signal-to-noise ratio is usually more appropriate for analyzing gene families that diversified hundreds of millions of years ago. Members of a given gene family in the published databases can be obtained by using BLAST searches (3); the position-specific iterated BLAST approach, available on-line from the National Center for Biotechnology Information, is particularly useful for finding distantly related members of a family that may be missed by single-pass similarity searches (available on-line at www.ncbi.nlm.nih.gov). For large, well-studied gene families there are often hundreds of sequences available; to use all would be extremely demanding of time and computer resources. In these cases, it is necessary to use only some of the many orthologous sequences, and selection should be made to ensure broad taxonomic sampling. When orthologs are highly conserved among closely related species, as in many gene families, little phylogenetic information is lost by choosing sequences from just one species in a taxonomic order (e.g. rodents, primates, cichlids, etc).

Sequences must then be aligned to produce a data matrix. Numerous methods and programs for sequence alignment are available; the most theoretically justifiable are those that perform multiple alignments in a phylogenetic context, such as Clustal (123), or the parsimony-based TreeAlign (50) or Malign (127). All alignment methods should be used with attention to the sensitivity of alignments to user-specified gap-change ratios and other parameters. Paralogous groups of genes in a family often significantly diverge from each other, and the alignment of less constrained regions often varies with the parameters. One technique to avoid the arbitrary selection of one of many plausible alignments is the "cull" procedure, in which alignment-ambiguous positions are removed (41). These sites often contain useful phylogenetic information, so their omission may reduce phylogenetic resolution and/or support. To avoid this problem, the "elision" procedure can be used to assemble numerous plausible alignments in a master data matrix—an approach that effectively gives higher weight to alignment-consistent positions without losing the information present in alignment-ambiguous sites (126).

Phylogenetic analysis of the aligned sequences proceeds as it does with organismal phylogeny. The diversity of most gene families usually requires heuristic search strategies, such as those available in Phylogenetic Analysis Using Parsimony [PAUP* (118)], to seek the most parsimonious tree(s) (MPT). Confidence in the phylogenetic relationships represented in the MPT can be evaluated by calculating Bremer supports, which express the relative character support for each node as the number of extra steps required for each node not to appear in the MPT (10). Bremer supports can be calculated automatically using

Auto-Decay (27). Bootstrapping—the assembly of a new data matrix that is the same size as the original by randomly sampling sequence positions from the original matrix, with replacement—is also frequently used to assess confidence in a phylogeny (35). Because this approach in essence measures the effect of random weighting of characters, it reveals only the degree to which phylogenetic signal is uniformly distributed throughout the data set, not statistical confidence in the MPT's nodes or the extent to which the data support that tree. In gene families, it is not uncommon for many sequence sites to be completely conserved and others to be highly diverged, with only a portion of sites exhibiting the intermediate degree of variability that makes them phylogenetically informative. Further, the sites that are informative at deep levels of the tree are often different from those that support resolution nearer the tips. High bootstrap values for many nodes in a gene family tree are thus not expected, even when there is substantial support for the MPT. (For other problems with bootstrapping as a measure of confidence, see 103.)

INTERPRETING TREE TOPOLOGY

Embedded Trees

A gene family phylogeny is less straightforward to interpret than an organismal tree. An accurate tree tracks recency of common ancestry among sequences in a gene family, but splitting and divergence of gene lineages can be caused either by duplication of genes within a genome (producing paralogs) or by the splitting of the taxonomic lineage that carries those genes (producing orthologs). When a gene family tree includes sequences from taxa whose most recent common ancestors existed before some but not all of the duplication events that produced the paralogs in the family, then orthologs and paralogs will be interleaved on the phylogeny in a complex pattern that hierarchically reflects the order in which gene duplication and cladogenetic events occurred.

Consider the example shown in Figure 1. The evolutionary process depicted in Figure 1*a* will yield the species tree in Figure 1*b* and the gene phylogeny in Figure 1*c*, if sequences are available for all relevant genes from all the taxa used in this analysis. Each branch of the tree that diverges from a node representing a gene duplication contains a replica of that part of the taxonomic tree produced by cladogenic events that occurred after the gene duplication. If all of the gene duplication events happened before any of the taxa on the tree diverged from each other, then each branch leading to a group of orthologs will contain the entire species tree. If some duplications occurred before and some after the relevant cladogenic events, then subtrees of various sizes, one for each paralog, will be arranged in nested fashion in the gene tree (Figure 1*c*). This master phylogeny of species trees within paralog trees is analogous to other kinds of trees in which lineages duplicate at more than one level, such as biogeographic area cladograms and phylogenies of parasites from multiple host taxa (92).

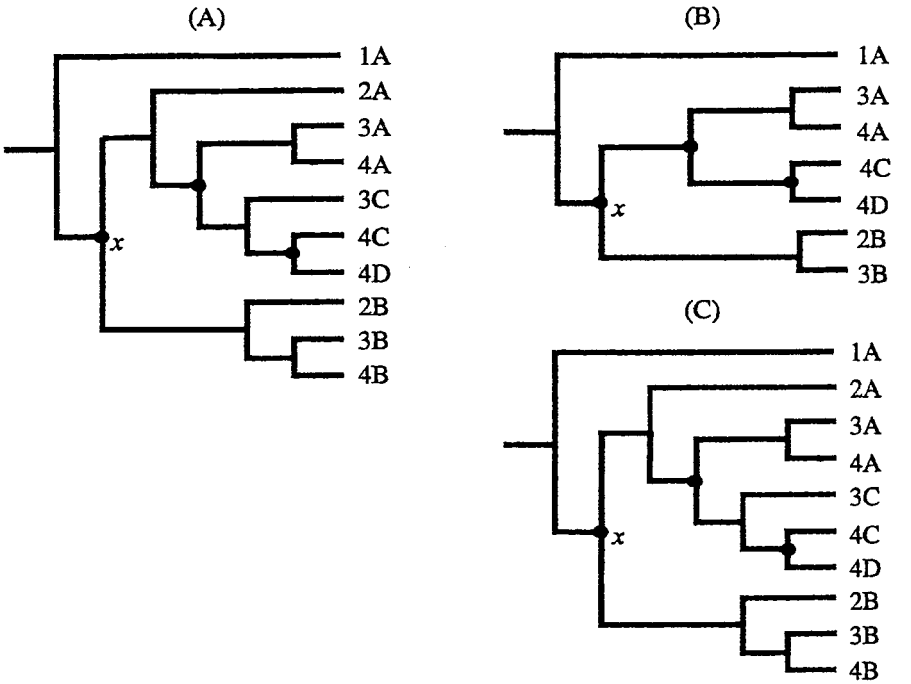


Figure 3 Inferring gene duplications and losses from a gene family tree. *A.* Gene duplications must be inferred at nodes where the gene tree is incompatible with the species tree (*red circles*). *B.* Gene duplications can be inferred even when some sequences are lost or missing. *C.* Reconciled tree for gene tree in *B*, with hypothetical branches leading to lost or missing sequences shown in *blue*.

Gene Duplication and Loss

Comparison of the inferred gene tree to a well-supported taxonomic phylogeny allows gene duplication events to be inferred and roughly dated. As Figure 3*a* (see color insert) shows, a gene duplication must be postulated at the base of any clade that contains a lineage whose branching order is incompatible with the taxonomic phylogeny. The location of the duplication event on the gene tree gives a lower bound of its age, because the duplication must have occurred prior to the divergence of all the lineages represented within that clade on the gene family tree. For example, Figure 3*a* suggests that the duplication labeled *x*—the event that created the paralogous gene groups A and B (and ultimately C too)—occurred prior to the taxonomic divergence of the lineage that contains species 2 from that containing species 3 and 4. Using this approach, it has been shown that a large number of gene families—including the *Hmg/Sox* genes, the nuclear receptors, *Wnt* genes, and several families of growth factors and their receptors—have diversified in two major phases. One wave of gene duplication occurred early in the metazoan

lineage, before the divergence of protostomes and deuterostomes, and another took place in the chordate lineage, before or during the early emergence of vertebrates (18, 25, 87, 90, 110, 111, 115, 124).

This kind of reasoning can be used even when gene sequences are not included in an analysis because they have been lost or have not been sequenced, as shown in Figure 3*b*. If sequences for genes 2A, 3C, and 4B are all missing, it is still necessary to postulate that duplication *x* occurred before the divergence of these three species, because the subtree ((2,3),4) is incompatible with the given taxonomic tree (1,(2,(3,4))). This conclusion is possible despite the fact that no species is known to possess all three paralogs in its genome, as is required to infer a gene duplication by phenetic approaches.

Losses or incomplete sampling can be inferred in an analogous way. Once the appropriate duplications are postulated in Figure 3*b*, for instance, it is possible to construct a "reconciled tree" (47, 93, 95) that resolves conflict between the species tree and the gene tree by including hypothetical branches for sequences that must have been lost or are not yet discovered. The reconciled tree (Figure 3*c*) makes it clear that 2A, 3C, and 4B must exist, either as unsequenced genes or unrecognizable pseudogenes. The visual reasoning that is facilitated by a reconciled tree can be formalized in the following rule: each branch that diverges from a gene duplication node *i* must lead immediately to another node *j* that contains one or more genes from all taxa descended from the taxonomic ancestor in which the gene duplication occurred. If it does not, then an intervening branch that leads to the lost gene or genes must be added between *i* and *j*.

Using the reconciled tree, it is also possible to predict which gene family members are likely to be found in species whose genomes have not yet been completely sequenced, providing guidance to laboratory work; given a species *y* that is known to contain gene *z*, all species that share a common ancestor with *y* more recently than the gene duplication that led to the appearance of *z* will also contain *z* (or, for gene losses, a pseudogene of *z*). Thus, for example, the discovery in the *D. melanogaster* genome of an estrogen-related receptor, previously known to exist only in vertebrates, implies that all protostomes and deuterostomes will also possess the gene, except for those descended from organisms in which the gene has been lost.

These rules for inferring gene duplication and loss have been automated in the programs Component (91) and GeneTree (94). These tools reconstruct the most parsimonious hypothesis of the process of gene creation and loss, given a gene family phylogeny and a species tree. If either the gene phylogeny or the taxonomic tree is weakly supported, interpretation may be more ambiguous, requiring the investigator to weigh the parsimony criterion for inferring gene losses against the parsimony criterion by which the gene tree and species tree were constructed. For example, the alignable portions of genes in the steroid receptor family are so conserved that the taxonomic relationships of some orthologs within and among mammalian orders are unorthodox and have very low Bremer supports

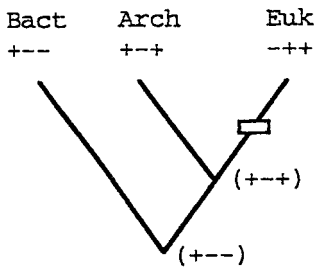
(125). Rather than to postulate gene losses each time an anomaly like this occurs, it may be preferable to interpret the topology of the gene tree at these nodes as inaccurate. This approach can be formalized by choosing a ratio that expresses the cost of a gene loss relative to an amino acid character change and then, using the method of Goodman et al (47), to choose the reconciled tree that minimizes the weighted sum of amino acid changes and gene losses. Although determining the proper value for this relative weight is ultimately a subjective and somewhat arbitrary choice (37), this procedure ensures consistency and allows the sensitivity of evolutionary reconstructions to the weight chosen for gene losses versus parallel or reverse sequence changes to be analyzed. Ultimately, this problem is seldom of great consequence. In most gene family phylogenies, the nodes that are interesting from an evolutionary perspective—those that reveal the timing of gene duplications and the relations among paralogs—are generally at relatively deep taxonomic levels; unexpected relationships near the tips of the gene tree—such as in the relationships among orthologs within the same taxonomic class or order—have no effect on the evaluation of most hypotheses about gene family evolution.

Inferring Ancestral Gene Sets

Once the series of gene duplication and loss events has been inferred from a gene family phylogeny, it is possible to infer what genes in the family were present in ancestral organisms at any specified level. Repeated for many or all gene families in the genome, it becomes possible to reconstruct a minimal version of the complete gene set of ancestral organisms, such as the ancestor of eukaryotes and archaea or the last common cellular ancestor of all three major domains in the tree of life.

Several comparative genomic studies have attempted to reconstruct ancestral gene ensembles by listing the presence and absence of orthologs in the genomes of descendant taxa and taking the intersection of these sets (62, 66). This approach, however, can be compromised by the loss of genes in some lineages or by nonorthologous gene displacement. As a result, the intersection of the sets of descendants' genes will underestimate the ensemble of genes or functional capacities that the common ancestor truly had.

A better way to approach this problem is to view genes as evolving characters in the hierarchical context implied by phylogeny. Methods for reconstructing the characteristics of hypothetical ancestral taxa on a phylogeny are well developed, have been automated in PAUP* and MacClade (75), and can be used to infer the presence or absence of genes in the genome of an ancestor (Figure 4). Consider for example the implications of the presence in archaea but not eukarya of similar orthologs of many bacterial genes, despite the currently accepted phylogeny (bacteria, (archaea, eukarya)) (6, 86, 98). The intersection-of-lists approach would conclude that these genes were absent from the common ancestor of all three



Taxon	Gene 1	2	3
Bacteria	+	-	-
Archaea	+	-	+
Eukarya	-	+	+
∩(AE)	-	-	+
∩(BAE)	-	-	-

Figure 4 Inferring the presence or absence of genes in the genomes of ancestral organisms. The tree shows the most parsimonious reconstruction for the presence/absence of three hypothetical genes in the ancestors of the major domains of life (character states in parentheses), given the phylogeny shown. State changes for gene 1 are mapped on the tree. When genes are lost in some lineages, the intersection of the gene sets of descendant organisms (∩) will underestimate the gene set of their last common ancestors. For gene 1, it is more parsimonious to infer a single gene loss in the eukaryotic lineage (*open box*) than to invoke two independent gains or a lateral gene transfer.

domains and were therefore gained independently in bacteria and archaea. It is more parsimonious to infer that these genes were present in the common ancestor and were then lost in the eukaryotes. A hierarchical understanding of the inheritance of orthologs as characters also suggests a simple reason why some genes are more similar between archaea and bacteria than between archaea and eukarya (6, 86); many such genes may have been highly conserved from their ancestral state in the archaeal and bacterial lineages and later diverged or were lost under unique selection pressures in the eukaryotes.

The estimation of an ancestral gene set allows functional inferences to be made about the integrated capacities of ancestral organisms. For example, the presence or absence of genes that are critical to certain biochemical or regulatory pathways implies whether the ancestor's metabolic and physiological processes included that pathway. Using this approach, for instance, Brown & Doolittle (13) have argued that the last common ancestor to all contemporary cells synthesized transfer RNA (tRNA)-glutamine complexes by charging tRNAs with glutamate and then transamidating the glutamate to glutamine, although this reaction is not found in contemporary eukaryotes, which charge tRNAs directly with glutamine.

Rooting

Drawing inferences about evolutionary history from a phylogeny requires that the tree be properly rooted, which is not always straightforward in gene family analyses. The only scenario in which empirical evidence would support the designation of one gene family member as an outgroup would be if a stem taxon were revealed to contain a single member of a gene family, thus designating this gene as the "ancestor." Except when whole genomes are available, the possibility that

gene loss has occurred in other lineages ensures some ambiguity in rooting by this method. There is another criterion, however, by which a tree may be rooted: the degree to which it preserves the expected taxonomic structure within each group of orthologs. Each of the possible rooted trees that can be derived from an unrooted gene family tree disrupts the expected taxonomic subtrees within each group of orthologs to a varying extent, and each thus requires a different number of gene losses to be assumed. The rooted tree(s) that requires the fewest assumed gene losses is the most parsimonious and therefore the preferred hypothesis of the process of gene duplication and loss. This approach is an adaptation of a method developed for rooting the deepest nodes in the tree of life, using duplicated genes, such that the root is placed to preserve the expected structure among pairs of highly conserved orthologous genes (14, 59, 67, 98).

Consider the example in Figure 5, which shows three of many ways of rooting an unrooted gene tree. The tree in Figure 5*B* requires no gene losses, that in Figure 5*C* requires one loss, and that in Figure 5*D* requires four losses. Given the sequence data and the taxonomic tree, then, the tree in Figure 5*B* is a slightly better hypothesis than that in Figure 5*C*, and both are considerably better than that in Figure 5*D*. To prefer a tree that is more parsimonious in terms of gene losses is well justified when missing branches are caused by the actual loss of a gene; such a tree offers the most complete explanation of the distribution of gene family members in various species as the result of inheritance from common ancestors, with a minimum of ad hoc hypotheses of additional gene duplications and losses. When missing branches may be caused by incomplete sampling, however, the parsimony criterion is less persuasive; to suggest that there may be unidentified gene family members in the genome of a species that has not been thoroughly studied is not a burdensome ad hoc hypothesis. Rooting may thus remain somewhat ambiguous except in cases in which complete genomes have been sequenced or exhaustive searches for gene family members have been conducted by polymerase chain reaction and library screening. Like the ambiguity in inferring losses, however, this problem is not an overwhelming one; even in the absence of an unambiguous root, the two plausible trees in Figure 5*B* and Figure 5*C* are congruent at all nodes but two, and they imply nearly identical evolutionary processes, differing only in the timing of the first gene duplication. Rooting may thus remain incompletely determined without precluding a substantial degree of evolutionary inference.

ANALYZING PROTEIN EVOLUTION IN A PHYLOGENETIC CONTEXT

With a well-supported gene family phylogeny in hand, detailed analysis of the mechanisms and dynamics of gene and protein evolution can begin. Numerous methods are available to reconstruct at the molecular level the series of events by which gene families diversified and took on specific functions.

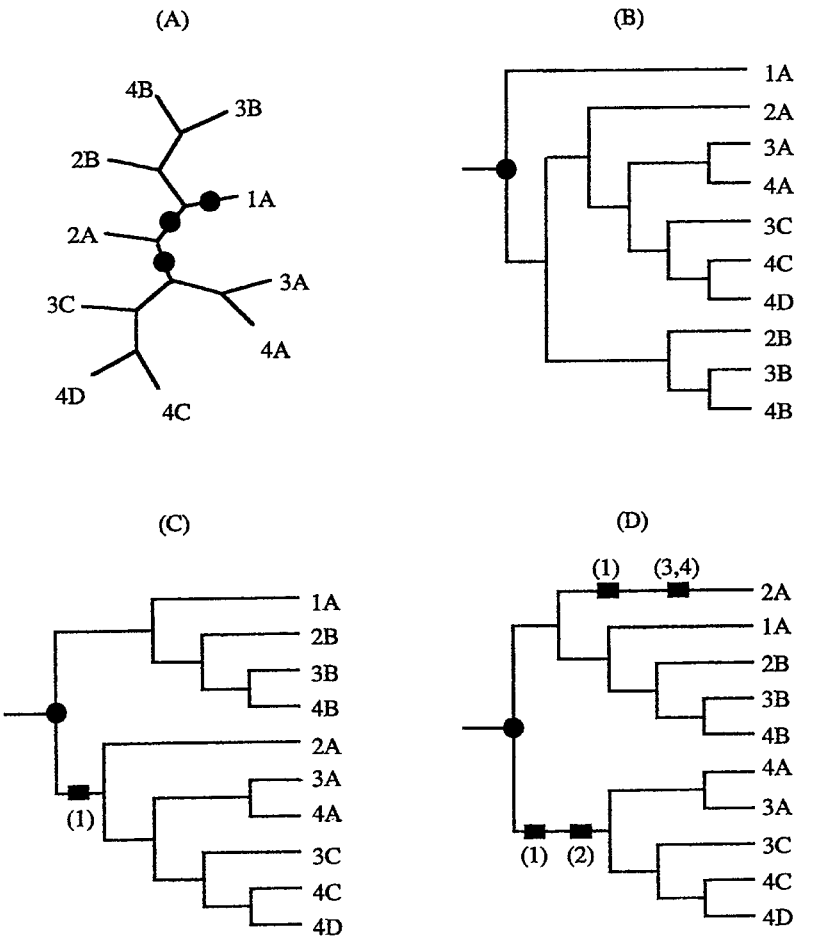


Figure 5 Inferring the root of a gene tree by minimizing ad hoc hypotheses of gene losses. *A*. Unrooted gene family tree. *B–D*. Gene trees rooted on branches labeled in *A*. Gene losses that must be hypothesized are marked with *dark boxes*; labels show the taxa in which the gene has been lost. Tree *B* is the most parsimoniously rooted tree.

Genome Mapping and Mechanisms of Gene Proliferation

Gene duplication can occur by tandem duplication (due to replication slippage or unequal recombination), duplications of whole genomes or large parts thereof, transposition of DNA sequences, or retrotransposition of RNA transcripts. When genomic information is available to specify the chromosomal location of genes in a superfamily, these mechanisms leave unique traces in the genome (see 128, for example).

Where tandem duplication has created paralogs, closely related gene family members will be tightly linked on a single chromosome. Large-scale duplications, such as those caused by polyploidization, result in gene family members that are scattered in the genome, and each descendant of such an event will be linked to members of the other gene families that have been duplicated in the same event. For example, the existence of one cluster of linked homeobox genes in amphioxus, four in mammals, and as many as seven in teleosts has led to the suggestion that there were two genome-wide duplications early in the vertebrate lineage and another after ray-finned fish diverged from the lobe-finned lineage that led to tetrapods (53, 79, 80). But the increase in copies of a single gene cluster could be the result of local rather than global DNA duplications. More suggestive evidence for the hypothesis of two successive genome duplications in the vertebrate lineage comes from the existence of numerous "tetralogous" gene groups—linked assemblages of members of several gene families that are repeated three or four times throughout the genomes of humans and mice (45, 116, but see 112). The number of genes in such groups, however, provides only weak evidence in favor of the hypothesis of serial genome duplication, which predicts specific phylogenetic relationships among paralogs. Gene trees with the topology ((A,B,),(C,D)) would support this hypothesis, but any other set of relations will refute it. Preliminary analysis of several families of developmentally important genes suggests that the phylogeny predicted by the genome duplication hypothesis seldom occurs (57).

Gene family members created by retrotransposition can be detected by the lack of introns in their sequences. The *CDY* gene on the human Y chromosome, for instance, appears to be an integrated retrosequence derived early in the primate lineage from the messenger RNA transcribed from an autosomal, intron-containing gene called *CDYL* (68). Fragmented or intact retroelements may also be detectable in positions up- and downstream from genes that have been inserted by retrotransposition (43). DNA transpositions have none of these characteristics, but the presence of conserved terminal sequences of mobile genetic elements in regions flanking the genes would provide evidence of such a process.

Domain Shuffling and Lateral Gene Transfer

New gene family members can also be created by "horizontal transfer" of genetic information between more ancient paralogous genes. Many proteins, particularly those in recognizable gene families, are composed of domains—discrete structural units with specific and often autonomous functions. Domain shuffling, which can occur by transposition of gene fragments or nonhomologous recombination, is thought to have been a major mechanism in the evolution of new proteins (22, 23). Domain shuffling in the history of a gene family can be examined by analyzing separately the phylogeny of protein domains. If some members of a family have been created as evolutionary chimeras by the shuffling of domains from more ancient members, the phylogenies of the domains will be incongruent, with chimeric

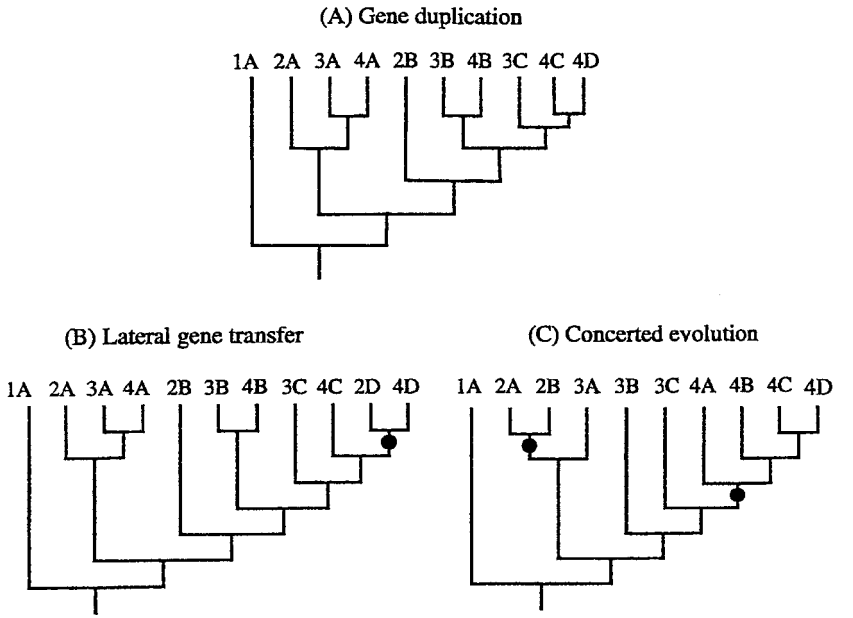


Figure 6 Lateral gene transfer and concerted evolution cause distinct patterns of phylogenetic incongruence. *A*. Hypothetical gene family phylogeny without lateral gene transfer or concerted evolution. *B*. Inferred gene phylogeny from the same process but with lateral gene transfer of gene D from taxon 4 to taxon 2. *C*. Inferred gene phylogeny from the same process but with concerted evolution in species 2 and 4. *Blue* and *red circles* mark nodes that are incongruent with the phylogeny in *A* due to lateral gene transfer and concerted evolution, respectively.

proteins grouping with one set of orthologs in the phylogeny inferred from the sequence of one domain and with another set in the tree from the other domain.

Incongruence will also occur with lateral gene transfer from one organism to another, but in this case the relevant partition of the data set is not between domains but between genes or groups of genes. The incongruence will place transferred genes at nodes that are incongruent with the tree expected based on vertical descent (Figure 6, see color insert). On this basis, several reports have argued that there has been massive transfer of genes involved in energy metabolism between archaea and thermophilic eubacteria (60, 86) and between gram-negative bacteria and eukarya (101).

Incongruence alone is extremely weak evidence of horizontal transfer. Some incongruence is expected simply as a chance result of partitioning of a data set into parts, because the effect of noise on topology is greater in smaller data sets than large ones. Hypotheses of domain shuffling or lateral gene transfer must therefore be tested to ascertain whether the observed incongruence among domains is greater than would be expected by chance alone. The incongruence length difference

(ILD) test of Farris and colleagues (31) provides a parsimony-based nonparametric statistical test of incongruence between two data subsets. It computes the number of extra steps imposed by analyzing the subsets together as compared with their separate analysis, and it compares this degree of incongruence to that observed for a large number of random partitions in the same data set. Significant incongruence in the ILD test cannot be explained by random, nondirected noise (homoplasy) in the sequence; it provides evidence of a substantially conflicting phylogenetic signal between domains or genes.

One limit of Farris' test is that it examines the entire tree at once; significant global incongruence can be present if a single gene has been created by domain shuffling or transferred horizontally, so it would be useful to know which nodes contribute to the overall incongruence. The local incongruence length difference (LILD) test applies the ILD technique to each clade in a tree, allowing the incongruence at each node to be quantified and its statistical significance to be tested (124). Other methods developed to assess recombination between alleles (reviewed in 19) can be adapted for assessing the horizontal transfer of information between pairs of genes in a family. For example, Huelsenbeck & Bull's test compares the likelihood of a tree without recombination to that of a tree in which recombination—or in this case domain shuffling or lateral transfer—occurred (55).

All of these tests assume a priori knowledge of where gene sequences should be divided into subsets whose phylogenies can be separately inferred. The partition between domains can be based on a priori structural and functional data about the gene structure. Protein domains are identified biochemically with deletion experiments or the creation of chimeric proteins, and they are often separated from each other by introns or more variable coding regions, justifying the placement of analytical partitions between them. Alternatively, potential partitions can be derived from the sequence data itself. Crandall & Templeton (19) have reviewed several methods for locating recombination sites, including a phylogenetic approach based on a statistical test of the linear distribution of diagnostic characters in the sequence. With this method, the characters that support a node with potential incongruence caused by domain shuffling are plotted along the length of the sequence, as are those that support alternative topologies. If domain shuffling occurred, the characters supporting the node are expected to be contiguous, with a discrete point in the linear sequence at which support for an alternative phylogeny begins to dominate. The probability that the observed clustering of synapomorphies for each tree could have arisen by chance can be calculated by reference to the hypergeometric distribution of clustering that would be expected by chance alone. One caveat with these methods is that the statistical power of the LILD and other tests for evaluating incongruence becomes weaker as the domains evaluated become shorter.

Concerted Evolution

Unlike domain shuffling, which serves as a mechanism for rapid creation of new proteins, concerted evolution tends to homogenize paralogs within a genome

(24, 104). Concerted evolution can be caused by unequal recombination, gene conversion, or replication slippage. It has been important in some gene families—particularly those, like the ribosomal RNAs, that occur in tandem arrays and cause dosage repetition, a selective advantage to the organism conferred by having additional copies of a gene (reviewed in 48). Gene conversion appears to be much less frequent in families of transcription factors and other regulatory proteins, which seldom occur in tandem and for which dosage repetition is of little selective value. One study of gene families in the *Caenorhabditis elegans* genome, for instance, found evidence of concerted evolution for only 2% of genes (108), and there is no evidence of concerted evolution in the steroid receptor family (125).

Concerted evolution will create specific forms of incongruence between gene family trees and taxonomic trees (Figure 6). If pairs or groups of paralogous genes from a single genome cluster together—particularly when the same pattern is repeated for the same genes in several taxa—concerted evolution may be the cause. The same pattern, however, could be caused by independent duplication of the same gene late in each lineage. This ambiguity can be resolved by a detailed examination of homogeneity in different regions of the gene. The known mechanisms of concerted evolution affect DNA segments of relatively short length; outside these homogenized stretches, the sequences of family members should remain unhomogenized. In contrast, genes that are similar within a genome due to recent duplications should be relatively homogeneous along their entire lengths (24). Sawyer provides a technique to identify local gene conversion events in nucleotide sequences (105). Clades at which concerted evolution may have occurred are identified on the tree as those that unite genes from a single genome. Sites that are synonymous (do not affect the protein sequence) and contain the same nucleotide in the two candidate genes and a different one in the closest outgroup sequence may have been homogenized by concerted evolution; the clustering of such sites in contiguous regions of the gene is evidence for gene conversion rather than homoplasy. Semple & Wolfe (108) have adapted a statistical method that was originally used to identify potential recombination sites for testing whether the observed contiguity of homogenized sites is significantly greater than expected by chance alone.

Sequence Divergence and the Evolution of Function

The ultimate goal of gene family studies is an understanding of how duplicated genes have taken on novel biochemical and organismal functions. Domain shuffling aside, it remains a mystery how the undirected process of mutation, combined with natural selection, has resulted in the creation of thousands of new proteins with extraordinarily diverse and well-optimized functions. This problem is particularly acute for tightly integrated molecular systems that consist of many interacting parts, such as ligands, receptors, and the downstream regulatory factors with which they interact. In these systems, it is not clear how a new function for any protein might be selected for unless the other members of the complex are already present, creating a molecular version of the ancient evolutionary riddle of the chicken and the egg. Detailed studies of gene family evolution promise to shed some light on

the process by which changes at the genetic level have led to the diversification of function for members of such integrated molecular systems.

To understand the evolution of structure-function relationships, we must first independently reconstruct the evolution of structure and of function. Phylogenetic methods for tracing the evolution of characters on cladograms and reconstructing ancestral sequences make it possible to infer the evolutionary history by which various aspects of protein function—catalytic activity, spatial and temporal regulation of expression, or affinity for a certain type of substrate, ligand, response element, or cofactor—have been gained, lost, or transformed on each branch of the gene tree. In particular, we can ask whether such functions evolved consistently on a phylogeny (were created once and conserved thereafter in all descendants) or whether they evolved in parallel, by convergence, or with reversals/loss in some lineages (5, 28, 69, 115, 124). When functions have evolved consistently, it is also possible to infer the functional characteristics of hypothetical ancestral proteins, even at the deepest nodes near the root of a gene family tree. Applying this approach to a tree of the tRNA family, for example, Fitch & Upper corroborated the hypothesis that the genetic code evolved by a progressive reduction in ambiguity and a gradual increase in specificity of association between smaller and smaller classes of codons and amino acids (38).

The same techniques can be used to trace the evolution of the primary structure of proteins on a cladogram. One useful strategy is to identify specific amino acids that "diagnose" a clade of proteins, particularly a group whose members share an important functional characteristic. These synapomorphic amino acids are expected to include those that are required for the function that appeared on the same branch. Genetic and biochemical studies of structure-function relations have found that many of these phylogenetically diagnostic amino acids are indeed essential to function, validating this technique for the prediction of structure-function relationships (125).

Ancestral sequences can also be reconstructed by using ML methods on a tree inferred by parsimony or ML. Yokoyama & Radlwimmer (129), for example, used an ML algorithm to reconstruct the sequences of ancestral red-green opsin proteins and predict the color sensitivity of those sequences based on empirical evidence of the effect of specific amino acids at critical sites. They showed that mammalian color vision has evolved from an ancestral green-sensitive opsin, with rampant parallelism and reversal at both the sequence and functional levels.

Once the evolution of its individual components has been traced on a gene tree, the evolution of structure-function relationships can be analyzed. In particular, correlations in the evolution of structure and of function can be located on the cladogram, and the evolutionary sequence of mutations that led to the appearance of protein clades with unique functions can be inferred. In some cases, specific amino acid changes can be associated with the emergence of specific functions (65, 129). Ultimately, with a densely sampled gene family phylogeny, it should be possible to characterize the dynamics of the evolutionary relationship between primary structure and specific aspects of protein function. Of particular interest is the extent to which functional changes take place gradually due to

accumulated mutations at individual sites (46) or whether they are emergent properties of complex combinations of sequence changes.

The role of selection in gene family evolution can be investigated by estimating relative and absolute rates of sequence divergence for different branches in the gene family tree. If the rate of mutation is more or less constant, then differences in divergence rates should indicate the strength of selection acting on any branch in the tree. Comparing the rates at which two paralogs have diverged since the same cladogenetic events suggests the relative importance of selection on each paralog, indicating whether some proteins in a family have been more constrained by selection than others (61, 69, 125). Absolute rates of divergence among orthologs can be calibrated with cladogenesis dates inferred from the fossil record (42). If taxa have been sampled densely enough, it should be possible to test the hypothesis that paralogs have evolved faster immediately after duplication events, followed by a slowing of sequence divergence thereafter (73). Likelihood tests have also been developed that provide a more computationally sophisticated means for characterizing divergence rates (102).

Finally, a gene family phylogeny can be compared to the phylogeny of other gene families with which its members interact at the molecular level to understand the coevolution of interacting proteins. Fryxell has examined the evolution of peptide hormones, growth factors, and cytokines with their receptors, using a comparative phylogenetic approach (40). These interacting gene families have diversified in a coordinated fashion, so that newly duplicated receptors gain affinity for newly duplicated ligands (see also the study of fibroblast growth factors and their receptors in 18). According to this model, simultaneous diversification of interacting protein families creates the conditions under which a duplicated receptor can take on a novel role and avoid the otherwise likely fate of transformation into a pseudogene. But some other gene families do not seem to follow this pattern; phylogenies of the transforming growth factor-β growth factors, their receptors, and the intracellular Smad proteins that transduce their signals are incongruent (87). This finding suggests that interacting proteins diversified independently, a scenario that is consistent with the subfunctionalization hypothesis (39) but not coevolution.

Gene Family Evolution and Organismal Novelty

Some investigators have examined phylogenetic correlations between the timing of gene duplication events and major evolutionary changes in developmental and physiological programs. For example, the extreme diversification in the arthropod lineage of the cytochrome p450 superfamily of enzymes for oxidative detoxification appears to have occurred at about the same time that arthropods began to feed on land plants, with their wide variety of deterrent and poisonous compounds (72).

More ambitiously, numerous investigators have proposed that the expansion of *Hox* genes, growth factors, or nuclear receptors early in the chordate or vertebrate lineage caused or enabled the increased morphological and regulatory

complexity of the crown vertebrates (8, 18, 28, 44, 53, 107, 111). Of course, a phylogenetic correlation between the expansion of a gene family and the appearance of new organismal features does not in itself imply causality, especially because genome-wide duplications would have led to the simultaneous expansion of many gene families. Recent findings in gene family evolution argue against such grand causal links between genetic and morphological evolution. First is the discovery of large numbers of *Hox* clusters in fish and priapulids, neither of which is by any clear measure more physiologically complex than tetrapods, which have considerably fewer *Hox* genes (21, 80). Second, many gene families appear to have undergone their major expansions before the divergence of sponges from the eumetazoan line (90, 117), refuting the hypotheses that gene family diversification was directly and causally linked to the Cambrian explosion (44, 111), in which the diversity and complexity of animal body plans are thought to have suddenly increased.

TOWARD A SAMPLING OF GENOMIC BIODIVERSITY

The age of high-throughput sequencing promises to revolutionize gene family studies and to make phylogenetic techniques and principles indispensable for genome analysis. The availability of complete gene sequences will allow researchers to "see" the absence of a protein from a genome, removing a major source of ambiguity in the interpretation of gene family phylogenies. For a representative picture of protein evolution, however, it will be necessary to have genomic sequence data from more than the handful of organisms now being sequenced (71). These organisms have been chosen for their biomedical or agricultural importance or their utility as model organisms for genetic and developmental analysis. But genomes from flies, worms, mice, zebrafish, yeasts, and arabodopsis on one hand—and corn, cotton, cows, trypanosomes, and humans on the other—are ultimately inadequate to inform rigorous inference about gene family evolution.

The complete genomes of all of the millions of species in nature will not be sequenced in the foreseeable future. But it is not unreasonable to hope that the choice of organisms to be studied in depth will be informed by phylogenetic relationships. From the perspective of evolutionary inference, the greatest gains will come by obtaining sequences from highly informative stem taxa—sister lineages to groups of major biological interest, such as vertebrates, chordates, bilaterians, and metazoa. For example, the additional taxa that are critical for metazoan comparative genomics are not more rodents, more primates, more teleosts, or more nematodes but the far less glamorous lamprey, hagfish, amphioxus, tunicates, echinoderms, cnidarians, sponges, and choanaflagellates. No matter what taxa become the focus of future sequencing projects, however, this much is clear: as the genomic data come pouring in, the best way to make sense of them—at conceptual, functional, and historical levels—is to begin with phylogenetics.

ACKNOWLEDGMENTS

We thank Darcy Kelley for helpful comments on the manuscript. This work was supported by National Science Foundation grant DEB 9870055 and by Columbia University's Center for Environmental Research and Conservation.

LITERATURE CITED

1. Abouheif E, Akam M, Dickinson WJ, Holland PWH, Meyer A, et al. 1997. Homology and developmental trends. *Trends Genet.* 13:432–33
2. Agosti D, Jacobs D, DeSalle R. 1996. On combining protein sequences and nucleic acid sequences in phylogenetic analysis: the homeobox protein case. *Cladistics* 12: 65–82
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–10
4. Amero, SA, Kretsinger RH, Moncrief ND, Yamamoto KR, Pearson WR. 1992. The origin of nuclear receptor proteins: a single precursor distinct from other transcription factors. *Mol. Endocrinol.* 6:3–7
5. Applebury ML. 1994. Relationships of G-protein-coupled receptors: a survey with the photoreceptor opsin subfamily. See Ref. 28a, pp. 235–48
6. Aravind LA, Tatusov RI, Wolf YI, Walker R, Koonin EV. 1998. Evidence for massive gene exchange between archael and bacterial hyperthermophiles. *Trends Genet.* 14:442–44
7. Atchley WR, Fitch WM. 1997. A natural classification of the basic helix-loop-helix class of transcription factors. *Proc. Natl. Acad. Sci. USA* 94:5172–76
8. Baker ME. 1997. Steroid receptor phylogeny and vertebrate origins. *Mol. Cell. Endocrinol.* 135:101–7
9. Bonci A, Chiesurin A, Muscas P, Rossolini GM. 1997. Relatedness and phylogeny within the family of periplasmic chaperones involved in the assembly of pilli or capsule-like structures of gram-negative bacteria. *J. Mol. Evol.* 44:299–309
10. Bremer K. 1995. Branch support and tree stability. *Cladistics* 10:295–304
11. Brendel V, Brocchieri L, Sandler SJ, Clark AJ, Karlin S. 1997. Evolutionary comparisons of RecA-like proteins across all major kingdoms of living organisms. *J. Mol. Evol.* 44:528–41
12. Brower AVZ, Schawaroch V. 1996. Three steps of homology assessment. *Cladistics* 12:265–72
13. Brown JR, Doolittle WF. 1999. Gene descent, duplication, and horizontal transfer in the evolution of glutamyl- and glutaminyl-tRNA synthetases. *J. Mol. Evol.* 49:485–95
14. Brown JR, Robb FT, Weiss R, Doolittle WF. 1997. Evidence of the early divergence of tryptophanyl- and tyrosyl-tRNA synthetases. *J. Mol. Evol.* 45:9–16
15. Burt DW, Bruley C, Dunn IC, Jones CT, Ramage A, et al. 1999. The dynamics of chromosome evolution in birds and mammals. *Nature* 402:411–13
16. Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, et al. 1998. Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* 282:2022–28
17. Chiu J, DeSalle R, Lam HM, Meisel L, Coruzzi G. 1999. Molecular evolution of glutamate receptors: a primitive signaling mechanism that existed before plants and animals diverged. *Mol. Biol. Evol.* 16: 826–38
18. Coulier F, Pontarotti P, Roubin R,

Hartun G, Goldfarb M, Birnbaum D. 1997. Of worms and men: an evolutionary perspective on the fibroblast growth factor (FGF) and FGF receptor families. *J. Mol. Evol.* 44:43–56

19. Crandall KA, Templeton AR. 1999. Statistical approaches to detecting recombination. In *The Evolution of HIV*, ed. KA Crandall, pp. 153–76. Baltimore, MD: Johns Hopkins Univ. Press
20. De Pinna MCC. 1991. Concepts and tests of homology in the cladistic paradigm. *Cladistics* 7:367–94
21. de Rosa R, Grenier JK, Andreeva T, Cook CE, Adoutte A, et al. 1999. Hox genes in brachiopods and priapulids and protostome evolution. *Nature* 399:772–76
22. Doolittle RF, 1995. The multiplicity of domains in proteins. *Annu. Rev. Biochem.* 64:287–314
23. Doolittle RF, Bork P. 1993. Evolutionarily mobile modules in proteins. *Sci. Am.* 269(4):50–56
24. Dover GA, Linares AR, Bowen T, Hancock HM. 1993. Detection and quantification of concerted evolution and molecular drive. *Methods Enzymol.* 224: 525–41
25. Ebendal T. 1992. Function and evolution in the NGF family and its receptors. *J. Neurosci. Res.* 32:461–70
26. Enright AJ, Illiopoulos I, Kyrpides, Ouzounis CA. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402:86–90
27. Eriksson T. 1996. Auto-Decay, version 2.9.5 (software). Stockholm, Swed: Stockholm Univ.
28. Escriva H, Safi R, Hanni C, Langlois MC, Saumitou-Laprade P. 1997. Ligand binding was acquired during evolution of nuclear receptors. *Proc. Natl. Acad. Sci. USA* 94(13):6803–8

28a. Fambrough DM, ed. 1994. *Molecular Evolution of Physiological Processes.* New York: Rockefeller Univ. Press

29. Farris JS. 1982. Distance data in phylogenetic analysis. *Adv. Cladistics* 1:3–23
30. Farris JS. 1983. The logical basis of phylogenetic analysis. *Adv. Cladistics* 2:7–36
31. Farris JS, Kallersjo M, Kluge AG, Bult C. 1995. Constructing a significance test for incongruence. *Syst. Biol.* 44:570–72
32. Felsenstein J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. *Syst. Zool.* 27: 401–10
33. Felsenstein J. 1978. The number of evolutionary trees. *Syst. Zool.* 27:27–33
34. Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17:368–76
35. Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783–91
36. Fitch WM. 1970. Distinguishing homologous from analogous proteins. *Syst. Zool.* 19:99–113
37. Fitch WM. 1979. Cautionary remarks on using gene expression events in parsimony procedures. *Syst. Zool.* 28:375–79
38. Fitch WM, Upper K. 1987. The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. *Cold Spring Harbor Symp. Quant. Biol.* 52:759–67
39. Force A, Lynch M, Pickett FB, Amores A, Yan Y-L, Postlethwaite JH. 1996. Preservation of duplicate genes by complementary degenerative mutations. *Genetics* 151:1531–45
40. Fryxell KJ. 1996. The coevolution of gene family trees. *Trends Genet.* 12:364–69
41. Gatesy J, DeSalle R, Wheeler W. 1993. Alignment-ambiguous nucleotide sites and the exclusion of systematic data. *Mol. Phylogenet. Evol.* 2:152–57
42. Gatesy JC, Hayashi C, DeSalle R, Vrba E. 1994. Rate limits for mispairing and compensatory change: the mitochondrial ribosomal DNA of antelopes. *Evolution* 48:188–96

43. Gaudieri S, Leelayuwat C, Townsend DC, Kulski JK, Dawkins RL. 1997. Genomic characterization of the region between HLA-B and TNF: implications for the evolution of multicopy gene families. *J. Mol. Evol.* 44(Suppl.1):S147–54
44. Gellon G, McGinnis W. 1998. Shaping animal body plans in development and evolution: modulation of Hox expression patterns. *BioEssays* 20:116–25
45. Gibson TJ, Spring J. 1998. Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends Genet.* 14:46–49
46. Golding GB, Dean AM. 1998. The structural basis of molecular adaptation. *Mol. Biol. Evol.* 15:355–69
47. Goodman M, Czelusniak J, Moore GW, Matsuda G. 1979. Fitting the gene lineage into its species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.* 28: 132–63
48. Graur D, Li WH. 1999. *Fundamentals of Molecular Evolution.* Sunderland, MA: Sinauer. 2nd ed. 481 pp.
49. Harvey PH, Pagel MD. 1991. *The Comparative Method in Evolutionary Biology.* Oxford, UK: Oxford Univ. Press
50. Hein J. 1990. Unified approach to alignment and phylogenies. *Methods Enzymol.* 183:626–44
51. Hendy MD, Penny D. 1989. A framework for the quantitative study of evolutionary trees. *Syst. Zool.* 38:297–309
52. Hennig W. 1963. Phylogenetic systematics. *Annu. Rev. Entomol.* 10:97–116
53. Holland PWH, Garcia-Fernandez J, Williams NA, Sidow A. 1994. Gene duplications and the origins of vertebrate development. *Development* (Suppl.) pp. 125–33
54. Holm L. 1998. Unification of protein families. *Curr. Opin. Struct. Biol.* 8:372–79
55. Huelsenbeck JP, Bull JJ. 1996. A likelihood ratio test for detection of conflicting phylogenetic signal. *Syst. Biol.* 42: 247–64
56. Huelsenbeck JP, Crandall KA. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst.* 28:437–66
57. Hughes AL. 1999. Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *J. Mol. Evol.* 48:565–76
58. Hull DL. 1988. *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science.* Chicago, IL: Univ. Chicago Press
59. Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. USA* 86:9355–59
60. Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. USA* 96:3801–6
61. Kissinger JC, Hahn J-H, Raff RA. 1997. Rapid evolution in a conserved gene family: evolution of the actin gene family in the sea urchin genus Heliocidaris and related genera. *Mol. Biol. Evol.* 14:654–65
62. Koonin EV, Muhegian AR, Bork P. 1996. Non-orthologous gene displacement. *Trends Genet.* 12:334–36
63. Koonin EV, Tatusov RL, Galperin MY. 1998. Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol.* 8:355–63
64. Koonin EV, Tatusov RL Rudd KE. 1996. Protein sequence comparison at genome scale. *Methods Enzymol.* 266:295–323
65. Kornegay JR, Schilling JW, Wilson AC. 1994. Molecular adaptation of a leaf-eating bird: stomach lysozyme of the hoatzin. *Mol. Biol. Evol.* 11:921–28
66. Kyripides N, Overbeek R, Ouzonis C. 1999. Universal protein families and the functional content of the last universal common ancestor. *J. Mol. Evol.* 49: 413–23

67. Labedan B, Boyen A, Baetens M, Charlier D, Chen P, et al. 1999. The evolutionary history of carbamoyltransferases: A complex set of paralogous genes was already present in the last universal common ancestor. *J. Mol. Evol.* 49:461–73
68. Lahn BT, Page DC. 1999. Retrotransposition of autosomal mRNA yielded testis-specific gene family on human Y chromosome. *Nat. Genet.* 21:429–33
69. Laudet V. 1997. Evolution of the nuclear receptor superfamily: early diversification from an ancestral orphan receptor. *J. Mol. Endocrinol.* 19:207–26
70. Laudet V, Hanni C, Coll J, Catzeflis F, Stehelin D. 1992. Evolution of the nuclear receptor gene superfamily. *EMBO J.* 11:1003–13
71. Leape DD. 1996. Biodiversity, genomes, and DNA sequence databases. *Curr. Opin. Genet. Dev.* 6:686–91
72. Lewis DFV. 1996. *Cytochromes P450: Structure, Function and Mechanism.* London: Taylor & Francis. 348 pp.
73. Li WH. 1985. Accelerated evolution following gene duplication and its implications for the neutralist-selection controversy. In *Population Genetics and Molecular Evolution*, ed. T Ohta, pp. 335–52. Berlin: Springer-Verlag
74. Li WH. 1997. *Molecular Evolution.* Sunderland, MA: Sinauer. 487 pp.
75. Maddison WP, Maddison DR. 1997. MacClade 3.07 (software). Sunderland, MA: Sinauer
76. Marcotte EM, Pellegrini M, Ng H-L, Rice DW, Yeates TO, Eisenberg D. 1999. Detecting protein function and protein interactions from genome sequences. *Science* 285:751–53
77. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature* 402:83–86
78. Mayr E. 1982. *The Growth of Biological Thought.* New York: Belknap Press
79. Meyer A. 1998. Hox gene variation and evolution. *Nature* 391:225–27
80. Meyer A, Malaga-Trillo E. 1999. Vertebrate genomics: more fishy tales about Hox genes. *Curr. Biol.* 9:R210–13
81. Murzin AG. 1998. How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol.* 8:380–87
82. Muse SV. 1999. Modeling the molecular evolution of HIV sequences. In *The Evolution of HIV*, ed. KA Crandall, pp. 122–52. Baltimore, MD: Johns Hopkins Univ. Press
83. Mushegian AR, Garey JR, Martin J, Liu LX. 1998. Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode and yeast genomes. *Genome Res.* 8:590–98
84. Nadeau JH, Sankoff D. 1998. Counting on comparative maps. *Trends Genet.* 14:495–501
85. Nelson G. 1994. Homology and systematics. In *Homology: The Hierarchical Basis of Comparative Biology*, ed. BK Hall, pp. 102–38. New York: Academic
86. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, et al. 1999. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* 399: 323–29
87. Newfeld SJ, Wisotzkey RG, Kumar S. 1999. Molecular evolution of a developmental pathway: phylogenetic analyses of transforming growth factor-β family ligands, receptors, and Smad signal transducers. *Genetics* 152:783–95
88. Nixon KC, Davis JL, Goloboff PA. 1998. Search strategies for large dataset: an example using rbcL. *Am. J. Bot.* 85:148–53
89. Ohno S. 1970. *Evolution by Gene Duplication.* Berlin: Springer-Verlag. 160 pp.
90. Ono K, Suga H, Iwabe N, Kuma K, Miyata T. 1999. Multiple protein tyrosine phosphatases in sponges and explosive gene duplication in the early evolution of

animals before the parazoan-eumetazoan split. *J. Mol. Evol.* 48:654–62
91. Page RDM. 1993. Component 2.0 (software). London: Nat. Hist. Mus.
92. Page RDM. 1993. Genes, organisms, and areas: the problem of multiple lineages. *Syst. Biol.* 42:77–84
93. Page RDM. 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol.* 43:58–77
94. Page RDM. 1998. Genetree 1.0 (software). Glasgow, Scotland: Univ. Glasgow
95. Page RDM, Charleston MA. 1997. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol. Phylogenet. Evol.* 7:231–40
96. Patterson C. 1988. Homology in classical and molecular biology. *Mol. Biol. Evol.* 5:603–25
97. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein function by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* 96: 4285–88
98. Phillippe H, Forterre P. 1999. The rooting of the universal tree of life is not reliable. *J. Mol. Evol.* 49:509–23
99. Raymond CS, Shamu CE, Shen MM, Seifert KJ, Hirsch B, et al. 1998. Evidence for evolutionary conservation of sex-determining genes. *Nature* 391:691–95
100. Reeck GR, de Haen C, Teller DC, Doolittle RF, Fitch WM, et al. 1987. Homology in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell* 50:667
101. Rice KA, Donoghue MJ, Olmstead R. 1997. Analyzing large data sets: rbcL 500 revisited. *Syst. Biol.* 46:554–63
102. Sanderson MJ. 1994. Reconstructing the history of evolutionary processes using maximum likelihood. See Ref. 28a, pp. 13–26
103. Sanderson MJ. 1995. Objections to bootstrapping phylogenies: a critique. *Syst. Biol.* 44:299–320
104. Sanderson MJ, Doyle JJ. 1992. Reconstruction of organismal and gene phylogenies from data on multigene families: concerted evolution, homoplasy and confidence. *Syst. Biol.* 41:4–17
105. Sawyer S. 1989. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* 6:526–38
106. Schledzewski K, Brinkmann H, Mendel RR. 1999. Phylogenetic analysis of components of the eukaryotic vesicle transport system reveals a common origin of adaptor protein complexes 1, 2 and 3 and the F subcomplex of the coatomer COPI. *J. Mol. Evol.* 48:770–78
107. Schwartz JH. 1999. Homeobox genes, fossils and the origin of species. *Anat. Rec.* 257:15–31
108. Semple C, Wolfe KH. 1999. Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. *J. Mol. Evol.* 48:555–64
109. Siddall ME, Kluge AG. 1997. Probabilism and phylogenetic inference. *Cladistics* 13:313–36
110. Sidow A. 1992. Diversification of the Wnt gene family on the ancestral lineage of vertebrates. *Proc. Natl. Acad. Sci. USA* 89:5098–102
111. Sidow A. 1996. Genome duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.* 6:715–22
112. Smith NGC, Knight R, Hurst LD. 1999. Vertebrate genome evolution: a slow shuffle or a big bang? *BioEssays* 21:697–703
113. Soltis PS, Soltis DE, Chase M. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* 402: 402–4
114. Song Y, Fambrough D. 1994. Molecular evolution of the calcium-transporting ATPases analyzed by the maximum parsimony method. See Ref. 28a, pp. 271–83
115. Soullier S, Jay P, Poulat F, Vanacker

J-M, Berta P, Laudet V. 1999. Diversification patterns of the HMG and SOX family members during evolution. *J. Mol. Evol.* 48:517–27

116. Spring J. 1997. Vertebrate evolution by interspecific hybridisation: are we polyploid? *FEBS Lett.* 400:2–8
117. Suga H, Koyanagi M, Hoshiyama D, Ono K, Iwabe N, et al. 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. *J. Mol. Evol.* 48:646–53
118. Swofford DL. 1999. PAUP*: Phylogenetic Analysis Using Parsimony (* and Other Methods) (software). Sunderland, MA: Sinauer
119. Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. 1996. Phylogenetic inference. In *Molecular Systematics*, ed. DM Hillis, C Moritz, BK Mable, pp. 407–509. Sunderland, MA: Sinauer. 2nd ed.
120. Tannenbaum DM, Wang Y, Williams SP, Sigler PB. 1998. Crystallographic comparisons of the estrogen and progesterone receptor's ligand binding domains. *Proc. Natl. Acad. Sci. USA* 95:5998–6003
121. Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. *Science* 278:631–37
122. Tautz D. 1998. Evolutionary biology: debatable homologies. *Nature* 395:17–19
123. Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673–80
124. Thornton JW, DeSalle R. 2000. A new method to localize and test the significance of incongruence: detecting domain-shuffling in the nuclear receptor superfamily. *Syst. Biol.* 49(2):183–201
125. Thornton JW, Kelly DB. 1998. Evolution of the androgen receptor: structure-function implications. *BioEssays* 20:860–68
126. Wheeler WC, Gatesy J, DeSalle R. 1995. Elision: a method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. *Mol. Phylogenet. Evol.* 4:1–9
127. Wheeler WC, Gladstein DS. 1995. Malign 2.7 (software). New York: Am. Mus. Nat. Hist.
128. Wilkie TM, Gilvert DJ, Olsen AS, Chen XN, Amatruda TT, et al. 1992. Evolution of the mammalian G protein alpha-subunit multigene family. *Nat. Genet.* 1:85–92
129. Yokoyama S, Radlwimmer FB. 1999. The molecular genetics of red and green color vision in mammals. *Genetics* 153: 919–32

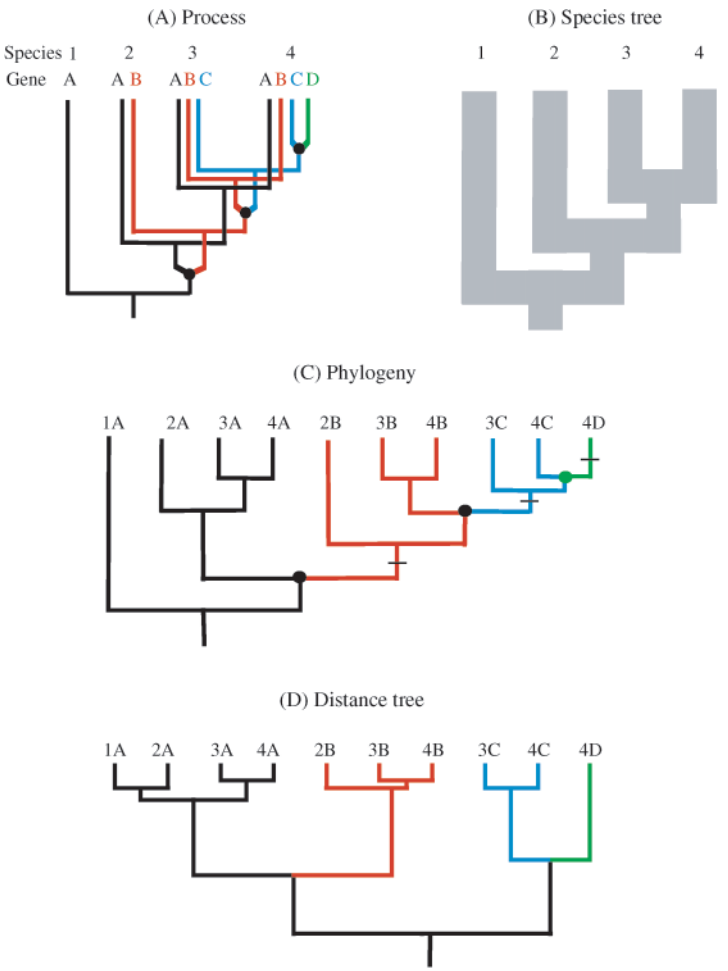


Figure 1 Cladistic and phenetic reconstructions of a gene family phylogeny. A. Evolutionary scenario of gene duplications (marked with *dark circles*) and cladogenesis (*unmarked nodes*) that generates a hypothetical gene family. Each group of colored branches leads to a group of similar orthologs. B. Species tree for the process in A. C. Gene phylogeny for the same process, correctly inferred by using the parsimony criterion. D. UPGMA tree for the process in A, assuming no homoplasy and 10-fold-higher rates of sequence change on branches on which gene duplications lead to new paralogous genes (hashed on the phylogeny in C) than on branches leading to conserved orthologs. This tree does not accurately represent evolutionary relationships.

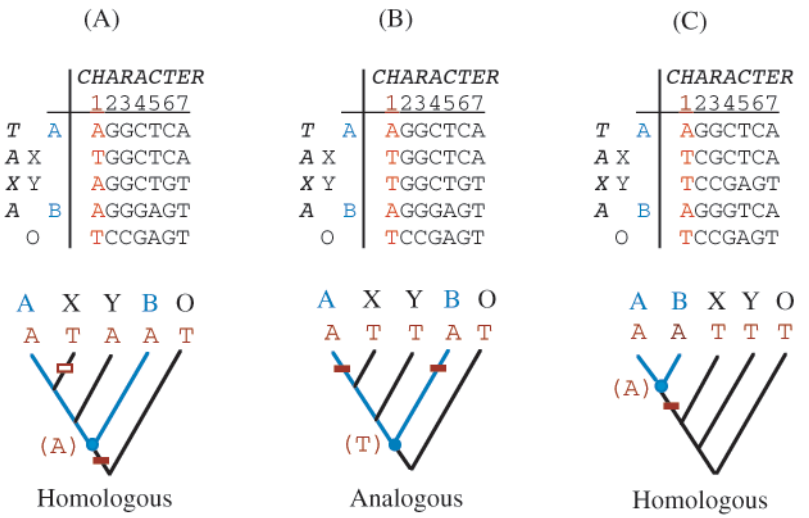


Figure 2 Homology assessment must include more characters and taxa than those being tested. By definition, the hypothesis that state A for character number 1 (*red*) is homologous in species A and B (*blue*) implies that the common ancestor of A and B (*blue circle*) had state A. Testing this hypothesis requires enough taxa to support the reconstruction that the ancestor did not have state A under at least some combinations of character states and enough characters to resolve the phylogeny of these taxa. Each tree shows the most parsimonious phylogeny for the data given, with the most parsimonious reconstruction of state changes for character 1. Gain of the state A is represented as a *filled box* and losses as *open boxes*. Trees A and B show that the outcome of homology assessment for character 1 in species A and B depends on the state of that character in other taxa. Trees C and D show that the outcome depends on the phylogeny inferred from all available characters.

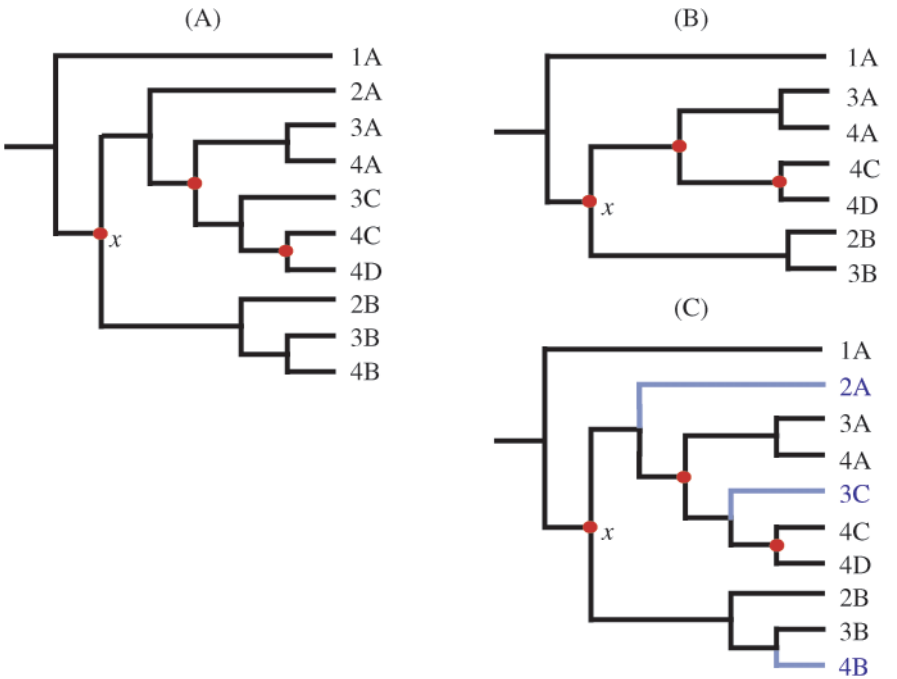


Figure 3 Inferring gene duplications and losses from a gene family tree. A. Gene duplications must be inferred at nodes where the gene tree is incompatible with the species tree (*red circles*). B. Gene duplications can be inferred even when some sequences are lost or missing. C. Reconciled tree for gene tree in B, with hypothetical branches leading to lost or missing sequences shown in *blue*.

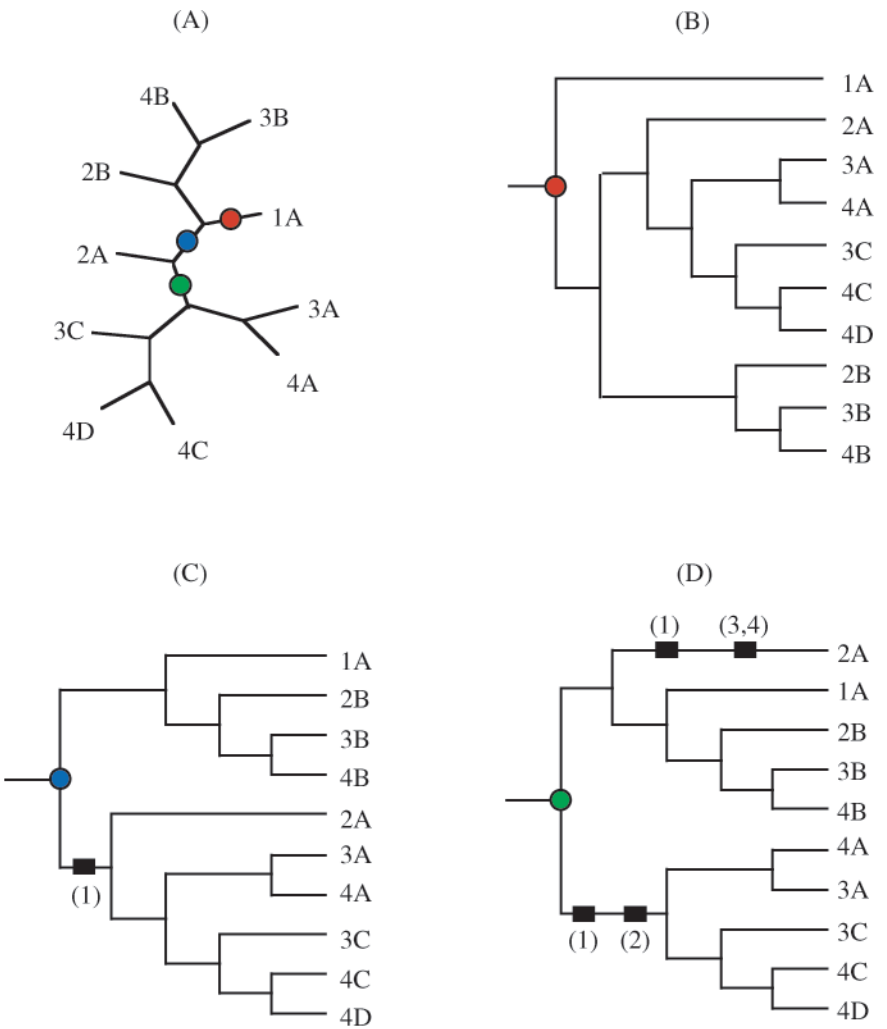


Figure 5 Inferring the root of a gene tree by minimizing ad hoc hypotheses of gene losses. A. Unrooted gene family tree. B–D. Gene trees rooted on branches labeled in A. Gene losses that must be hypothesized are marked with *dark boxes*; labels show the taxa in which the gene has been lost. Tree B is the most parsimoniously rooted tree.

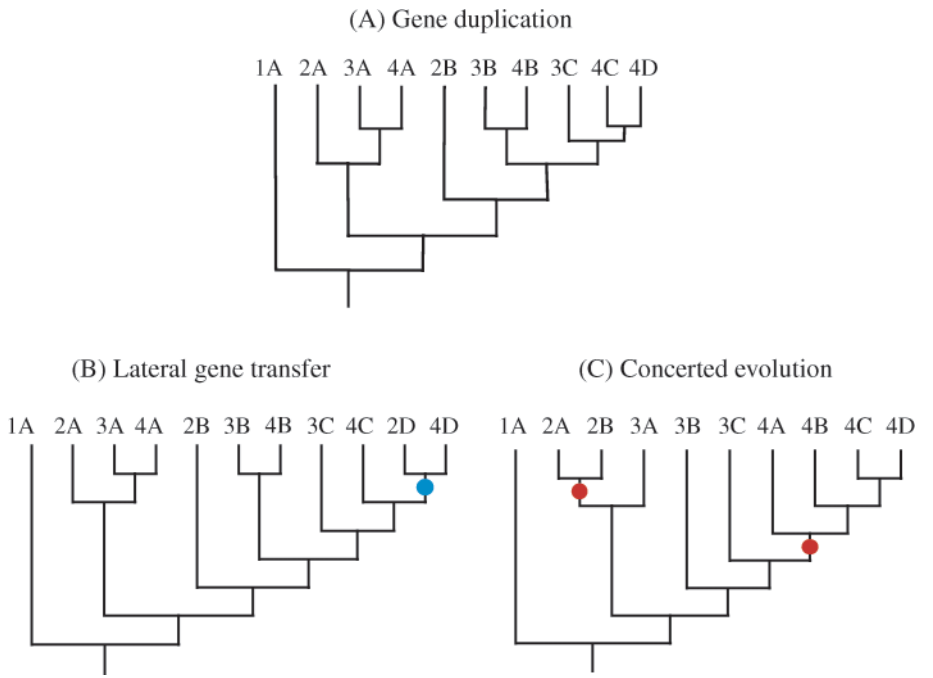


Figure 6 Lateral gene transfer and concerted evolution cause distinct patterns of phylogenetic incongruence. A. Hypothetical gene family phylogeny without lateral gene transfer or concerted evolution. B. Inferred gene phylogeny from the same process but with lateral gene transfer of gene D from taxon 4 to taxon 2. C. Inferred gene phylogeny from the same process but with concerted evolution in species 2 and 4. *Blue* and *red circles* mark nodes that are incongruent with the phylogeny in A due to lateral gene transfer and concerted evolution, respectively.