Plant Molecular Biology **42:** 1–23, 2000.

Evolution of genes and taxa: a primer

Jeff J. Doyle[1] and Brandon S. Gaut[2]
[1]*L. H. Bailey Hortorium, 466 Mann Library Building, Cornell University, Ithaca, NY 14853, USA (e-mail: jjd5@cornell.edu);* [2]*Dept. of Ecology and Evolutionary Biology, 321 Steinhaus Hall, U.C. Irvine, Irvine, CA 92697-2525, USA (e-mail: bgaut@uci.edu)*

Key words: homology, molecular population genetics, multigene families, phylogenetic methods, rates of molecular evolution

Abstract

The rapidly growing fields of molecular evolution and systematics have much to offer to molecular biology, but like any field have their own repertoire of terms and concepts. Homology, for example, is a central theme in evolutionary biology whose definition is complex and often controversial. Homology extends to multigene families, where the distinction between orthology and paralogy is key. Nucleotide sequence alignment is also a homology issue, and is a key stage in any evolutionary analysis of sequence data. Models based on our understanding of the processes of nucleotide substitution are used both in the estimation of the number of evolutionary changes between aligned sequences and in phylogeny reconstruction from sequence data. The three common methods of phylogeny reconstruction – parsimony, distance and maximum likelihood – differ in their use of these models. All three face similar problems in finding optimal – and reliable – solutions among the vast number of possible trees. Moreover, even optimal trees for a given gene may not reflect the relationships of the organisms from which the gene was sampled. Knowledge of how genes evolve and at what rate is critical for understanding gene function across species or within gene families. The Neutral Theory of Molecular Evolution serves as the null model of molecular evolution and plays a central role in data analysis. Three areas in which the Neutral Theory plays a vital role are: interpreting ratios of nonsynonymous to synonymous nucleotide substitutions, assessing the reliability of molecular clocks, and providing a foundation for molecular population genetics.

Introduction

Molecular systematics and evolutionary biology are dynamic disciplines, with their own research goals, journals, and jargon. The complexity of these fields can be daunting to those who do not routinely employ their methods, and it is apparent that molecular biologists occasionally misapply the tools of the disciplines. One simple example is the term 'homology', which is persistently misused in molecular biological literature. Another example are phylogenies based on molecular data; papers in even the best molecular biology journals present gene trees whose methods of construction are not specified and that may not be robust. Such trees may constitute a poor basis for interpretation and discussion.

This chapter provides a short primer on what we consider to be some of the key concepts in molecular evolution and systematics. Our goal is to help dispel some of the confusion over the basic principles of molecular evolution for a target audience of plant molecular biologists. We make no effort to be comprehensive in our citations, seeking instead to highlight important (and often controversial) issues and to point readers in the direction of some key references. More exhaustive treatment of these topics is available from a variety of texts and reviews (e.g. [81, 132]). Additional resources in this issue include the paper by Soltis and Soltis, which provides information about phylogenetic analyses with large data sets, and the paper by Muse, which discusses statistical aspects of molecular evolutionary analysis.

This paper begins with a discussion of the concept of homology as applied to genes, gene families and nucleotide sequences. One important aspect of homology is sequence alignment, which we briefly discuss. We then describe the importance of unobserved nucleotide substitutions and the use of models of molecular evolution that seek to account for them. These models play an important role in how gene phylogenies are inferred from nucleotide sequences; the process of phylogeny reconstruction is described in some detail. After this description, we consider obstacles to inferring species' relationships from phylogenetic trees based on nucleotide sequences. We then turn to Neutral Theory, which is an important theoretical construct underlying the study of population genetics and molecular evolution, focusing on three uses of the Neutral Theory as a source of prediction. We conclude with a brief exposition of the importance of the integration of evolutionary and molecular biology.

Homology, orthology and paralogy

Homology is a central concept in evolution and systematics – indeed in all of biology. Prior to the theory of evolution, the term was applied to convey structural or functional commonality, such as organs that performed similar functions in different organisms. The term eventually assumed an evolutionary meaning, that of similarity due to common ancestry.

Genes are homologous, therefore, if they are derived from the same gene in a common ancestor. Homology is an all-or-nothing concept – either two genes are homologous or they are not [112]. In the molecular biology literature, the term homology (or 'homologous') is often mis-used as a synonym for 'similar'. For example, when two aligned DNA sequences are identical at 90% of their nucleotide sites, a researcher will report that they are '90% homologous'. Strictly speaking, the phrase '90% homologous' implies that 90% of the nucleotide sites have shared a common ancestor but that the remaining 10% are evolutionarily unrelated. This could be true if the genes in question each had two functional domains but shared only one of them [51], but this is rarely the intended meaning. Usually the intended meaning is that the two DNA sequences are homologous over their entire length but that 10% of the bases have diverged and are no longer identical. If this is the intended meaning, it is more correct to say that the sequences are '90% identical'. For protein sequences, where amino acids can be classified into functional groups with similar chemical or structural properties, 'identity' can be distinguished from 'similarity'. Neither similarity nor identity is synonymous with homology, however.

The concept of homology extends to multigene families. All the members of a multigene family are homologous whether they were sampled from a single species or from several species. They are homologous because they derive from a common ancestor – i.e., a single gene in a single common ancestor. It is important to recognize, however, that divergence between any two homologous genes in a multigene family can be traced backward to one of two kinds of events. If the event that generated the two genes was a duplication event, then the two genes are *paralogous* [34] (Figure 1); for example, genes coexisting within the same genome and representing different subfamilies of a gene family are paralogous. In contrast, *orthologous* genes are derived from speciation events. Members of a single subfamily that are found in different species and are derived from the same duplicate copy are orthologous (Figure 1). As we shall see, an understanding of orthology and paralogy is crucial for interpreting multigene family data.

The coexistence of genes from the same gene family in a single genome is definitive evidence of paralogy, but orthology is more difficult to determine. Orthology can be hypothesized from commonality of function between genes of the same gene family in two different species, but this functional approach is not foolproof because paralogues may show functional convergence. Orthology is determined more convincingly by reference to an explicit phylogenetic hypothesis for the genes in question, from which it can be demonstrated that genes of different species belong to the same subfamily (Figure 1B). An additional source of evidence for hypothesizing orthology is to demonstrate shared chromosomal position and linkage relationships (synteny) between species. Ideally, determination of orthology should rely on both phylogeny and synteny, but orthology has rarely been proven with such rigor (but see [127]). Syntenic approaches to deducing orthology will likely become more common with the growth of plant genomic research.

When paralogues evolve in a strictly divergent manner after a duplication event, there is a clear distinction between orthologous and paralogous genes. However, paralagous copies do not evolve separately in many gene families [4]. Instead, in extreme cases such as the nuclear ribosomal gene families (nrDNA), paralogues evolve in concert, so that each of the often

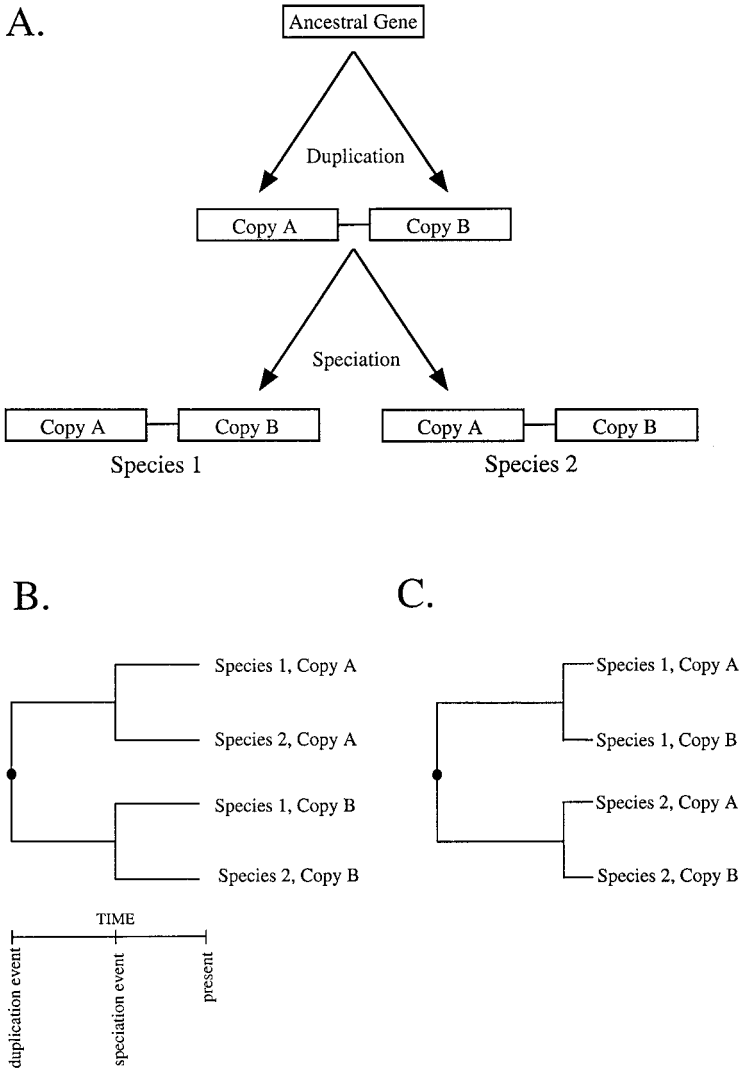


Figure 1. A. A diagram of paralogous and orthologous events. Gene duplication leads to two paralogous gene copies A and B. After speciation, both copies are present in both species. The copy A sequences are orthologous because they diverged by speciation and not duplication. The copy B sequences are also orthologous. B. A gene phylogeny of the paralogous and orthologous sequences in the absence of gene conversion, showing how orthologous copies reflect the pattern of speciation and paralagous copies date back to the time of the duplication event. C. The effect of gene conversion on the phylogeny of the multi-gene family shown in B. Gene conversion has caused paralagous copies to be each other's closest relative.

thousands of copies within an individual is identical or nearly so [3]. This 'concerted' evolution is not confined to highly repetitive ribosomal genes; homogenization is also observed among the copies of small multigene families of protein-coding genes such as *rbcS* [92]. Mechanisms hypothesized to be responsible for this concerted evolution are unequal crossing over, particularly in tandemly repeated families such as the nrDNA, and gene conversion, which can operate even across non-homologous chromosomes. When a locus fully homogenized by concerted evolution is compared across species, all of the paralogues within a species appear as each others' closest relatives in a gene tree (Figure 1C).

Full concerted evolution, on the one hand, and retention of paralogy/orthology relationships (in which there is no gene conversion) on the other, are two endpoints of a continuum of evolutionary possibilities. The plant actin gene family represents an example of one end of the continuum, because orthology relation-

ships have been retained in this gene family over long evolutionary periods in most species. Even so, gene conversion is evident among some members of the gene family in maize [94]. The fact that both divergence and concerted evolution may occur in the same gene family is a complicating factor in reconstructing and interpreting gene phylogenies [120].

Alignment and multiple hits

Evolutionary analysis of DNA sequence data begins with the alignment of two or more sequences that are at least provisionally hypothesized to be homologous (either paralogous or orthologous, in the case of gene family members). Alignment involves determining which positions along the DNA or protein sequence are derived from a common ancestral position – in other words, which *positions* are homologous within a set of homologous genes. It has been argued that without alignment, the DNA sequences deposited in GenBank are, from an evolutionary perspective, a random collection of nucleotides [76]. Despite its fundamental importance, alignment remains perhaps the most difficult and poorly understood step in molecular evolutionary analysis.

Alignment would be simple if sequences only diverged by base substitution, or even if they diverged only by insertions and deletions (indels). In the first case, alignment would be trivial no matter how divergent the sequences; in the second case, gaps could be added to either sequence to preserve complete identity of aligned positions. Unfortunately, sequences evolve both by substitution and by length mutation, and the gaps that must be added to a sequence to achieve a reasonable alignment cannot be read from a sequencing gel the way one reads nucleotides. Instead, gaps must be hypothesized and inserted according to some formula that minimizes the number and length of gaps but also maximizes the overall number of nucleotide matches. There is little to guide the researcher in this process [141], because the balance between substitution and length mutation varies with the gene, the gene region, and the organism in question. The same gene may evolve primarily by length mutation in one species, but by substitution in another [124].

The philosophies and methods employed in alignment range from the pairwise algorithmic methods of Needleman and Wunsch [99] that form the core of most multiple sequence alignment programs, to Markov Chain Monte Carlo methods [24], to methods

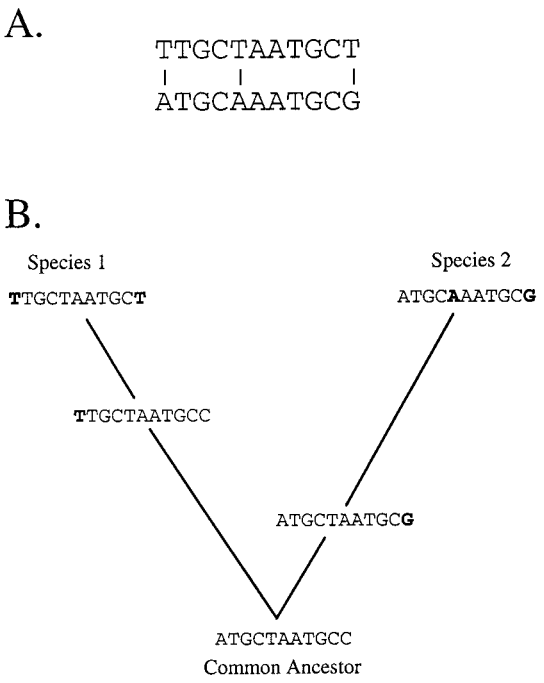


Figure 2. A. Two aligned sequences; lines between sequences indicate observable differences. B. The evolution of the two sequences, which shows the history of substitution events. Bold letters represent substituted sites. There were a total of 4 substitution events after the divergence of the two sequences from their common ancestor. Two of the substitutions occurred in the 3′ nucleotide site, where the ancestral state 'C' was substituted to 'T' in one lineage and to 'G' in the second lineage.

that simultaneously consider alignment and phylogeny [49, 147]. It is not clear which of these approaches is best. Furthermore, all of the approaches require a good deal of trial and error. When it comes to alignment, it is wise to test various combinations of gap and substitution cost parameters, and it may also be wise to manually adjust the resulting alignments. There is no magic substitute for a researcher's knowledge and intuition about the molecule(s) under study. Familiarity with conserved motifs or secondary structure, for example, can guide researchers to 'reasonable' alignments [76]. Although it is difficult to give concrete general advice about alignment, it is clear that it is unwise to accept uncritically alignments produced with the default settings of any computer algorithm.

Once it is understood that an alignment is a hypothesis about homology, it is important to recognize that alignments do not report all of the evolutionary change between DNA sequences. Figure 2A provides an example of two aligned homologous sequences. With three differences between them, the sequences

are identical at 73% (8 out of 11) sites. On a superficial level, the three differences imply that three – and only three – nucleotide substitutions have occurred since the sequences last shared a common ancestor, but this is not necessarily the case. As a source of comparison, we have provided the 'true' evolutionary history of the sequences in Figure 2B. (Of course, a true evolutionary history is unknowable in practice, but it is a helpful construct for the purposes of illustration.) The 'true' history of the sequences reveals that the 3′ nucleotide site has experienced two substitutions over time, resulting in a total of four nucleotide substitutions between the sequences since they shared a common ancestor. The occurrence of two or more substitutions at the same site is known as a *multiple hit* or *superimposed substitution*. When there have been multiple hits, the number of observed differences between sequences is always an underestimate of the true number of evolutionary changes that have taken place. In general, the greater the number of substitutions observed between two sequences, the greater the number of unobserved multiple hits there have been.

How does one correct for multiple hits? One must model the stochastic nature of DNA sequence substitutions, using the tools of probability and statistics. Multiple hits are explicitly built into these probability models. The simplest model of nucleotide substitution is that of Jukes and Cantor [66]. In addition to assuming that multiple hits can occur, the Jukes-Cantor model assumes that any nucleotide – either A, C, G, or T – can be substituted by any other nucleotide with equal probability. Many variations of the Jukes-Cantor model have been formulated, including models that permit different probabilities of change among bases [70, 137], models that assume that some nucleotide sites evolve more rapidly than others [150], and models that explicitly partition amino acid altering (nonsynonymous or 'replacement') nucleotide substitutions from non-amino-acid altering (synonymous or 'silent') nucleotide substitutions [101, 43, 95]. A detailed derivation of these models can be found in various sources [153, 132, 81], and the models are also explained more fully in the paper by Muse.

Insertions and deletions can also undergo multiple changes over time, and it is possible, in theory, to model the evolution of insertions and deletions. In practice, however, it has proven difficult to model the behavior of insertions and deletions [139, 2], and hence they are not usually incorporated into models of DNA evolution.

Is it necessary to correct for multiple hits? The answer to this question depends on context. For example, the parsimony method of phylogenetic inference does not explicitly correct for multiple hits, and yet simulation studies show that the parsimony method identifies the correct phylogenetic tree under many evolutionary conditions [57, 125]. Thus, it may not always be important to correct for multiple hits when the goal is to construct a phylogenetic hypothesis, although we should note that two other prominent phylogenetic methods – distance methods and likelihood methods – employ substitution models that attempt to correct for multiple hits (see below for an exposition on phylogenetic methods). In other contexts, it is essential to correct for multiple hits. For example, it is inappropriate to make conclusions about rates of change in molecular sequences without correcting for multiple hits. Without such a correction, the rate of change in the sequences will always be underestimated (see below for a discussion of rates of change in sequences).

Substitution models have two advantages over simple counts of differences between sequences. First, they allow the estimation of the actual – not just the *observed* – number of nucleotide substitutions that have occurred between sequences. For example, when the Jukes-Cantor model is applied to the DNA sequences in Figure 2A, we estimate that 3.8 substitutions have occurred between the sequences. This estimate is known as the 'distance' between two sequences. Distances are often expressed either as the total number of estimated substitutions (in this case, 3.8 substitutions) or, more commonly, as the number of base substitutions per nucleotide site (in this case, 3.8 substitutions over 11 sites $= 0.346$ substitutions per site). It is worth noting that the Jukes-Cantor distance estimate is not correct in this case, because we know the actual number of substitutions to be 4.0 in this hypothetical example (Figure 2B). However, the estimate of 3.8 total substitutions is much closer to reality than the observed count of 3.0 (Figure 2A). Although our model-based estimate is not correct for this hypothetical example, model-based estimates are expected to converge on the correct number of substitutions when the model is accurate and the DNA sequences are long.

Second, nucleotide substitution models form the basis of statistical tests. Statistical tests can be used to address a litany of questions. One example from the literature is: do two genes evolve at significantly different rates? Two groups have recently asked this

question to learn whether downstream genes within biosynthetic pathways are more evolutionarily conserved than either upstream genes in the same pathway [111] or the genes that regulate the pathway [108]. To address the question, the researchers needed estimates of rates of evolution in different genes and standard deviations of the estimates. Nucleotide substitution models provide both of these quantities. Nucleotide substitution models have statistical utility far beyond the question of genes evolving at different rates. For example, nucleotide substitution models form the basis of statistical inference in distance-based and maximum-likelihood phylogenetic methods [e.g. 58, 60].

Nucleotide substitution models are implemented in numerous software programs, including MEGA [78], PHYLIP [33], and PAUP* ([132]; Dave Swofford, Sinauer Associates, 1999). These multi-faceted evolutionary analysis programs can estimate the distance between sequences as well as estimate phylogenies.

Phylogeny reconstruction

Rooted trees, unrooted trees and searching tree space

It has been said that nothing makes sense except in the light of evolution, but it is also true that very little in evolution makes much sense without a phylogenetic context. Whether distinguishing orthology from paralogy, or creating realistic alignments, or estimating rates of evolution, ancestor-descendant relationships play an important role. Phylogenetics is a dynamic field in its own right, vibrant both with ongoing theoretical controversies and with empirical advances driven by the computer revolution. There is much more to constructing a phylogeny than just clicking the 'phylogenetic tree' button in a commercial program such as the MEGALIGN option of DNASTAR, or inputting raw sequences into the UPGMA algorithm of a handy statistical package.

The basic problem is to take a set of aligned sequences and to find a bifurcating tree that describes their ancestor-descendant relationships most accurately. An important assumption is that such a tree exists in the first place. Recombination, whether among alleles at a single locus or among paralogous loci, violates the assumption of bifurcation by bringing together gene regions with different phylogenetic histories [55, 20].

Even without recombination, the challenge of finding a best tree is daunting. For n sequences there exist $(2n-5)!/[2^{n-3}(n-2)!]$ different tree shapes (topologies) if the tree is unrooted [31]. An unrooted tree is a tree in which groupings are inferred but no direction to evolutionary change is implied (Figure 3A). The number of possible unrooted tree topologies becomes astronomical very quickly: there are only three different tree topologies that relate four sequences, nearly 1000 topologies for seven sequences, over two million topologies for 10 sequences and roughly 1×10^{27} topologies for 25 sequences. With under 55 sequences, the number of possible tree topologies is around 10^{79}, which exceeds the estimated number of electrons in the observable universe! One of the challenges of phylogenetic inference is to sample trees as thoroughly as possible throughout this dauntingly large 'tree space' in order to insure that a large proportion of the reasonable trees have been evaluated.

The number of tree topologies increases dramatically when a tree is rooted. Unlike an unrooted tree, a rooted tree implies directionality in time (Figure 3B). Such directionality is necessary to evaluate the history of the characters under study. Rooting a set of 'ingroup' sequences is usually accomplished by also including one or more sequences assumed to fall outside the ingroup. This 'outgroup' defines the base of the ingroup, identifying the sequences that branch off first within the ingroup. Rooting therefore requires the assumption that at least something is known about the relationships of the set of sequences in question. Assumptions about sequence relationships can sometimes be made relatively safely – especially with single-copy genes – and this knowledge can be used to root a tree. For example, in a taxonomic data set that includes orthologous sequences from pine and several flowering plants, the pine sequence is clearly the most distantly related, and so can be used as an outgroup to root the tree.

Rooting is not so simple for multigene families (Figure 3B, C). Consider a data set from a gene family having two paralogous loci, and each locus includes one sampled sequence from pine and one sequence from each of a few flowering plant genera. If the gene duplication that led to paralagous genes occurred before the divergence of gymnosperms and angiosperms, then the orthologous sequences of pine and flowering plants share a common ancestor that is more recent than the duplication event. The root should be placed so as to group these orthologous sequences. If the gene tree is rooted with the single pine gene, the tree would artificially group all flowering plant sequences, ignoring the fact that the earliest evolutionary

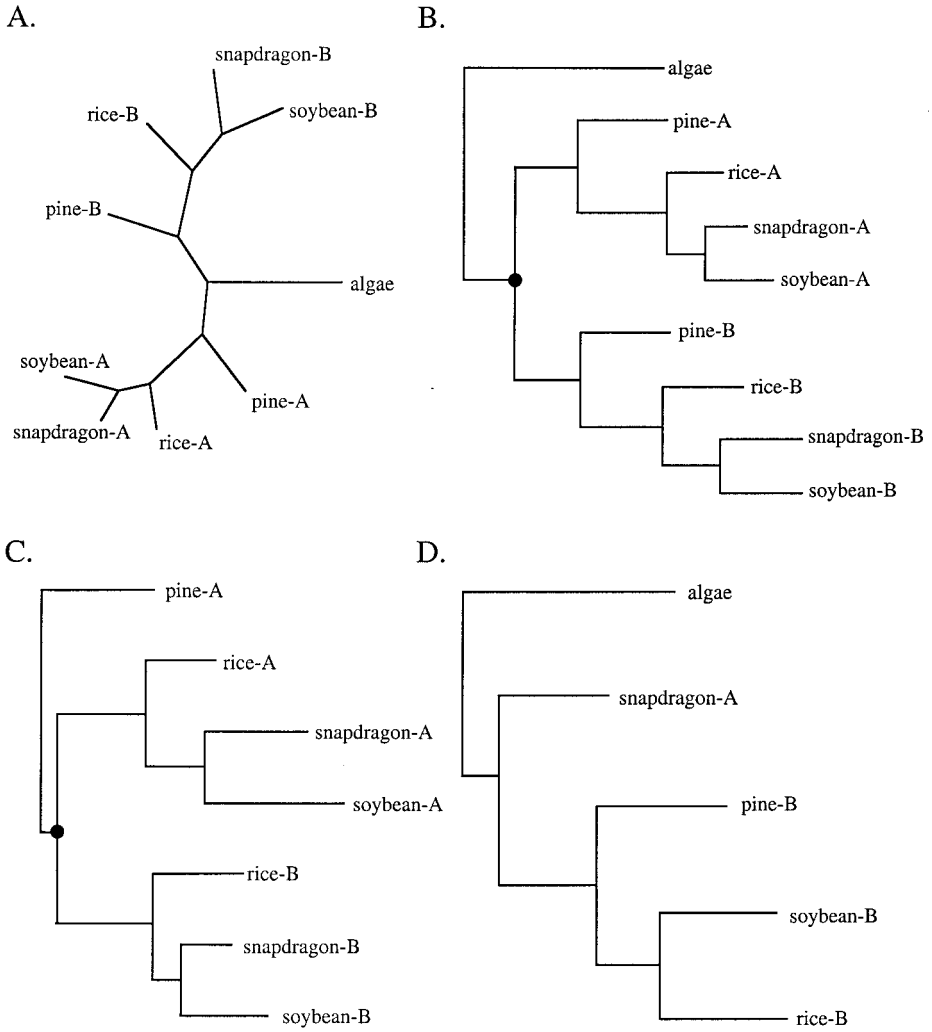


Figure 3. A. An unrooted phylogeny of a multigene family. In this phylogeny, groups can be hypothesized – for example, the soybean-A and the snapdragon-A sequences group together – but no evolutionary direction is implied. Therefore, one cannot make inferences about which nodes are old and which are more recent. B. A rooted tree showing the 'true' phylogeny of a multigene family with paralagous subfamilies A and B. A rooted tree allows inference about directionality, so that older events are deeper within the phylogeny. With the correct outgroup, the duplication of the A and B genes, as indicated by the dark circle, is properly inferred to have occurred before the divergence of the gymnosperms (pine) and the angiosperms. C. The effect of improper rooting: if a single pine sequence is assumed to be the outgroup for this phylogeny of the gene family, all of the angiosperm sequences cluster together. The duplication event is now inferred to have occurred *after* the divergence of gymnosperms and angiosperms. D. The effect of sampling paralagous genes for inferring species' phylogeny: with a mixture of paralagous and orthologous genes, snapdragons are improperly inferred to be basal to the higher plants.

event is a gene duplication rather than a speciation event (Figure 3C). Of course, if concrete knowledge of paralogy/orthology relationships for a gene family is available, this can be used for rooting, as has been done for the tree of life, where there is no taxonomic outgroup [42].

Finding a best tree in the tree space requires a criterion for judging trees and then evaluating many trees to find one or more that fulfills this criterion. Commonly used criteria are discussed at length below. The process of choosing an optimal tree varies somewhat with the criterion chosen, but all have some things in common. First, for very small data sets the obvious approach of evaluating every possible tree topology and choosing the optimal tree(s) can be taken. This exhaustive search rapidly becomes impractical, however. For slightly larger data sets (the size depends on the computer hardware, the software and attributes of

the data set, but around 30 sequences is probably a reasonable current ceiling), a branch-and-bound search [50] provides a mathematically defined shortcut that guarantees finding the optimal trees.

For most data sets, however, methods must be used that do not guarantee finding optimal trees. When using heuristic methods, it is up to the investigator to decide what constitutes a thorough search of tree space. Typical heuristic searches consist of two phases: finding a starting tree and branch-swapping. In the first stage, a tree is very quickly produced that is usually far from optimal. In the second stage, this crude first approximation is modified by mathematically defined switches of groups of sequences from one part of the tree to another. At each swap, the tree is evaluated according to the optimality criterion. If the tree is better, it is kept; if it is not better, it is discarded. Branch-swapping continues until any additional modifications to the topology result in less optimal trees. The outcome of this type of search is strongly dependent on the starting tree and on the thoroughness of the branch-swapping algorithm used. The starting tree, in turn, is often dependent on the order in which the sequences are added to the tree, and so it is common to conduct many (often a thousand or more) searches using randomized sequence entry order to ensure that many different starting trees are evaluated. In this way it is hoped that searches do not terminate prematurely on 'islands' in which suboptimal trees are surrounded by still less optimal topologies [85], and that all islands of optimal trees are identified. Heuristic search strategies are described further in the paper by Soltis and Soltis.

Phylogenetic methods

The three commonly used optimality criteria for phylogeny reconstruction from sequence data are parsimony, distance, and likelihood. Their relative merits are the focus of often acrimonious debate in the molecular evolution community. In the following sections we discuss each of these approaches briefly, with particular reference to their treatment of superimposed substitutions (multiple hits).

Parsimony

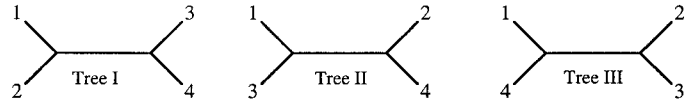
Parsimony can be viewed simply as an optimality criterion for which the optimal (most parsimonious) tree is the tree with the smallest number of mutational changes. Ideally, each possible tree topology is given a parsimony score, although, as noted above, it is not practical to evaluate every possible tree with large data sets. The score for each tree is based on the minimum number of changes in character states that are required to explain the data.

An example of parsimony inference with four sequences is provided in Figure 4. Figure 4A shows the three unrooted trees that can relate the four sequences, and Figure 4B provides an example data set. For each possible tree, the number of character state changes is computed separately for each nucleotide site, and the parsimony score for the tree is the sum of changes over all nucleotide sites. As an example, consider Figure 4C, which details the number of changes inferred for the first nucleotide site. When character states are placed on tree I (sequences 1 and 2 have an 'A', while sequences 3 and 4 have a 'G'), only one change in character state is needed to explain the data. This one change is a change of character state from 'A' to 'G' that occurred in the middle branch. Implicit in this inference is the prediction that the two nodes, which represent common ancestors to extant sequences, had character states of 'a' and 'g' (Figure 4C). Because only one change, or 'step', is needed to explain the data, tree I is given a score of '1' for the first nucleotide site. When the same character states are mapped onto trees II and III, at least two changes are required to explain the data (Figure 4C). Note that the two changes could have occurred on any pair of branches on trees II and III; in short, the location of changes in character state cannot be determined with certainty without the addition of a root. Because two changes are required, trees II and III are assigned scores of '2' for the first nucleotide site. When one applies the same method to all five sites (Figure 4B) and the results are summed across sites, tree I has a total score of 5, tree II has a total score of 7 and tree III has a score of 6. Tree I has the lowest total score, requires the fewest evolutionary steps, and is thus the most parsimonious tree.

Many proponents of parsimony defend the approach on philosophical grounds [28], because in minimizing extra steps it also minimizes the number of additional ad hoc hypotheses (parallel or reversed nucleotide substitutions, in the case of sequences) required to explain the data. In its fundamental form, parsimony with equal weighting of all characters (often called unweighted parsimony) ignores superimposed substitutions and treats multiple hits as an inevitable source of false similarity (homoplasy) that adds extra mutational steps to shortest trees.

The basic method can be modified by various forms of weighting to compensate for multiple hits.

A.



B.

		Site 1	Site 2	Site 3	Site 4	Site 5
Data	Species 1	A	A	A	A	C
	Species 2	A	T	A	T	C
	Species 3	G	T	A	T	T
	Species 4	G	A	A	T	T
Scores	Tree I	1	2	0	1	1
	Tree II	2	2	0	1	2
	Tree III	2	1	0	1	2

C.

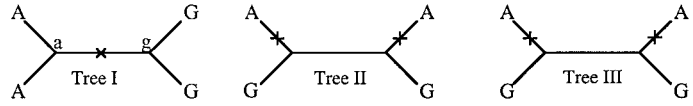


Figure 4. An example of the parsimony optimality criterion. A. Each of the three possible unrooted trees relating the four sequences. B. The table contains a hypothetical data matrix, consisting of a nucleotide sequence of 5 bases representing each of the four species and also a score for each of the three trees at each of the nucleotide sites. C. An example of scoring for the first nucleotide site; the ×'s on the trees mark hypothesized changes in character states (see text for details). Lower-case letters represent hypothesized character states at the nodes.

Character or character state changes can be weighted *a priori* to approximate the evolutionary models used explicitly in other approaches. For example, because silent substitutions usually greatly outnumber replacement substitutions and most silent substitutions occur at third codon positions, it is anticipated that most multiple hits will occur at third positions. If so, then changes at third positions may in general be less phylogenetically reliable than those at other positions (but see [105]), and they could be made to have less influence on the analysis by down-weighting. If synonymous sites are down-weighted, a mutational step at a third position site does not contribute as much to the score of the tree as a change involving a first or second codon position. Similar arguments can be made for character state changes rather than positions. If, as is often true, transitions (base changes between A and G or between C and T) are more prevalent than transversions (changes between purines and pyrimidines), then it might be assumed that most multiple hits, and hence the preponderance of misleading phylogenetic signal, involve transitions. Under these assumptions, the down-weighting of all transitions is justified.

Other weighting methods compensate implicitly for multiple hits by down-weighting all characters that show homoplasy. This can be done during tree searching [44, 45] or *a posteriori*, through iterative cycles of weighting and tree searching [27, 10]. Clearly, if multiple hits are responsible for parallelisms and reversals at particular characters, down-weighting of these characters compensates for multiple hits. Proponents of these methods point out that no particular *a priori* assumptions about the value of character transformations are required, which is useful in that no generalizations need be made about, for example, all third codon positions or all transitions being particularly labile [21].

Because parsimony algorithms count discrete steps, it is common for searches to identify numerous equally most parsimonious trees. The strict consensus of such a set of trees shows only the groups (clades) shared among all of the trees, and will therefore be poorly resolved if equally parsimonious topologies are in conflict with one another. The strict consensus tree thus is usually not itself one of the most parsimonious trees, but is useful in identifying groupings supported by all optimal trees. The majority rule consensus tree, as the name suggests, shows groupings that occur in the majority of equally most parsimonious trees. It is often far more resolved than the strict consensus tree. However, there is no particular biological significance to the majority rule tree, because all equally most par-

simonious trees, even those in the minority, represent equally tenable phylogenetic hypotheses.

Distance

The distance approach to phylogeny reconstruction begins with estimation of pairwise distances between nucleotide sequences. Pairwise distances compensate for multiple hits by transforming observed percent differences between aligned sequences into an estimate of the actual number of nucleotide substitutions, using one of the various models of molecular evolution described above. One can think of distance estimates as containing two terms: the observed dissimilarity between sequences plus a compensation term. The compensation term becomes larger as the observed dissimilarity increases, by an amount that is dependent on the model invoked.

Minimum evolution is a commonly used distance criterion for choosing an optimal tree. The minimum evolution tree is the tree for which the sum of all branch lengths is smallest. Although this may sound superficially like parsimony, the approach is not explicitly character-based, because a pairwise distance matrix, rather than changes at individual nucleotide positions, is used in the tree building process. The minimum evolution method thus does not count individual mutational steps as does parsimony. The very fast Neighbor-Joining algorithm [118] provides a good approximation of the minimum evolution tree and is available in many software packages, such as MEGA and PAUP*. However, like the heuristic search strategies described above, different results may be obtained depending on the entry order of sequences [29], and it is therefore advisable to perform several searches with random entry order.

Neighbor-Joining is preferred for distance analyses over the older and even more widely available UPGMA algorithm, because unlike UPGMA it does not assume that all the sequences evolve at the same rate. The UPGMA algorithm has been shown to be inaccurate under a wide variety of conditions (e.g. [59]) and should not be used with DNA sequences.

Pairwise distances and the branch lengths estimated from them can be calculated to many significant figures. Thus, unlike parsimony with its discrete steps, there are rarely ties among distance trees. However, it is misleading to argue that distance methods should be preferred over parsimony because the former identifies a single tree as optimal. In fact, many distance trees may be nearly equally optimal and should probably be considered as suitable phylogenetic hypotheses for a given data set. The set of equally optimal trees can be determined with statistical comparison of the length of distance trees (e.g. [115]).

Likelihood

Maximum likelihood is a family of statistical approaches commonly used throughout the biological sciences. Its application in phylogenetics involves estimating the likelihood of observing a particular set of aligned sequences given a model of nucleotide substitution and a tree topology. As with distance methods, choice of the substitution model and its particular parameters provides the means of allowing for multiple hits. However, maximum likelihood is more like parsimony in that each character (aligned nucleotide position) contributes directly and individually to the overall optimality score, in this case the probability of data given the model and tree. Given a tree topology, branch lengths are estimated according to the model and parameters chosen, and the probability of obtaining the particular character states is estimated for each nucleotide position. These probabilities are multiplied together to obtain the overall likelihood for the topology in question, which is usually presented as its negative logarithm. Like distances, these values can be expressed to several decimal places, so there is little chance of ties among nearly equally optimal trees. One advantage of likelihood methods is that a statistical test called the likelihood ratio can be used to evaluate many properties of trees [75, 58, 76]. A good introduction to the methods and philosophy of likelihood can be found in Lewis [79].

The likelihood method is very flexible, in that models can range from general to highly specific and can adjust for many sources of variation. Perhaps most importantly, nucleotide substitution models are used directly in the estimation process, rather than indirectly as in *a priori* weighted parsimony searches. Likelihood thus has the advantage of being character-based, like parsimony, but model-based, like distance methods. However, likelihood methods are computationally very intensive, much more so than either distance or parsimony. Likelihood searches are usually practical with only relatively small numbers of sequences, even using heuristic search strategies.

Robustness and reliability

Whatever the method used to construct a tree, it is common to present some numerical assessment of the reliability of the groupings it depicts. The most

commonly used method for this is the phylogenetic bootstrap [32], which simulates obtaining new data on the relationships among a group of sequences by resampling (with replacement) the same set of characters and performing a new phylogenetic analysis. This is done many (100–10 000 or more) times and a majority rule consensus tree is constructed for the resulting trees. The frequency with which particular groupings appear on the majority rule tree gives a measure of their support by the sequence data.

A number of studies have investigated the theoretical and empirical behavior of bootstrapping [154, 155, 52, 126]. These studies generally report that the bootstrap value is a conservative estimate of the level of support for a cluster of sequences. For example, Hillis and Bull [52] used computer-generated phylogenies to investigate the behavior of the bootstrap test; they found that a bootstrap value of 70% reflected an empirical cut-off wherein the group of sequences was almost always (95% of the time or better) a true phylogenetic group. Although the significance of bootstrap values varies with the data set and the method of phylogenetic inference, a bootstrap value of 70–80% is often taken to indicate strong support for a cluster of sequences.

The bootstrap is not the only method to assess reliability. For distance methods, confidence intervals can be estimated for particular groupings, using variances that are calculated along with pairwise distances [115–117]. For parsimony, the parsimony jackknife [29] is a fast alternative for identifying well supported groupings, and has recently been used with data sets of over 2000 plant *rbcL* sequences [67]. Another widely used estimate of support is the Bremer support index, also called 'decay analysis' [8, 9]. In this method, trees one step longer than the most parsimonious tree are retained along with most parsimonious trees, and the strict consensus of these trees is constructed. Comparison of this strict consensus with the strict consensus of most parsimonious trees reveals which clades collapse (decay) after one additional step. This process is continued for two steps, three steps, and so on until all clades have collapsed in the strict consensus tree. The advantage of the Bremer support index is that it can be computed without multiple searches; in contrast, bootstrapping can be a time-intensive procedure because it requires a new search with each resampled data set. The disadvantage of Bremer support statistics is that they vary substantially from data set to data set, and there is therefore no objective basis for interpretation of Bremer support indices with any given data set. For any method and any data set, it is important to apply some measure of reliability, because the measures can indicate which groupings within a tree have strong support.

There have also been efforts to compare different methods of phylogeny reconstruction, mainly using computer simulations (e.g. [57]) and occasionally using experimental biological systems with known phylogenies (e.g. [15]). In general, all of the methods are fundamentally reliable for most data sets. Some molecular phenomena cause problems for all methods, but in many cases the problem can be alleviated if an evolutionary model matching the phenomenon is used to compensate for the problem [151, 37]. Principal examples of problems are unequal amounts of divergence in different lineages [30], as can occur when DNA sequences evolve very rapidly in one lineage but not another, or radically different base compositions in different sequences of the same data set [130]. It is largely how the different phylogenetic approaches handle these extreme cases that fuels the controversy about choice of method. In general, however, it is commonly observed that although different methods may identify different topologies as optimal, the differences among these topologies usually involve poorly resolved groupings. Groupings strongly supported in the optimal tree from one method often seem to be robust to choice of method.

Inferring the relationships of organisms from molecular data: incongruence among phylogenies

Gene trees vs. species trees

For many molecular phylogenetic studies the goal is to infer species relationships. This adds a layer of complexity to the problem of resolving tree topologies, because a perfectly good gene tree may not faithfully depict relationships among the species from which the genes were sampled. Three common sources of incongruence between gene trees and taxon (species) trees are: (1) mixing paralogous and orthologous sequences, (2) introgression of genes among species, and (3) sorting of ancestral polymorphisms (for reviews see e.g. [100, 146]).

As detailed above, orthologous sequences in a group of taxa (species, genera, etc.) track speciation events, whereas paralogous sequences trace the history of the gene duplication event. Thus, only orthologous sequences can be used to infer taxonomic relationships. Consider the case of an ancient gene duplication

that took place in the ancestor of higher plants (Figure 3B), with two paralogous sequences (genes A and B) sampled from each of four species. The reconstructed phylogeny for the higher-plant sequences shows a divergence between A and B sequences, with two identical subtrees: one for paralogue A and one for paralogue B. Each subtree is a faithful depiction of species' relationships, but consider what happens when a mixture of A and B sequences is sampled (Figure 3D). Sampling the A paralogue from snapdragon and the B paralogue from pine, soybean and rice produces a tree with soybean and rice, rather than soybean and snapdragon, as sister species (Figure 3D). This is clearly incorrect, yet the tree itself is in no way wrong as a gene tree – the soybean and rice B paralogues do indeed share a more recent common ancestor than either does with the A paralogue of snapdragon. This example illustrates that it is difficult to place much confidence in a gene tree as representative of species relationships without rigorous demonstration of orthology among sequences.

Introgression and sorting of ancestral polymorphisms are common problems at lower taxonomic levels, and both produce similar patterns [102]. Introgression is a familiar phenomenon to plant breeders, but it also occurs in nature (see the paper by Rieseberg *et al.* in this volume). Hybridization and recurrent backcrossing in one direction move one or more genes from one species into another. For the systematist, the problem is that an introgressed gene moves physically but not phylogenetically. For example, a resistance gene from *Lycopersicon pennellii* introgressed into *L. esculentum* remains a *L. pennellii* gene. A phylogeny reconstructed from this locus will place *L. esculentum* with *L. pennellii* regardless of the true relationships among these and other *Lycopersicon* species.

The sorting of ancestral polymorphisms (also known as 'lineage sorting') is a stochastic process that is predicted from population genetic theory. Consider a species with alleles A and B at a locus (Figure 5). This species gives rise to two daughter species, one of which (species 1) inherits only allele A whereas the other inherits both alleles. This polymorphic daughter species in turn gives rise to two daughter species, one of which (species 2) inherits only allele A whereas the other (species 3) inherits only allele B. Clearly, species 2 and 3 share a more recent common ancestor than either does with species 1. However, if sequences from this locus are used to reconstruct the phylogeny of these species, the resulting tree will show species

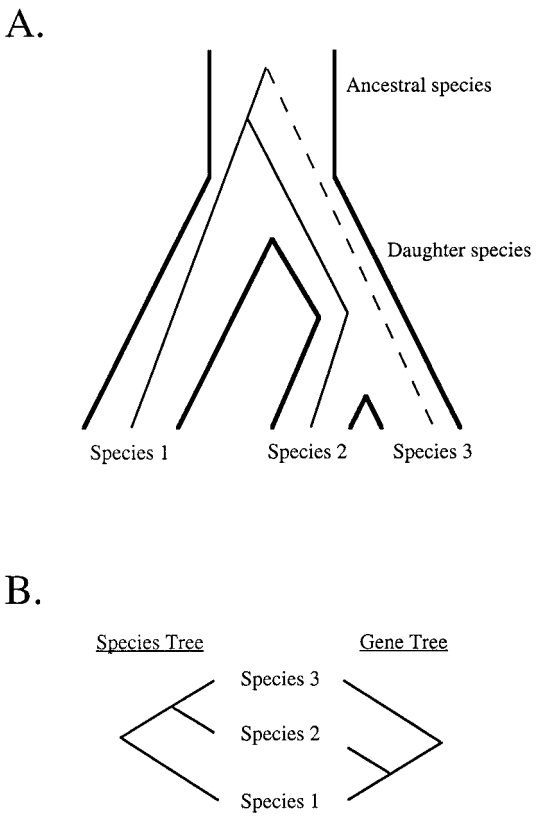


Figure 5. A. An example of lineage sorting. The thick outer lines represent the genetic boundaries of species; the thin inner lines represent lineages of alleles at one locus. The solid thin line represents the evolutionary history of allele A, while the dashed line represents the history of allele B. B. The effect of lineage sorting on phylogenetic inference. Because of lineage sorting, the sequences in species 1 and 2 are more closely related than the sequences in species 3, but the organismal history is such that species 2 and 3 are more closely related to each other than to species 1.

1 and 2 as sister species, because they both share allele A. The ancestral polymorphism has been sorted into the three descendant species in such a way as to be phylogenetically misleading as to species relationships, though it is a perfectly accurate depiction of the relationships of the alleles themselves. It is worth noting that the pattern of incongruence could also be explained by introgression among taxa; it is generally very difficult to distinguish between lineage sorting and introgression [142].

Incongruence among phylogenetic trees

It is not uncommon for systematists to find that different genes – or, alternatively, molecular vs. morphological data sets – do not yield identical phylogenies. In one sense, incongruence among the topologies of

different trees is inconvenient, and incongruence is often seen as an error that must be accounted for. However, topological disagreement can arise from biologically meaningful phenomena, and therefore can provide important insights when understood properly [146]. For example, incongruence between a chloroplast gene tree and the trees for several nuclear genes may suggest that introgression has occurred involving only the chloroplast genome [114]. This is significant for inferring organismal phylogenies, and also potentially for understanding the evolution of interactions between nuclear and organellar genes.

Disagreements between gene topologies and organismal phylogenies are critical in understanding the homologies of gene families. It is a basic premise in evolutionary biology – and therefore in any comparative approach – that commonality of function across species is underlain by commonality of genes. This can be rephrased in evolutionary terms to state that structural and functional homologies across species should involve orthologous and not paralogous genes. The difficulties inherent to identifying true orthologues have already been discussed, but an underlying obstacle is undersampling of most plant genomes. Many gene families have been studied largely by the accumulation of cDNA sequences from studies with diverse goals, often on different tissues. This leads to the likelihood that many family members are not represented in gene trees. Even if comprehensive attempts are made to identify all members of a gene family, the possibility exists that some will be missed. This will be true even when full genome sequences are available, because genes may be lost in the course of evolution.

The fact that some genes are 'missing' may be hypothesized by comparing gene trees with other phylogenetic information. As noted above, orthologous genes should track organismal phylogenies. If discrepancies are observed between the relationships among gene family members for several species and the relationships of those species, then one explanation is that some genes have not been identified. In the above example of duplicate genes in rice, soybean, snapdragon and pine (Figure 3), full sampling of both paralogues in all four taxa should produce a tree from which the correct relationships can be inferred from either subtree. However, if one is confident in the taxonomic relationships, and in the quality of the molecular data, then it may be possible to infer, from a partially sampled set of genes, which members have *not* yet been sampled from which taxa. This can be critical knowledge for using information from a well-studied model system to guide the search for orthologues underlying a phenomenon of interest in a less-studied species.

Of course, it is also possible that the true orthologue may never be found. Redundancy of genes, in some (many?) cases due to the polyploid nature of most plants (see the paper by Wendel in this volume), means that some members of gene families are likely to have become pseudogenes. Such genes are expected to accumulate substitutions at the neutral rate, meaning that over time they may not retain a great deal of sequence similarity with homologous functional members of their family. Moreover, such genes also accumulate deletions and insertions, hastening their divergence from homologues, and in some cases pseudogenes may be eliminated from the genome entirely. This type of gene absence is also of interest, however, because the complete absence of an orthologue to a gene of known critical function leads to the hypothesis that some other gene has been recruited to perform this function. Such a gene may be paralogous or perhaps not even homologous; finding such a gene may reveal much about such phenomena as pleiotropy and epistasis.

For simple cases, it is often apparent where genes are missing. For more complex situations, where the relationships among the plants themselves may be unclear, methods are being developed that use the principle of parsimony to hypothesize duplications, gene losses, and speciation events simultaneously [106]. The hypothesis of incongruence should be tested rigorously in all but those cases where relationships are so clear as to make this trivial. Various phylogenetic and statistical methods are available for testing the significance of incongruence between two data sets or the trees inferred from them [65].

Neutral Theory and the tempo and mode of molecular evolution

The Neutral Theory

Phylogenetic inference is crucial for understanding the tempo and mode of evolutionary change as well as for documenting evolutionary relationships. Most studies of tempo and mode are explicitly designed to provide information about the pattern and strength of natural selection. Much of this inference relies on comparing real data to predictions that are based on the Neutral Theory of molecular evolution. Because the Neutral Theory plays such a critical role in molecular evolutionary analysis, we will briefly explain some aspects

of the Neutral Theory before outlining methods to examine the tempo and mode of molecular evolution.

The Neutral Theory was postulated independently by Kimura [68] and King and Jukes [73]. The theory was radical because it stipulated that most evolutionary change was invisible to natural selection and therefore evolutionarily neutral. The fate of neutral mutations is determined by the stochastic process known as random genetic drift. Under the process of genetic drift, a new neutral mutation – which is typically found in only one individual in a population of individuals – will usually be lost to evolution, but occasionally by chance a neutral mutation can become the predominant variant in a population.

Contrary to common misconception, the Neutral Theory of molecular evolution neither ignores natural selection nor constitutes an alternative to natural selection. The Neutral Theory recognizes that negative selection or selective constraint is a potent force in the evolution of genes [71]. Negative selection is the culling of deleterious mutations from populations. The most obvious example of negative selection is that of the lethal mutant, which cannot contribute to the next generation by virtue of its lethality. Kimura also recognized that positive (or adaptive) mutations must occur and can motivate evolutionary change. The radical aspect of Neutral Theory was the assertion that positive selection events are rare relative to the number of evolutionary changes brought about by neutral mutations and random genetic drift [72]. On the whole, this assertion has proven to be true at the molecular level.

The Neutral Theory of molecular evolution was formulated before the advent of DNA sequence data, yet it was remarkably prescient. For example, the Neutral Theory predicted that rates of nonsynonymous nucleotide substitutions would generally be slower than rates of synonymous nucleotide substitution [69]. The neutralist interpretation of this phenomenon is that most synonymous mutations are neutral (or nearly-neutral) because they do not change amino acids and therefore have little or no effect on gene function. In contrast, most nonsynonymous mutations are removed by negative selection pressure because the mutations have a detrimental impact on protein function. The Neutral Theory also predicted that: (1) gene regions of functional importance should evolve more slowly than less important regions and (2) duplicated genes should differ in their evolutionary rate [72]. These predictions have been confirmed by a number of studies.

The Neutral Theory is quantitatively tractable and makes many predictions that apply both to population genetics and to molecular evolutionary analyses. These predictions form the basis of null hypotheses that can be tested; in fact, the most valuable role of Neutral Theory may be its role as a null hypothesis. In the following sections, we outline the use of the Neutral Theory as a null hypothesis for studies of molecular evolution, with particular emphasis on detecting the effects of natural selection at the molecular level.

Rates of nonsynonymous and synonymous evolution as a measure of adaptive evolution

One interesting prediction of Neutral Theory concerns the ratio of nonsynonymous to synonymous nucleotide substitutions. This ratio can provide important evidence about gene function, and can provide insights into specific amino acid residues that are functionally important. We will refer to the ratio as d_n/d_s, where d_n is the distance estimate of the number of nonsynonymous substitutions separating two sequences and d_s is the distance estimate of the number of synonymous substitutions between sequences.

In pairwise comparisons between most homologous protein coding sequences, d_n/d_s is less than 1.0. The neutralist explanation for this phenomenon is that most nonsynonymous substitutions are deleterious and are thus retained infrequently relative to synonymous substitutions. A d_n/d_s ratio less than 1.0 is often used as evidence to argue that a gene evolves with constraint on amino acid replacements and is therefore functional. On the other hand, if replacing amino acids has no effect on the function of the protein, then both nonsynonymous and synonymous substitutions are completely neutral. In this case, d_n/d_s will equal 1.0 because selection does not discriminate between nonsynonymous and synonymous substitutions. Finally, if there is adaptive selection for nonsynonymous substitutions, so that it is advantageous when the gene experiences nonsynonymous substitutions, then d_n/d_s can be greater than 1.0. A d_n/d_svalue greater than 1.0 is commonly taken as evidence that the protein is under diversifying selection for increased amino acid diversity.

The latter phenomenon has been characterized, but very rarely. The classic example of $d_n/d_s > 1.0$ is the MHC class I antigen recognition site [61]. The adaptive advantage to diversity at the antigen-recognition site is clear: the greater the diversity at the receptor

site, the better the ability to recognize, bind and defend against a broad array of pathogens. In fact, a high d_n/d_s ratio appears to be a general feature of pathogen and defense interactions [25], including plant-pathogen interactions [98] (also see the paper by Richter and Ronald in this volume). In this context, identification of the amino acid residues with high d_n/d_s ratios can provide evidence of the regions of molecules that are contact points for pathogen-defense interactions [103]. High d_n/d_s ratios have also been found in systems that do not play a role in pathogen interactions**,** such as vertebrate lysozyme [93], but it appears that high d_n/d_s ratios are rare in genes that are not involved in pathogen interactions.

The characterization of d_n/d_s ratios constitutes a test for adaptive selection events that is largely based on the expectations of Neutral Theory, but it is important to interject two notes of caution regarding the test. First, the test for $d_n/d_s > 1.0$ detects only diversifying selection, which constitutes a limited subset of adaptive selection events. Based solely on d_n/d_s ratios, it would be inappropriate to conclude that there has been *no* adaptive selection between sequences but rather that there have not been enough adaptive events to elevate d_n/d_s higher than 1.0. For example, changes to a few key residues at the active site of a protein could be strongly adaptive, but these changes would represent too small a fraction of the overall change to elevate d_n substantially. The second word of caution has to do with testing statistically whether d_n/d_s is greater than 1.0. Simply observing that d_n/d_s is greater than 1.0 is not sufficient to demonstrate adaptive selection, because the variation inherent in estimating d_n/d_s will occasionally inflate the ratio higher than 1.0. The lack of a statistical approach has been evident in the plant molecular biology literature [107]. Statistical methods to test whether $d_n/d_s > 1.0$ are available [103, 152] and have been applied to plant sequences [98, 143].

The molecular clock and the study of rates of molecular evolution

The molecular clock hypothesis

The molecular clock was born of empirical observation but gained an underlying theoretical basis from Neutral Theory. The empirical observations came from the early 1960s, when researchers were first able to compare homologous amino acid sequences from different species. These comparisons revealed that the number of differences between amino acid sequences varied roughly linearly with the time of divergence between species [156, 86, 157], suggesting that amino acid replacements were accruing at a regular 'clock-like' rate over time. These observations led to the formulation of the 'molecular clock' hypothesis [157], which posits that either nucleotide substitutions or amino acid replacements occur at a regular rate *per year*. An important corollary prediction of the molecular clock hypothesis is that rates of molecular evolution are equal among diverse evolutionary lineages.

The molecular clock hypothesis has had important implications for the study of evolutionary phenomena, and it can be especially useful for estimating divergence times between taxa in the absence of a fossil record. For example, Sarich and Wilson [123] used a molecular clock argument to hypothesize that man diverged from other higher primates about 5 million years ago. This estimate was four-fold lower than contemporary divergence estimates based on the fossil record, but subsequent paleontological work has confirmed the estimate based on the molecular clock argument. Molecular clock arguments have also been used to estimate the origin of angiosperms and the date of the monocot-dicot divergence [89, 148]. It is worth noting, however, that molecular clock estimates do not always agree with each other or with fossil-based estimates.

The early work of Sarich and Wilson [123] illustrates the potential utility of the molecular clock for dating evolutionary events. Because of this potential, a good deal of effort has been invested into testing the clock hypothesis, but there are two additional reasons for examining the rates at which macromolecules – particularly DNA sequences – change over time. The first additional reason to study rates of evolution is because molecular clocks are a central issue in debates over the mode of molecular evolution. The Neutral Theory provides theoretical justification for a molecular clock by stating that a molecular clock is expected if rates of mutation are constant *per year* in different evolutionary lineages [71]. Subsequent arguments have claimed that mutation rates in different evolutionary lineages should be constant *per generation* rather than per year. Under these conditions, Neutral Theory predicts that there should be a generation-time clock, wherein organisms with fast generation times have fast rates of molecular evolution [104, 149]. Both time-calibrated and generation-time-calibrated molecular clocks are consistent with Neutral Theory. However, some have criticized the biological assumptions underlying the neutral argument [41], and there

is still debate as to whether it is reasonable to expect molecular clocks to hold. It is thought that empirical characterization of molecular clocks will shed light on these debates and make it possible to infer the mechanisms underlying evolutionary change.

The extent to which either the time-calibrated clock or the generation-time clock holds is unclear, despite extensive empirical efforts [149, 82, 23, 83]. In plants, it appears that time-calibrated molecular clocks do not hold, but there may be generation-time clocks [7, 38, 36, 40]. In addition to molecular clocks based on time and generation-time, researchers have also hypothesized that rates of molecular change correlate with metabolic rates [88, 1, 87] and rates of speciation [91, 7]. The eventual resolution of the behavior (or the complete lack) of molecular clocks will rely on continued empirical characterization of nucleotide substitution rates.

A second and equally important reason to study rates of change in DNA sequences is that they can provide insights into gene function. It has long been known that amino acids in structurally important protein regions – for example, a catalytic site of an enzyme or an important functional motif – tend to evolve more slowly than amino acids in less important regions [72]. Similar arguments apply to promoter regions embedded in non-coding DNA; in some cases, promoter regions have been identified because of their slow rate of evolution [158]. Study of evolutionary rates can also help provide clues into the evolution and function of multigene families. Empirical studies have shown that different members of plant gene families often evolve at different rates [145, 39, 90, 127]. Such differences could indicate slightly different functions among paralogous gene copies [138], low functional constraint on redundant gene copies [72], or different mutational dynamics in different regions of the genome.

Measuring evolutionary rates

It is instructive to define two basic measures of nucleotide substitution rates: *absolute rates* and *relative rates*. The measurement of absolute rates requires homologous nucleotide sequences from at least two taxa and an estimate of the divergence time between the taxa. Given these data, k, the absolute rate of nucleotide substitution per site per year, is estimated by

$$\hat{k} = \hat{d}/2\hat{T}$$

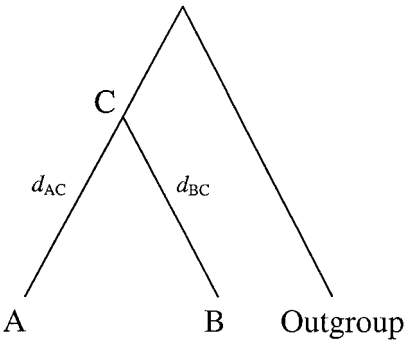


Figure 6. The three-sequence phylogeny used in the relative rate test to examine the molecular clock hypothesis. Under a time-calibrated molecular clock, the amount of evolution in the lineages leading from the common ancestor (C) to species A and B should be equal. The outgroup is needed to root the tree.

where $\hat{d}$ is the estimated number of nucleotide substitution events per nucleotide site between homologous sequences, and $\hat{T}$ is the estimated divergence time between taxa. The estimation of d is usually based on the application of nucleotide substitution models, and the divergence time T must be estimated from fossil records or from indirect measures of divergence times such as vicariance events. An example of an absolute rate estimate comes from the *adh1* gene in grasses. The fossil record suggests that maize and rice diverged 50 million years ago [148], and the number of DNA substitutions between maize and rice *adh1* sequences is estimated to be 0.53 substitutions per synonymous site, using the distance model of Kimura to estimate distances [70]. Thus, the rate of *adh1* evolution in this example is estimated to be 5.3×10^{-9} substitutions per synonymous site per year.

Three characteristics of absolute rate estimates deserve comment. First, estimates of k can be compared among independent evolutionary lineages. For example, *adhC* in cottons has been measured to evolve at a synonymous rate of 1.5×10^{-9} substitutions per site per year [128], suggesting that *adhC* in cottons has evolved more slowly than *adh1* in grasses. Second, in many cases it is not possible to estimate absolute rates because gaps in the fossil record preclude reasonable estimation of T. Third, absolute rates can be averaged across nucleotide sites, but substitution rates commonly vary among sites. This is especially important for rates of nonsynonymous and synonymous nucleotide substitutions; it is often wise to employ a nucleotide substitution model that allows separate estimation of nonsynonymous and synonymous rates.

In the absence of fossil time estimates, rates of evolution can be compared between evolutionary lineages by the *relative rate* method. The relative rate method does not lead to an estimate of k, the number of nucleotide substitutions per site per year, but it does provide a framework for testing the time-calibrated molecular clock. There are many variations on the relative rate method, but the simplest approach requires three homologous nucleotide sequences: two ingroup sequences (represented by sequences A and B in Figure 6) and an outgroup sequence. Given the phylogeny of these three sequences (Figure 6), rates of nucleotide substitution can be compared between the evolutionary lineages leading to the two ingroup sequences. Specifically, the relative rate between ingroup sequences from taxa A and B is defined as

$$\hat{r} = \frac{\hat{d}_{\mathrm{AC}}}{\hat{d}_{\mathrm{BC}}}$$

where $\hat{r}$ represents the relative rate estimate, and $\hat{d}_{\mathrm{AC}}$ and $\hat{d}_{\mathrm{BC}}$ represent the estimated number of substitution events on the branches leading from the common ancestor C to ingroups A and B, respectively. It is important to note that the parameter r is independent of the time dimension, because both d_{AC} and $\mathrm{d_{BC}}$ are functions of the time of divergence between taxa A and B; when the ratio is taken, divergence time cancels out.

When r is not significantly different from one, then rates of evolution are similar in two evolutionary lineages, and it can be reasonably concluded that sequence divergence is linearly related to time. In other words, if $r = 1$ then there is evidence for a time-calibrated molecular clock.

The null hypothesis that $r = 1$ (or, more exactly, that $d_{\mathrm{AC}} = d_{\mathrm{BC}}$) is testable by the relative rate test [123]. There have been several implementations of the test for application to nucleotide sequences, and we refer readers to these publications for greater detail [149, 80, 97, 134, 95, 136, 46]. It should also be noted that the lack of a time-calibrated molecular clock makes it difficult, but not impossible, to use nucleotide substitution rates to estimate divergence dates between taxa [119, 140].

Selection at the population level: molecular population genetics

Molecular population geneticists focus on measuring the amount and pattern of genetic diversity in a species or population. There are several reasons to measure genetic diversity, but the primary reason is to investigate the strength and effects of natural selection. Neutral Theory plays an important role in this task, because Neutral Theory serves as the null hypothesis to test against actual data. When the data do not fit the predictions of Neutral Theory, it is often appropriate to infer that genetic variation has been affected by natural selection.

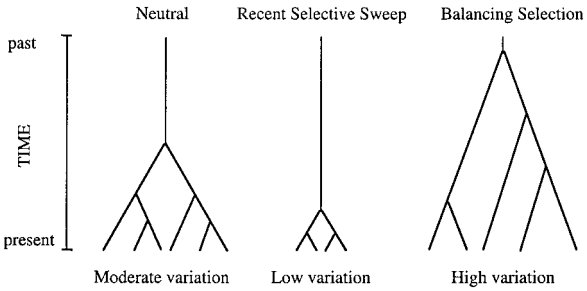


Figure 7. Example of intra-population genealogies, wherein each branch on the tree represents the lineage of an allele in the population. The figure represents a genealogy sample from each of three loci from the same species; each locus is assumed to have the same mutation rate. The diagram conveys the dependence of the depth of the genealogy (in either time or genetic variation) as a function of the type of natural selection acting on the locus.

The *modus operandi* of molecular population genetics is to sample genetic diversity at the DNA level, usually by sequencing the same gene from several different individuals of the same species. The gene may be chosen for any of several reasons. For example, in a cultivated plant there may be interest in genetic diversity among alleles of a gene that encodes an important agronomic trait [144]. Researchers may also want to contrast genetic diversity between genes from different chromosomal regions [5, 22]. Whatever gene(s) are chosen, DNA sequence data provide a wealth of information about the amount of genetic diversity in the gene, the frequency of variants in the sample and the genealogical (or phylogenetic) relationships among alleles.

A genealogy is a summary of the genetic or phylogenetic relationships among sequences (alleles) drawn from a single gene or locus (Figure 7). Each branch in the genealogy represents an allelic lineage. The branching structure of the genealogy is shaped by evolutionary processes such as mutation, the effective (or long-term) population size of the species, the pattern and strength of selection acting on the locus, recombination, and demographic factors. One important feature of a genealogy is the depth in time of the

deepest node (Figure 7). This time-depth is a measure of the persistence of allelic diversity.

Coalescent theory provides a theoretical framework to study features of genealogies [74, 55]. For example, coalescent theory predicts that the time depth of the genealogy is expected to be roughly four times the long-term population size of the species, when many sequences are sampled and the gene is evolving neutrally [74]. The time-depth of the genealogy is closely related to the amount of genetic diversity in the genealogy, because long-lived allelic lineages have more time to accrue genetic mutations.

Selection can affect the amount of genetic diversity at a locus, and this, in turn, affects the depth of the genealogy. There are many kinds of selection to consider, but we will discuss only two here: selective sweeps and balancing selection. The two kinds of selection have opposite effects on both the amount of variation within a locus and on the time-depth of the genealogies sampled from that locus.

A selective sweep refers to the fixation of an adaptive allele. As the adaptive allele is driven to fixation by selection, it 'sweeps away' genetic variation that is not linked to the adaptation. A gene that has experienced a recent selective sweep has less variation than an unlinked 'neutral' gene, and it also has a more shallow genealogy (Figure 7). One way to test for a selective sweep is to compare genetic diversity between loci. The first test of this type was the Hudson, Kreitman and Aguade (or HKA) test [56], which has been applied widely. Some tests for selective sweeps do not rely on comparisons between loci but instead use information about the frequency distribution of variants within the sample of sequences [133, 35]. Altogether, there is good evidence that selective sweeps do occur (e.g. [47, 54]), but it appears that they occur infrequently. Thus far, there are few examples of selective sweeps in plant genes, but two have been documented in maize genes that appear to have been associated with domestication [48, 144].

In contrast to selective sweeps, balancing selection acts to maintain genetic variation, with the net result that loci under balancing selection tend to have high levels of diversity and deep genealogies relative to neutral loci (Figure 7). Classic examples of loci under balancing selection include *Drosophila adh* [77] and the self-incompatibility alleles of solanaceous plants [64, 12]. One feature of loci undergoing balancing selection is that allelic lineages can be very long-lived (e.g. [64]). This feature has been utilized to make inferences about population events in the ancient past [135, 113]. The paper by Richman and Kohn in this volume describes studies of the genealogy of sequences from a locus under balancing selection, and Richman and Kohn use these genealogies to try to infer demographic events (population bottlenecks) during the evolution of species.

One of the more interesting findings of molecular population genetics is that the level of DNA sequence diversity varies throughout the genome as a function of recombination rate. In *Drosophila*, for example, loci near centromeres tend to have low recombination rates and also tend to have low levels of genetic diversity, but both recombination rate and genetic diversity increase toward the tips of chromosomes [5]. The relationship between diversity and recombination is not because recombination is mutagenic, rather it reflects an interdependence between natural selection and recombination [5, 11]. In regions of low recombination, linkage between nucleotide sites ensures that selection for or against a single nucleotide substitution will affect a large region of the genome. In regions of high recombination, nucleotide sites are nearly independent, so selection on a single site affects a much smaller region of the genome. The net result of the interdependence between selection and recombination is that: (1) levels of genetic diversity can be a function of chromosomal position and (2) large chromosomal regions can be genetically depauperate. These phenomena have been documented in plants as well as *Drosophila* [22, 131], but there is still much to learn about the dynamics of genome evolution as it relates to diversity.

Thus far, the molecular population genetics of plants have been most studied in maize and arabidopsis [63, 109, 110]. These two systems make an interesting contrast, because their differences in breeding system (outcrossing vs. inbreeding) result in different patterns and types of genetic variation. The contrast between breeding systems has also been made in other taxa [14, 84]. Molecular population genetic approaches have also proven fruitful for studying the effects of selection [48, 144] and population bottlenecks [26, 53] during the domestication of maize.

The future: an integration of molecular biology and evolutionary analysis

Perhaps the most important interface between molecular genetics and evolutionary biology is achieving an understanding of the molecular basis of phenotypic

change. With the advent of genomic technologies, most of the genes of a few select taxa will be identified. The challenge will be to elucidate the function of these genes; to this end, comparative – and therefore evolutionary – approaches will prove important.

Examples of the comparative approach to elucidate function are becoming more frequent, but the body of work by Doebley and coworkers constitutes a particularly instructive example of the interplay of molecular and evolutionary approaches. The work began in the background of debates over the origin of maize. The phenotypic differences between maize and its wild relatives are pronounced [62], which had made it difficult to positively identify the wild relatives of maize. The application of molecular systematic approaches largely resolved the debates, because molecular phylogenetic studies of maize and its wild relatives clearly indicate that maize is closely related to one particular *Zea* taxon (for a review, see [16]).

Given the close relationship of maize and its wild relatives, the next logical question was: What are the genes responsible for the huge phenotypic differences between maize and its wild relatives? Doebley and coworkers used a comparative quantitative genetics approach to answer this question. They crossed maize with its wild ancestor and discovered that five quantitative trait loci (QTL) segregating in the crosses explained most of the phenotypic variation between the two taxa [17]. They followed this work with attempts to isolate the QTLs. Using molecular approaches – most notably, genetic screens with transposable element-induced mutations – they isolated the *tb1* gene [19], which was previously shown to contribute to the phenotypic differences between maize and its wild ancestor [18].

Molecular analysis has shown that the *tb1* gene acts as a repressor of organ growth. The gene is upregulated in the lateral-branch primordia of maize, resulting in short lateral branches in maize relative to its wild ancestors [19]. However, functional studies alone do not prove the role of *tb1* in domestication; since domestication is a historical event, an evolutionary approach can provide additional evidence of the role of *tb1*. The reasoning for an evolutionary investigation of *tb1* is as follows: if *tb1* was important to domestication, then the gene was under strong selection by the domesticators. If it was under strong selection, then the *tb1* locus should contain much less genetic variability than maize loci that were not under selection. A recent population genetic study has shown that this gene contains little genetic variation and has experienced a selective sweep associated with domestication [144]; this sweep is consistent with the role of *tb1* in domestication. Surprisingly, this selective sweep covers only the promoter region of the gene, suggesting that selection during domestication focused on a regulatory variant (rather than a protein variant) at the *tb1* locus [144]. Altogether, the study of genes conferring domestication-associated traits has demonstrated the power and utility of molecular genetics coupled with evolutionary analysis.

Other examples of the interplay between molecular genetics and evolutionary analysis are beginning to surface. For example, Bennetzen and Kellogg [6] recently mapped a phenotypic character (genome size) onto a phylogeny of the grass family. They used this analysis to argue that genome size fluctuations are largely unidirectional, with an evolutionary trend toward ‘genomic obesity’. This argument was especially compelling in the background of Bennetzen and coworker’s studies of retrotransposon distributions in the maize genome [129, 121]. These studies suggest that retrotransposon activity has been rampant in the maize genome during the past 3–6 million years and may, in fact, have led to a doubling of the size of the maize genome [122]. Arguments about retrotransposon activity relied extensively on molecular evolutionary tools to estimate the time of retrotransposon insertions. Thus, evolutionary analyses have provided substantial insight into the structure of grass genomes.

As this volume attests, evolutionary approaches have become particularly important in the study of the function and structure of multi-gene families. With the continued production of plant genomic data (such as genomic sequence and EST data), studies of gene families will expand dramatically to include more paralagous family members and more taxa. The role of evolutionary analysis will be to achieve a better understanding of the tempo and pattern of gene family diversification, including the forces that shape retention and loss of gene family members [13].

The explosion of genomic studies in arabidopsis, rice, maize, cotton and other model systems are providing unprecedented opportunities to wed molecular genetic results – for example, sequence data, comparative microarray data and structural genomic data – to evolutionary analysis. The union of evolutionary genetics and molecular genetics has the potential to provide countless insights into the evolutionary patterns underlying phenotypic function. However, the success of this marriage will rely on the cooperation of

molecular biologists and evolutionary biologists, each with an appreciation for the importance of the other's approach. The realization of the power of joint evolutionary and molecular genetic approaches in plant biology is just in its infancy; with continued cultivation, this joint approach will prove to be a powerful tool in the new millenium.

References

1. Adachi J, Cao Y, Hasegawa M: Tempo and mode of mitochondrial DNA evolution in vertebrates at the amino acid sequence level – rapid evolution in warm-blooded vertebrates. J Mol Evol 36: 270–281 (1993).
2. Allison L, Wallace CS: The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and to optimization of multiple alignments. J Mol Evol 39: 418–430 (1994).
3. Appels R, Honeycutt RL: rDNA: evolution over a billion years. In: Dutta SK (ed.), DNA Systematics, vol. 2, pp. 81–135. CRC Press, Boca Raton, FL (1986).
4. Arnheim N: Concerted evolution of multigene families. In: Nei M, Koehn RK (eds), Evolution of Genes and Proteins, pp. 38–61. Sinauer Associates, Boston (1983).
5. Begun DJ, Aquadro CF: Levels of naturally occurring DNA polymorphism correlate with recombination rates in *Drosophila melanogaster*. Nature 356: 519–520 (1992).
6. Bennetzen JL, Kellogg EA: Do plants have a one-way ticket to genomic obesity? Plant Cell 9: 1509–1514 (1997).
7. Bousquet J, Strauss SH, Doerksen AH, Price RA: Extensive variation in evolutionary rate of *rbc*L gene sequences among seed plants. Proc Natl Acad Sci USA 89: 7844–7848 (1992).
8. Bremer K: The limits of amino-acid sequence data in angiosperm phylogenetic reconstruction. Evolution 42: 795–803 (1988).
9. Bremer K: Branch support and tree stability. Cladistics 10: 295–304 (1994).
10. Carpenter JM: Successive weighting, reliability, evidence. Cladistics 4: 291–296 (1994).
11. Charlesworth D, Charlesworth B, Morgan MT: The pattern of neutral molecular variation under the background selection model. Genetics 141: 1619–1632 (1995).
12. Clark AG, Kao T-H: Excess nonsynonymous substitution at shared polymorphic sites among self-incompatibility alleles of Solanaceae. Proc Natl Acad Sci USA 88: 9823–9827 (1991).
13. Clegg MT, Cummings MP, Durbin ML: The evolution of plant nuclear genes. Proc Natl Acad Sci USA 94: 7791–7798 (1997).
14. Cummings MP, Clegg MT: Nucleotide sequence diversity at the alcohol dehydrogenase 1 locus in wild barely (*Hordeum vulgare* ssp. *spontaneum*): an evaluation of the background selection hypothesis. Proc Natl Acad Sci USA 95: 5637–5642 (1998).
15. Cunningham CW, Zhu H, Hillis DM: Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. Evolution 52: 978–987 (1998).
16. Doebley J: Molecular evidence and the evolution of maize. Econ Bot 44: 6–27 (1990).
17. Doebley J, Stec A, Wendel J, Edwards M: Genetic and morphological analysis of a maize-teosinte F2 population – implications for the origin of maize. Proc Natl Acad Sci USA 87: 9888–9892 (1990).
18. Doebley J, Stec A, Gustus C: *Teosinte branched 1* and the origin of maize: evidence for epistasis and the evolution of dominance. Genetics 141: 333–346 (1995).
19. Doebley J, Stec A, Hubbard L.: The evolution of apical dominance in maize. Nature 386: 485–488 (1997).
20. Doyle JJ: Trees within trees: genes and species, molecules and morphology. Syst Biol 46: 537–553 (1997).
21. Doyle JJ, Davis JI: Homology in molecular phylogenetics: a parsimony perspective. In: Soltis DE, Soltis PS, Doyle JJ (eds), Molecular Systematics of Plants, 2nd ed., pp. 101–131. Kluwer Academic Publishers, Dordrecht, Netherlands (1998).
22. Dvorak J, Luo M-C, Yang Z-L: Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing *Aegilops* species. Genetics 148: 423–434 (1998).
23. Easteal S, Collet C, Betty D: The Mammalian Molecular Clock. R.G. Landes, Austin, TX (1995).
24. Eddy SR: Hidden Markov models. Curr Opin Struct Biol 6: 361–365 (1996).
25. Endo T, Ikeo K, Gojobori T: Large-scale search for genes on which positive selection may operate. Mol Biol Evol 13: 685–690 (1996).
26. Eyre-Walker A, Gaut RL, Hilton H, Feldman DL, Gaut B.S.: Investigation of the bottleneck leading to the domestication of maize. Proc Natl Acad Sci USA 95: 4441–4446 (1998).
27. Farris JS: A successive approximations approach to character weighting. Syst Zool 18: 374–385 (1969).
28. Farris JS: The logical basis of phylogenetic analysis. In: Platnick NI, Funk VA (eds), Advances in Cladistics 2, pp. 7-36. Columbia University Press, New York (1983).
29. Farris JS, Albert VA, Källersjö M, DL, Kluge AG: Parsimony jackknifing outperforms neighbor-joining. Cladistics 12: 99–124 (1996).
30. Felsenstein J: Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27: 401–410 (1978).
31. Felsenstein J: The number of evolutionary trees. Syst Zool 27: 27–33 (1978).
32. Felsenstein J: Confidence limits in phylogenies: an approach using the bootstrap. Evolution 39: 783–791 (1985).
33. Felsenstein J: PHYLIP Manual. University Herbarium, University of California, Berkeley, CA (1990).
34. Fitch WM: Distinguishing homologous from analogous proteins. Syst Zool 19: 99–113 (1970).
35. Fu Y-X, Li W-H: Statistical tests of neutrality of mutations. Genetics 133: 693–709 (1993).
36. Gaut BS: Molecular clocks and nucleotide substitution rates in higher plants. Evol Biol 30: 93–120 (1997).
37. Gaut BS, Lewis PO: Success of maximum likelihood phylogeny inference in the four-taxon case. Mol Biol Evol 12: 152–162 (1995).
38. Gaut BS, Muse SV, Clark WD, Clegg MT: 1992. Relative rates of nucleotide substitution at the *rbc*L locus of monocotyledonous plants. J Mol Evol 35: 292–303 (1992).
39. Gaut BS, Morton BR, McCaig BM, Clegg MT: Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. Proc Natl Acad Sci USA 93: 10274–10279 (1996).

40. Gaut BS, Clark LG, Wendel JF, Muse SV: Comparisons of the molecular evolutionary process at *rbc*L and *ndh*F in the grass family (Poaceae). Mol Biol Evol 14: 769–777 (1997).
41. Gillespie JH: On Ohta's hypothesis: most amino acid substitutions are deleterious. J Mol Evol 40: 64–69 (1995).
42. Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ: Evolution of the vacuolar H^+ ATPase: implications for the origin of eukaryotes. Proc Natl Acad Sci USA 86: 6661–6665 (1989).
43. Goldman N, Yang ZH: Codon-based model of nucleotide substitution for protein coding DNA sequences. Mol Biol Evol 11: 725–736 (1994).
44. Goloboff PA: Estimating character weights during tree search. Cladistics 9: 83–91 (1993).
45. Goloboff PA: Tree searches under Sankoff parsimony. Cladistics 14: 229–237 (1998).
46. Gu X, Li W-H: Bias corrected paralinear and LogDet distances and tests of molecular clocks and phylogenies under nonstationary nucleotide frequencies. Mol Biol Evol 13: 1375–1383 (1996).
47. Guttman DS, Dykhuizen DE: Detecting sweeps in naturally occurring *Escherichia coli*. Genetics 138: 993–1003 (1994).
48. Hanson MA, Gaut BS, Stec AO, Fuerstenberg SI, Goodman MM, Coe EH, Doebley J: Evolution of anthocyanin biosynthesis in maize kernels: the role of regulatory and enzymatic loci. Genetics 143: 1395–1407 (1996).
49. Hein J: A tree reconstruction method that is economical in the number of pairwise comparisons used. Mol Biol Evol 6: 669–684 (1989).
50. Hendy MD, Penny D: Branch and bound algorithms to determine minimal evolutionary trees. Math Biosci 59: 277–290 (1982).
51. Hillis DM: Homology in molecular biology. In: Hall BK (ed), Homology: The Hierarchical Basis of Comparative Biology, pp. 339–368. Academic Press, New York (1994).
52. Hillis DM, Bull JJ: An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst Biol 42: 182–192 (1993).
53. Hilton H, Gaut BS: Speciation and domestication in maize and its wild relatives: evidence from the *Globulin-1* gene. Genetics 150: 863–872 (1998).
54. Hilton H, Kliman RM, Hey J: Using hitchhiking genes to study adaptation and divergence during speciation within the *Drosophila melanogaster* species complex. Evolution 48: 1900–1913 (1994).
55. Hudson RR: Gene genealogies and the coalescent process. Oxford Surv Evol Biol 7: 1–44 (1991).
56. Hudson RR, Kreitman M, Aguade M: A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153–159 (1987).
57. Huelsenbeck JP: Performance of phylogenetic methods in simulation. Syst Biol 44: 17–48 (1995).
58. Huelsenbeck JP, Crandall KA: Phylogeny estimation and hypothesis testing using maximum likelihood. Annu Rev Ecol 28: 437–466 (1997).
59. Huelsenbeck JP, Hillis DM: Success of phylogenetic methods in the four-taxon case. Syst Biol 42: 247–264 (1993).
60. Huelsenbeck JP, Rannala B: Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science 276: 227–232 (1997).
61. Hughes AL, Nei M: Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335: 167–170 (1988).
62. Iltis HH: From teosinte to maize: the catastrophic sexual transmutation. Science 222: 886–894 (1983).
63. Innan H, Tajima F, Terauchi R, Miyashita NT: Intragenic recombination in the *adh* locus of the wild plant *Arabidopsis thaliana*. Genetics 143: 1761–1770 (1996).
64. Ioerger TR, Clark AG, Kao T-H: Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. Proc Natl Acad Sci USA 87: 9732–9735 (1990).
65. Johnson LA, Soltis DE: Phylogenetic inference in Saxifragaceae sensu stricto and *Gilia* (Polemoniaceae) using *mat*K sequences. Ann Miss Bot Gard 82: 149–175 (1995).
66. Jukes TH, Cantor CR: Evolution of protein molecules. In: Munro HN (ed.), Mammalian Protein Metabolism, pp. 21–32. Academic Press, New York (1969).
67. Källersjö M, Farris JS, Chase MW, Bremer B, Fay MF, Humphries CJ, Petersen G, Seberg O, Bremer K: Simultaneous parsimony jackknife analysis of 2538 *rbc*L DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. Plant Syst Evol 213: 259–287 (1998).
68. Kimura M: Evolutionary rate at the molecular level. Nature 217: 624–626 (1968).
69. Kimura M: Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 276: 275–276 (1977).
70. Kimura M: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16: 111–120 (1980).
71. Kimura M: The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, UK (1983).
72. Kimura M, Ohta T: On some principles governing molecular evolution. Proc Natl Acad Sci USA 71: 2848–2852 (1974).
73. King JL, Jukes TH: Non-darwinian evolution: random fixation of selectively neutral mutations. Science 164: 788–798 (1969).
74. Kingman JFC: On the genealogy of large populations. J Appl Prob 19A: 27–43 (1982).
75. Kishino H, Hasegawa M: Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in the Hominoidea. J Mol Evol 29: 170–179 (1989).
76. Kjer KM: Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs. Mol Phyl Evol 4: 314–330 (1995).
77. Kreitman M, Hudson RR: Inferring the evolutionary histories of *Adh* and the *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics 127: 565–582 (1991).
78. Kumar S, Tamura K, Nei M: MEGA: molecular evolutionary genetic analysis, version 1.0. Penn State University, University Park, PA 16802, USA (1993).
79. Lewis PO: Maximum likelihood as an alternative to parsimony for inferring phylogeny using nucleotide sequence data. In: Soltis DE, Soltis PS, Doyle JJ (eds), Molecular Systematics of Plants II: DNA Sequencing, pp. 132–163. Kluwer Academic Publishers, Boston (1998).
80. Li P, Bousquet J: Relative-rate test for nucleotide substitutions between two lineages. Mol Biol Evol 9: 1185–1189 (1992).
81. Li W-H: Molecular Evolution. Sinauer Associates, Sunderland, MA (1997).

82. Li W-H, Tanimura M, Sharp P: An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J Mol Evol 25: 330–342 (1987).
83. Li W-H, Ellsworth DL, Krushkal J, Chang BH-J, Emmet DH: Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. Mol Phyl Evol 5: 182–187 (1996).
84. Liu F, Zhang L, Charlesworth D: Genetic diversity in a *Leavenworthia* population with different inbreeding levels. Proc R Soc Lond B 265: 293–301 (1998).
85. Maddison DR: The discovery and importance of multiple islands of most-parsimonious trees. Syst Zool 40: 315–328 (1991).
86. Margoliash E: Primary structure and evolution of cytochrome c. Proc Natl Acad Sci USA 50: 672–679 (1963).
87. Martin AP, Palumbi SR: Body size, metabolic rate, generation time, and the molecular clock. Proc Natl Acad Sci USA 90: 4087–4091 (1993).
88. Martin AP, Naylor GJP, Palumbi SR: Rate of mitochondrial DNA evolution is slow in sharks compared to mammals. Nature 357: 153–155 (1992).
89. Martin W, Gierl A, Saedler H: Molecular evidence for pre-Cretaceous angiosperm origins. Nature 339: 46–48 (1989).
90. Mathews S, Sharrock RA: The phytochrome gene family in grasses (Poaceae): a phylogeny and evidence that grasses have a subset of loci found in dicot angiosperms. Mol Biol Evol 13: 1141–1150 (1996).
91. Mayr E: Change of genetic environment and evolution. In: Huxley J, Hardy AC, Ford EB (eds), Evolution as a Process, pp. 157–180. George, Allen and Unwin, London (1954).
92. Meagher RB, Berry-Lowe S, Rice K: Molecular evolution of the small subunit of ribulose bisphosphate carboxylase: nucleotide substitution and gene conversion. Genetics 123: 845–863 (1989).
93. Messier W, Stewart C-B: Episodic adaptive evolution of primate lysozymes. Nature 385: 151–153 (1997).
94. Moniz de Sa M, Drouin G: Phylogeny and substitution rates of angiosperm actin genes. Mol Biol Evol 13: 1198–1212 (1996).
95. Muse SV, Gaut BS: A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11: 715–724 (1994).
96. Muse SV, Gaut BS: Comparing patterns of nucleotide substitution patterns among chloroplast loci using the relative ratio test. Genetics 146: 393–399 (1997).
97. Muse SV, Weir BS: Testing for equality of evolutionary rates. Genetics 132: 269–276 (1992).
98. Myers BC, Shen KA, Rohani P, Gaut BS, Michelmore RW: Receptor-like genes in the major resistance locus of lettuce are subject to divergent selection. Plant Cell 10: 1833–1846 (1998).
99. Needleman SB, Wunsch CD: A general method applicable to the search for similarities in the amino acid sequences of two proteins. J Mol Biol 48: 443–453 (1970).
100. Nei M: Molecular Evolutionary Genetics. Columbia University Press, New York (1987).
101. Nei M, Gojobori T: Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418–426 (1986).
102. Neigel JE, Avise JC: Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. In: Karlin S, Nevo E (eds), Evolutionary Processes and Theory, pp. 515–534. Academic Press, New York (1986).
103. Nielsen R, Yang ZH: Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929–936 (1998).
104. Ohta T, Kimura M: On the constancy of the evolutionary rate of cistrons. J Mol Evol 1: 18–25 (1971).
105. Olmstead R, Reeves PA, Yen AC: Patterns of sequence evolution and implications for parsimony analysis of chloroplast DNA. In: Soltis PS, Soltis DE, Doyle JJ (eds), Molecular Systematics of Plants II: DNA Sequencing, pp. 164–187. Kluwer Academic Press, Boston (1998).
106. Page RDM: GeneTree: comparing gene and species phylogenies using reconciled trees. Bioinformatics 14: 819–820 (1998).
107. Parniske M, Hammond-Kosack KE, Golstein C, Thomas CM, Jones DA, Harrison K, Wulff BBH, Jones JDG: Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the Cf-4/9 locus of tomato. Cell 91: 821–832 (1997).
108. Purugganan MD: The molecular evolution of development. BioEssays 20: 700–711 (1998).
109. Purugganan MD, Suddith JI: Molecular population genetics of floral homeotic loci: departures from the equilibrium neutral model at the APETALA3 and PISTILLATA genes of *Arabidopsis thaliana*. Genetics 151: 839–848 (1998).
110. Purugganan MD, Suddith JI: Molecular population genetics of the *Abrabidopsis* CAULIFLOWER regulatory gene: non-neutral evolution and wild variation in floral homoeotic function. Proc Natl Acad Sci USA 95: 8130–8134 (1999).
111. Rausher MD, Miller RE, Tiffin P: Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. Mol Biol Evol 16: 266–274 (1999).
112. Reeck GR, de Haën C, Teller DC, Doolittle RF, Fitch WM, Dickerson RE, Chambon P, McLachlan AD, Margoliash E *et al.*: 'Homology' in proteins and nucleic acids: a terminology muddle and a way out of it. Cell 50: 667 (1987).
113. Richman AD, Uyenoyama MK, Kohn JR: Allelic diversity and gene genealogy at the self-incompatibility locus in the Solanaceae. Science 273: 1212–1216 (1996).
114. Rieseberg LH, Soltis DE: Phylogenetic conseequences of cytoplasmic gene flow in plants. Evol Trends Plants 5: 65–84 (1991).
115. Rzhetsky A, Nei M: A simple method for estimating and testing minimum-evolution trees. Mol Biol Evol 9: 945–967 (1992).
116. Rzhetsky A, Nei M: Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol Biol Evol 10: 1073–1095 (1993).
117. Rzhetsky A, Kumar S, Nei M: Four-cluster analysis: a simple method to test phylogenetic hypotheses. Mol Biol Evol 12: 163–167 (1995).
118. Saitou N, Nei M: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–425 (1987).
119. Sanderson MJ: A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol Biol Evol 14: 1218–1231 (1997).
120. Sanderson MJ, Doyle JJ: Reconstruction of organismal and gene phylogenies from data on multigene families: concerted evolution, homoplasy and confidence. Syst Biol 41: 4–17 (1992).
121. SanMiguel P, Tikhonov A, Jin Y-K, Melake-Berhan A, Springer PS, Edwards KJ, Avramova Z, Bennetzen JL:. Nested retrotransposons in the intergenic regions of the maize genome. Science 274: 765–768 (1996).

122. SanMiguel PJ, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL: The paleontology of intergene retrotransposons of maize: dating the strata. Nature Genetics 20: 43–45 (1998).
123. Sarich VM, Wilson AC: Immunological time-scale for hominid evolution. Science 150: 1200–1203 (1967).
124. Schaeffer SW, Aquadro CF, Anderson WW: Restriction-map variation in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. Mol Biol Evol 4: 254–265 (1987).
125. Siddal ME: Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris Zone. Cladistics 14: 209–220 (1998).
126. Sitnikova T, Rzhetsky A, Nei M: Interior-branched and bootstrap tests of phylogenetic trees. Mol Biol Evol 12: 319–333 (1995).
127. Small RL, Ryburn JA, Cronn RC, Seelanan T, Wendel JF: The tortoise and the hare: choosing between noncoding plastome and nuclear *adh* sequences for phylogeny reconstruction in a recently diverged plant group. Am J Bot 85: 1301–1315 (1998).
128. Small RL, Ryburn JA, Wendel JF: Low levels of nucleotide diversity at homoeologous *Adh* loci in allotetraploid cotton (*Gossypium* L.). Mol Biol Evol 16: 491–501 (1998).
129. Springer PS, Edwards KJ, Bennetzen JL: DNA class organization on maize *adh1* yeast artificial chromosomes. Proc Natl Acad Sci USA 91: 863–867 (1994).
130. Steel MA, Lockhart PJ, Penny D: Confidence in evolutionary trees from biological sequence data. Nature 364: 440–442 (1993).
131. Stephan W, Langley CH: DNA polymorphism in *Lycopersicon* and crossing-over per physical length. Genetics 150: 1585–1593 (1998).
132. Swofford DL, Olsen GJ, Waddell PJ, Hillis DM: Phylogenetic Inference. In: Hillis DM, Moritz C, Mable BK (eds), Molecular Systematics, pp. 407–514. Sinauer Associates, Sunderland, MA (1996).
133. Tajima F: The effect of change in population size change on DNA polymorphism. Genetics 123: 597–601 (1989).
134. Tajima F: Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135: 599–607 (1993).
135. Takahata N: Evolutionary genetics of human paleopopulations. In: Takahata N, Clark AG (eds), Mechanisms of Molecular Evolution, pp. 1–21. Sinauer Associates, Sunderland, MA (1993).
136. Takezaki, Rzhetsky A, Nei M: Phylogenetic test of the molecular clock and linearized trees. Mol Biol Evol 12: 823–833 (1995).
137. Tamura K, Nei M: Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10: 512–526 (1993).
138. Theissen G, Kim JT, Saedler H: Classification and phylogeny of MADs-box multigene family suggest defined roles of MADs-box gene subfamilies in the morphological evolution of eukaryotes. J Mol Evol 43: 484–516 (1996).
139. Thorne JL, Kishino H, Felsenstein J: Inching toward reality: an improved likelihood model of sequence evolution. J Mol Evol 34: 3–16 (1992).
140. Thorne JL, Kishino H, Painter IS: Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol 15: 1647–1657 (1998).
141. Vingron M, Waterman S: Sequence alignment and penalty choice: review of concepts, case studies and implications. J Mol Biol 235: 1–12 (1994).
142. Wakeley J: Distinguishing migration from isolation using the variance of pairwise differences. Theor Pop Biol 49: 369–386 (1996).
143. Wang G-L, Ruan D-L, Song W-Y, Sideris S, Chen L, Pi L-Y, Zhang S, Zhang Z, Fauquet C, Gaut BS, Whalen C, Ronald PC: Xa21D encodes a receptor-like molecule with a leucine rich repeat domain that determines race-specific recognition and is subject to adaptive evolution. Plant Cell 10: 765–779 (1998).
144. Wang RL, Stec A, Hey J, Lukens L, Doebley J: The limits of selection during maize domestication. Nature 398: 236–239 (1999).
145. Waters ER: The molecular evolution of the small heat-shock proteins in plants. Genetics 141: 785–795 (1995).
146. Wendel JF, Doyle JJ: Phylogenetic incongruence: window into genome history and molecular evolution. In: Soltis DE, Soltis PS, Doyle JJ (eds), Molecular Systematics of Plants II: DNA Sequencing, pp. 265–296. Kluwer Academic Publishers, Boston (1998).
147. Wheeler WC, Gladstein DS: MALIGN: a multiple sequence alignment program. J Hered 85: 417–418 (1994).
148. Wolfe KH, Gouy M, Yang Y-W, Sharp PM, Li W-H: Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc Natl Acad Sci USA 86: 6201–6205 (1989).
149. Wu C-I, Li W-H: Evidence for higher rates of nucleotide substitution in rodents than in man. Proc Natl Acad Sci USA 82: 1741–1745 (1985).
150. Yang Z: Maximum-likelihood models for combined analyses of multiple sequence data. J Mol Evol 42: 587–596 (1996).
151. Yang Z, Goldman N, Friday A: Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol Biol Evol 11: 316–324 (1994).
152. Zhang Z, Kumar S, Nei M: Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. Mol Biol Evol 14: 1335–1338 (1998).
153. Zharkikh A: Estimation of evolutionary distances between nucleotide sequences. J Mol Evol 39: 315–329 (1994).
154. Zharkikh A, Li W-H: Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. 1.4. Taxa with a molecular clock. Mol Biol Evol 9: 1119–1147 (1992).
155. Zharkikh A, Li W-H.: Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. 2.4. Taxa without a molecular clock. J Mol Evol 35: 356–366 (1992).
156. Zuckerkandl E, Pauling L: Molecular disease, evolution, and genetic heterogeneity. In: Bryson B, Vogel HJ (eds), Horizons in Biochemistry, pp. 189–225. Academic Press, New York (1962).
157. Zuckerkandl E, Pauling L: Evolutionary divergence and convergence in proteins. In: Bryson B, Vogel HJ (eds), Evolving Genes and Proteins, pp. 97–116. Academic Press, New York (1965).
158. Zurawski G, Clegg MT: Evolution of higher-plant chloroplast encoded genes: implications for structure-function and phylogenetic studies. Annu Rev Plant Physiol 38: 391–418 (1987).