TREES WITHIN TREES: GENES AND SPECIES, MOLECULES AND MORPHOLOGY

JEFF J. DOYLE

L. H. Bailey Hortorium, 462 Mann Library Building, Cornell University, Ithaca, New York 14853, USA; E-mail: jjd5@cornell.edu

Abstract.—The construction and interpretation of gene trees is fundamental in molecular systematics. If the gene is defined in a historical (coalescent) sense, there can be multiple gene trees within the single contiguous set of nucleotides, and attempts to construct a single tree for such a sequence must deal with homoplasy created by conflict among divergent histories. On a larger scale, incongruence is expected among gene tree topologies at different loci of individuals within sexually reproducing species, and it has been suggested that this discordance can be used to delimit species. A practical concern for such topological methods is that polymorphisms may be maintained through numerous cladogenic events; this polymorphism problem is less of a concern for nontopological approaches to species delimitation using molecular data. Although a central theoretical concern in molecular systematics is discordance between a given gene tree and the true "species tree," the primary empirical problem faced in reconstructing taxic phylogeny is incongruence among the trees inferred from different sequences. Linkage relationships limit character independence and thus have important implications for handling multiple data sets in phylogenetic analysis, particularly at the species level, where incongruence among different historically associated loci is expected. Gene trees can also be reconstructed for loci that influence phenotypic characters, but there is at best a tenuous relationship between phenotypic homoplasy and homoplasy in such gene trees. Nevertheless, expression patterns and orthology relationships of genes involved in the expression of phenotypes can in theory provide criteria for homology assessment of morphological characters. [Gene trees; homoplasy; linkage; morphology; orthology; phylogeny; species delimitation.]

Page (1993, 1994) summarized several ways in which trees from one hierarchical level may "include" historically associated trees from a lower level. The trees of the lowest hierarchical and organizational level, the gene, are discussed here. Gene trees are fundamental to molecular systematics, i.e., to systematics generally. The primary use of such trees, constructed from DNA sequence variation at individual genetic loci, is to infer phylogenies of the taxa bearing those genes. As such, genes are little more than collections of individual nucleotides, each assumed to be an independent and representative character for organismal phylogeny reconstruction. The fact that systematists are speaking of "gene trees" at all and distinguishing them from trees of taxic relationships represents a significant advance over earlier— not long past—ways of thinking about molecular systematics. The full implications of that distinction, necessitated at each taxonomic level by different but often overlapping biological processes, are still being elucidated and remain to be incorporated into the day-to-day work of systematics.

Two aspects of "inclusion" that involve gene trees are explored here. The first is gene flow, which in sexually reproducing organisms brings together lineages with different histories. Gene trees thus become historically associated as genes are united in different combinations within the genomes of individual organisms. Two hierarchical levels are involved: the lower level is that of the gene, and the higher level is that of the organism. The presence of different alleles at a locus provides the substrate for effective recombination. Within a contiguous set of nucleotides composing a structural gene, recombination may produce new alleles that are phylogenetic mosaics, with significant implications for gene tree construction. Above the organizational level of single loci, recombinational mechanisms (crossing over, gene conversion, independent assortment) permit individual loci to evolve independently and potentially to track different histories. An im-
important consequence is the classic "gene trees versus species trees" problem, whose empirical effect is incongruence between different loci. Although these issues have received considerable attention (e.g., Maddison, 1995), their implications for species delimitation have been less well explored (but see Baum and Shaw, 1995; Doyle, 1995) and are discussed here. The existence of separate trees for different parts of a gene or for different genes within species raises the issue of how to accommodate potentially conflicting data sets, an area of considerable controversy (e.g., de Queiroz et al., 1995; Miyamoto and Fitch, 1995).

A second type of "inclusion" also takes place within the genomes of individual organisms but involves different organizational levels rather than different hierarchical levels. Genes underlie the morphological attributes that are the basis for most taxonomic treatments and, until the advent of macromolecular data, were the characters of most phylogenetic analyses. Gene trees are "included" in morphologies, and the relationship between their topologies and morphological character state trees is explored here.

Gene Trees and Species Trees

Asexually reproducing species represent a simplified case for species delimitation and phylogeny reconstruction (e.g., Lidén, 1990) because there is only phylogeny—not tokogeny—Hennig's (1966) term for reticulate relationships among sexually reproducing individuals. In such taxa, the entire genome is a single historical unit and is thus a single "gene" in terms of coalescence theory (e.g., Hudson, 1990). This is not the case with sexually reproducing species. For such taxa, consequently, every stage of molecular systematics is complicated, from the construction of a tree for alleles at a given locus, to delimiting species, to inferring relationships among organismal taxa.

Gene Trees within Genes: Recombination, Incongruence, and Homoplasy

A gene as defined in classical genetics, molecular biology, and empirical molecular systematics (e.g., Rieger, 1991) is a contiguous set of nucleotides, usually with a particular function: coxl, adh, rbcL, etc. However, as has been recognized for some time (e.g., Grant, 1957) and discussed explicitly more recently (e.g., Hudson, 1990), the historically relevant unit (the coalescent gene [c-gene]) may be either larger or smaller than this standard definition indicates. With no effective recombination, the c-gene can be an entire genome, as in the case of organellar genomes; with recombination, it can be as small as a single nucleotide, and multiple c-genes can therefore exist within a single conventional gene.

Hudson (1990) pointed out that the genealogy of a single contiguous set of nucleotides may be impossible to reconstruct with any confidence if recombination levels are sufficiently high. In discussing cladistic analyses, Nixon and Wheeler (1992) noted that homoplasy caused by recombination does not result from incorrect homology assessments in the usual sense but rather from conflict among the different trees supported by characters belonging to regions differing in their historical patterns (see example in Fig. 1). Thus, recombination is one case (others are hybridization and gene conversion among paralogous loci) of mixed-signal homoplasy (e.g., Doyle, 1996).

Thus the "gene" whose assumed unique history molecular systematists attempt to reconstruct may itself be historically heterogeneous, having several different "included" c-genes. Because history is the primary concern of those reconstructing phylogenies, it would seem that this should be the first and most significant process partition (Bull et al., 1993) treated by those concerned with data partitions and should be more important, for example, than the kinds of functional criteria outlined by Miyamoto and Fitch (1995). However, there seems to have been little thought given to this issue in the debate over methods of combining data. It falls under the broader heading of historical incongruence, a topic that, as noted recently by de Queiroz et al. (1995), has received
less attention in this debate than it deserves. The practical implications are obvious. While advocating strict separation of data sets, Miyamoto and Fitch (1995:66) commented that “a lower limit on the sizes of process partitions is effectively set by the need for sufficient numbers of characters to test statistically for significant heterogeneities among these partitions.” A sequence composed of c-genes that can be as small as a single nucleotide clearly poses significant challenges to those advocating data partitioning. Although few sequences within species may show this extreme a level of recombination, cases of recombination extensive enough to prohibit construction of well-resolved allele trees are known (e.g., \textit{adh} in \textit{Drosophila pseudoobscura}: Schaeffer and Miller, 1992; see Doyle, 1995).

\textit{Genes and Their Trees in Species Delimitation}

Reconstruction of a gene tree terminates the initial stage of the molecular systematics research program, the part that involves objective phylogenetic analysis (Doyle, 1995). The second stage of the program involves what I have elsewhere referred to as a leap of faith (Doyle, 1992), in which the systematist substitutes the names of organismal taxa (species, genera, families) for the alleles or haplotypes, sampled from individuals, that are the true terminals of the objective analysis. For the following discussion, I assume that the many technical and theoretical difficulties inherent in this first stage of the analysis have been surmounted and that accurate gene trees have been reconstructed.

The next question that must be addressed is whether there is a phylogenetic history of organisms and organismal taxa that can be recovered. This question seems uncontroversial with regard to taxa at, for example, the generic level and higher. However, it is demonstrably untrue for individuals within sexually reproducing organisms connected by gene flow, the relationships among which are tokogenetic (reticulate) rather than phylogenetic (divergent). Between these relatively clear extremes lies the species. Despite the controversy surrounding this hierarchical level, most concepts agree that actual, potential, past, or present gene flow is of critical importance in defining species. Thus, in at least a broad sense (and strictly speaking, in the case of some species definitions), relationships within species are primarily tokogenetic, and relationships between species are assumed to be phylogenetic and thus represent a historical sequence of cladogenic events. Whether any two individuals of species that diverged from one another most recently will be more closely related to one another than to an individual of a third species that diverged earlier is difficult to predict. This difficulty remains whether the criterion for “relationship” is overall similarity of their genomes, the gene tree at any particular locus from which species relationships could be inferred, or the overall number of loci at which gene trees correctly predict the sequence of cladogenic events.

Individual sexually reproducing organisms are not themselves species; their relationships are not phylogenetic. Consider the technical problem of reconstructing a gene tree for alleles at a single locus (an allele tree) for which heterozygous individuals exist. It is (theoretically) straightforward to construct a tree in which the operational taxonomic units are alleles sampled from conspecific individuals that are linked by gene flow. The stage at which organismal taxa are substituted for genes is not controversial when organellar haplotypes are sampled from a homoplasmic individual or when alleles at a nuclear locus are sampled from a homozygote. But an individual that is heterozygous for a given locus possesses two alleles at that locus, each with its own unique historical position in the gene tree. It is therefore impossible to make a unique substitution of “individual” for “allele” in the gene tree; such an individual must appear twice and at different positions on the tree. The interpretation of this situation in terms of either individual or species relationships is not apparent. Heterozygosity is a case of association among historical lineages, in
this case involving alleles with different histories that are brought together in a single diploid genome. Heterozygosity is a product of gene flow and thus is evidence that relationships among individuals are tokogenetic rather than phylogenetic.

The use of sexually reproducing individuals as terminal taxa in a molecular phylogenetic analysis therefore does not provide even a technical solution to the species problem, as has been suggested (Vrana and Wheeler, 1992). Some unit that includes many individuals is required, meaning that some criterion for recognizing "species" is necessary. It is reasonable to ask whether gene sequences could be used to provide such a criterion, given the current reliance of systematists on DNA data.

The most obvious approach is to somehow utilize gene trees, the stock-in-trade of molecular systematics. Baum and Shaw (1995) proposed a method for delimiting "genealogical" species on the basis of coalescence of alleles or haplotypes at multiple loci. Their method operates on the assumption that relationships among individuals within such species are tokogenetic and that individuals within a species will therefore possess alleles having different relationships at each of many loci. In the allele tree for one locus, the alleles of individuals 1 and 2 may be sisters, whereas in the tree for an unlinked second locus, the allele possessed by individual 1 may be more closely related to that of individual 5. Substituting individuals for alleles in the tree for each locus permits the construction of a strict consensus tree of individuals over all loci. This consensus tree will lack resolution among individuals that are linked by gene flow, collapsing to the point where groups of individuals do not share related alleles. This point of multilocus coalescence defines the genealogical species.

An alternative approach makes use of DNA sequence data in a different way, by using the distribution of alleles in individuals regardless of their positions in an allele tree. Individuals sharing alleles are expected to be close relatives if alleles are defined as DNA sequences that are identical at both coding and noncoding nucleotides. But how can alleles, and by extension the individuals bearing them, be grouped without reference to an allele tree? The method of Doyle (1995) takes advantage of the fact that a heterozygous individual provides evidence of gene flow between individuals that are homozygous for each of its two alleles. Alleles that coexist in heterozygous form are part of a common "allele pool" at their locus (Doyle, 1995), the critical property of which is that members of the same allele pool are potentially able to recombine. The group of individuals that possess alleles belonging to a single allele pool thus are a field for recombination (Carson, 1957; FFR of Doyle, 1995). The allele pool may include alleles that never occur together in heterozygous form: for example, there may be no a/f heterozygotes, but if the a/b and b/f heterozygotes join alleles a, b, and f in the same allele pool, then all individuals bearing any of these alleles will belong in the same FFR. Moreover, these three alleles need not bear any particular relationship to one another in the allele tree at this locus. Allele f may be only distantly related to alleles a and b, and other individuals may possess alleles more closely related to alleles a or b. But if none of these exist in heterozyg-

---

**FIGURE 1.** The effect of recombination on phylogeny reconstruction. L = length; CI = consistency index; RI = retention index. (a) Data matrix of four ingroup (A-D) alleles and one outgroup allele prior to recombination, with 12 binary characters numbered sequentially. (b) Allele phylogeny reconstructed from the data matrix shows no homoplasy. (c) Recombination of alleles A and C to produce a recombinant, R; the corresponding R, recombinant is not followed. (d) Phylogeny reconstruction of two coalescent genes whose boundary is the site of the recombination event between alleles A and C; neither of the two c-gene trees shows homoplasy. (e) Parsimony analysis of the whole data set, ignoring the presence of two historical units (c-genes), results in two trees and homoplasy. Homoplasy is due to the mixed signals of the two c-genes.
gous combinations with alleles $a$, $b$, or $f$, then such individuals will not belong to the same FFR. Once allele pools are defined, gene pools can be defined in much the same way, by considering the intersections of multiple allele pools: a multilocus genotype with allele $b$ at locus 1 and allele $x$ at locus 2 links alleles $a$, $b$, and $f$ with all alleles belonging to the same allele pool at locus 2 as does allele $x$. From the gene pool, a multilocus FFR can then be defined (Doyle, 1995).

These topological and nontopological methods both rely on the fact that individuals within larger entities (genealogical species or multilocus FFR) have different sets of reticulate (tokogenetic) relationships. Each suffers from different practical problems. Topology-based approaches must contend with loci at which ancient polymorphisms persist. The coalescence points of alleles at such a locus predate many cladogenic events leading to taxa among which gene flow has since ceased and that may be well differentiated morphologically, ethologically, etc. Classic examples are the major histocompatibility complex (MHC) of animals (e.g., Gaur et al., 1992) and self-incompatibility loci of plants (e.g., loerger et al., 1990; Clark and Kao, 1991; Richman et al., 1996). Using the MHC as an example, the ancestor in which all human DQB1 alleles coalesce predates the divergence of humans, chimpanzees, gorillas, macaques, and a number of other primates (Gaur et al., 1992). For a topological species definition, either this large and diverse group of taxa must be accepted as a single "species" or the locus must be excluded from consideration in species delimitation. If a goal of species delimitation is objectivity, then it would seem reasonable to justify either decision based on some set of criteria. Recognizing apes and humans as a single species may be unpalatable from a practical point of view, but the decision to do so does use an objective and repeatable approach to define a natural group with particular biological properties. It is more difficult to envision practical objective criteria for excluding a locus from consideration. One might decide to exclude MHC loci on the basis that these loci represent a classic case of selectively maintained polymorphism (e.g., Hughes and Nei, 1988), for which it is expected that polymorphisms will transcend accepted specific (or even generic) boundaries. However, it is difficult to see how this solution can be applied objectively for most loci, where often little or nothing is known about mechanisms maintaining polymorphism. Moreover, some of the best methods for detecting selection at a locus (e.g., Hudson et al., 1987) rely on comparisons of polymorphism within species versus divergence between species; clearly a species concept must be invoked and applied prior to using such tests. It does not appear that a practical criterion has yet been proposed for excluding loci that does not have as a prerequisite agreement with existing or "reasonable" taxonomic boundaries.

In contrast, trans-specific polymorphism is theoretically not a problem for a nontopological method because topological relationships are expected to persist far longer than will allele identities. Consider two extant DQB1 alleles, one sampled from a human and the other from a chimpanzee, that were identical immediately prior to the cladogenic event that gave rise to the two taxa. These sister alleles have persisted to the present in both taxa, having been prevented from either becoming extinct or going to fixation by the selectional regime at this locus. These alleles share a more recent common ancestor with each other than either does with any other DQB1 allele (except for those derived from each subsequent to the divergence of humans and chimpanzees), and so they will be sisters on the DQB1 gene tree, more closely related than either is to other alleles found in humans or chimpanzees. Yet these alleles have been diverged for some 6 million years, and so it is likely that they will differ from one another by at least one substitution, perhaps in a noncoding region or at a synonymous codon position. Despite being sisters on the allele tree, they are not the same allele (e.g., Rieger, 1991) and so will not provide a link between humans and chimpanzees. There will not be any
heterozygous individuals having these two alleles, because there has been no gene flow between chimpanzees and humans since their divergence from their common ancestor. Theoretically, therefore, even a locus like DQB1 can be accommodated in the nontopological method of Doyle (1995), although low levels of allele divergence coupled with parallelism could produce “shared” alleles that were not identical by descent and so would provide misleading estimates of relationship. The fine resolution provided by DNA sequences on which the nontopological approach relies may be a source of “problems” in the form of unpalatable results, although in a direction opposite that of the topological method. The finer the resolution, the less likely that identical alleles will be shared among individuals. For many taxa, the units suggested by such a procedure may be very small, and it may then be desirable to identify some objective means of grouping FFRs together (Doyle, 1995).

Gene Trees and Species Phylogeny

Once units of phylogeny have been delimited, the topology relating them can be inferred by the application of the nontopological molecular systematics program (Doyle, 1995). However, some major problems remain. First, not all loci are suitable for this purpose. For example, even if the DQB1 locus can be used to delimit human, chimpanzee, and gorilla species, the relationships of DQB1 alleles are tangled enough that no historical inference could be made from them for these species. For a locus to be suitable for phylogenetic inference, a minimum requirement is that its alleles be monophyletic within terminal taxa (Doyle, 1995). This phylogenetic “suitability” requirement can be applied objectively and uniformly to any locus once species have been delimited. Although it involves exclusion of some loci, it is not subject to the criticism of topological methods for species delimitation that they may require subjective decisions concerning which loci must be excluded to give a desired result.

More important than this technical concern is the fact that even if all alleles are monophyletic at a locus within the species being studied, there is no guarantee that the overall topology at that locus will reflect the cladogenic history of the species themselves. This problem is simply the now-familiar one of gene tree versus species tree. However, although the problem is usually phrased as one of incongruence between a particular reconstructed gene tree and “the” species tree, the true taxon tree is unknown and unknowable, with few exceptions (e.g., experimental phylogenies; Hillis et al., 1992). Therefore, to discuss incongruence in terms of the unknown species tree is irrelevant from an empirical perspective; although there is presumably a unique species history to be recovered by constructing gene trees for different loci, only the individual gene trees actually can be observed. If some significant percentage of gene trees within a set of species is expected to disagree with the species history (e.g., Pamilo and Nei, 1988), then it is incongruence among gene trees that will be observed when multiple loci are studied. This conflict must somehow be reconciled in arriving at a species-level phylogenetic hypothesis. The fact that incongruence among gene trees is expected is an important aspect of the gene trees versus species trees issue that should perhaps be more strongly stated than has often been the case.

In their review of issues concerning combining data, de Queiroz et al. (1995) distinguished three approaches to the problem of incongruent data sets. The first approach is to produce a collection of individual histories. However, species are more than merely collections of independent gene lineages, and usually they do have a singular history. Systematists are interested in historical associations of organismal lineages, not just in gene trees for each of the many thousands of sequences that comprise the genome. De Queiroz et al.’s (1995) second option is to depict relationships as a reticulating tree. However, lineage sorting, which is the primary stochastic source of incongruence, is not retic-
ulation per se, unlike secondary processes such as hybridization. The final alternative is the production of a hierarchical tree "representing the single dominant pattern among data sets" (de Queiroz et al., 1995: 674), i.e., combining data sets or topologies; whether some taxa can or should be excised (de Queiroz et al., 1995) is a separate issue. The entire spectrum of alternative ways of dealing with multiple data sets thus can be discussed under this third option. A consensus tree of gene trees from several loci in a taxonomic congruence study is clearly a means to this end, and if data (instead of topologies) are to be combined, the choice between direct combination and recoding methods (e.g., Doyle, 1992) boils down to a question of whether one should trust the species tree supported by the most nucleotides or the species tree that is supported by the most loci. Both are total evidence trees, and both are character congruence methods, but with different definitions of the relevant character.

Character definition is an important problem in any phylogenetic analysis, and a fundamental attribute of a useful character is independence (e.g., Davis and Heywood, 1973). Independence normally means that the state of one character does not predict the state of another. Systematists seek to explain observed character covariation in terms of phylogeny; other forces or phenomena that cause the states of several characters to vary together pose problems for phylogeny reconstruction (although they are of obvious interest evolutionarily; e.g., Sanderson and Hufford, 1996). Natural selection is the primary force leading to nonindependence, e.g., selection for particular amino acids in a protein such as lysozyme (Messier and Stewart, 1997) or selection to maintain secondary structure in ribosomal RNAs (Wheeler and Honeycutt, 1988). Constraints involving function, secondary structure, nearest-neighbor effects, base compositional biases, etc., are all concerns for the construction of individual gene trees.

The historical independence of c-genes is, paradoxically, a primary source of character nonindependence when multiple data sets are used to reconstruct species histories because the unit of incongruence among loci is not the single nucleotide but rather the entire locus or even larger regions of DNA that assort independently (e.g., nuclear or organellar chromosomes, c-genes). Organellar genomes represent a familiar and relatively simple example because all of their genes are irrevocably linked and travel as single historical units. When the organellar genome has a history different from that of the nuclear genome, every nucleotide in every one of its genes will give the "wrong" phylogenetic pattern for those taxa affected (e.g., in lineage sorting or introgression). Under such circumstances, combining organellar and nuclear sequences to produce a single tree in a character congruence approach will produce mixed-signal homoplasy, as does recombination at a single locus. In the case of multiple loci, larger regions of the genome—loci, blocks of loci, or whole chromosomes—rather than different regions along the same gene have different "included" histories. Because of their historical linkage to one another as a single c-gene, the characters involved are not independent; homoplasy at one predicts homoplasy at all others (Doyle, 1992). Numerous cases of incongruence between organellar and nuclear genomes are now known (reviewed by Wendel and Doyle, in press). However, incongruence among nuclear loci that are in linkage equilibrium with one another should be just as much a problem unless there is some general propensity for organellar and nuclear sequences to behave differently evolutionarily (see Moore, 1995; Hoelzer, 1997).

Some significant fraction of all neutrally evolving loci are expected to show patterns other than the true phylogeny and will therefore disagree with one another, which makes this topic important in the debate over data combination. This issue has received much less attention than it deserves (as noted by de Queiroz et al., 1995) perhaps because fundamental historical incongruence of different loci is a problem
for any method that deals with data from multiple genes and still has as its goal the construction of a single species tree. Each approach makes different compromises in dealing with this problem.

Consider three loci (A, X, Y) sampled from a group of five species related by a nonreticulating tree (Fig. 2). Loci X and Y are genes of 500 nucleotides each, and locus A is a gene of 1,500 nucleotides. The common ancestor of all five species was fixed for allele x at locus X and for allele y at locus Y but was polymorphic for alleles a and b at locus A. This polymorphism was carried through other ancestors, until fixation occurred for either an allele of type “a” (open bars in Fig. 2) or type “b” (solid bars). Suppose that all three loci yield gene trees with no homoplasy and robust support for all nodes, that the number of synapomorphies provided by each gene is in direct proportion to its length, and that an identical proportion of the total number of synapomorphies supports each clade in each gene tree. This classic example of lineage sorting leads to the inference of a species tree from locus A that does not agree with the true species tree; it is also incongruent with trees reconstructed for loci X and Y, which track the true species history.

In this situation, a taxonomic congruence approach using currently available methods (but see Miyamoto and Fitch, 1995) would combine the topologies of all three trees to produce a strict consensus tree, which in this case fails to resolve a dominant pattern among the data sets (Fig. 2b). In a character congruence approach (e.g., Kluge, 1989), all three loci are combined to form a single data set. In this case, parsimony analysis recovers the topology of locus A because this locus provides more nucleotide characters than the combined X and Y data sets (Fig. 2c). A dominant pattern is resolved, but it is not the pattern of the true species tree. A third alternative is to treat each locus as a single ordered multistate character whose states are the alleles and whose transformational order is given by the gene tree (Doyle, 1992). This approach is analogous to Brooks parsimony analysis (e.g., Brooks, 1990) as used in biogeography and host/parasite coevolution studies. In an analysis of these three “characters,” two trees are identified, one with a topology identical to that of the tree from loci X and Y and the second with a topology found in no other analysis (Fig. 2d). The three approaches differ markedly in their sensitivities to different factors. Taxonomic congruence emphasizes topology, and strict consensus is sensitive to any single locus that shows topological differences, regardless of how many loci are studied. Perhaps other consensus methods (e.g., majority rule consensus of the strict consensus trees from each
locus?) would be preferable as an approach to taxonomic congruence, once enough loci were sampled. Character congruence is sensitive to the number of characters supporting contrasting patterns, regardless of the number of historically distinct linkage groups (c-genes) involved.

The recoding method is intermediate, emphasizing individual c-genes as the units of analysis but at the cost of ignoring the relative amounts of the genome represented by each and the relative support for individual clades in each tree.

To the extent that linkage (historical association) is a key consideration, then none of these options is adequate by itself. Consider the following simple example. A group of sexually reproducing eukaryotic species each has a strongly bimodal karyotype composed of four chromosomes in the haploid set. The single large chromosome is considerably longer than the sum of the smaller chromosomes, and each chromosome bears a number of genetic loci proportional to its length. There is independent assortment but no recombination within chromosomes; therefore, the units of lineage sorting are whole chromosomes, each of which behaves as a single c-gene. Lineage sorting produces a situation in which all but one linkage group tracks the correct species tree, but that linkage group, by chance, is the largest chromosome. The entire genomes of the species are sequenced, and trees are produced for each of their thousands of loci. Because all loci on each chromosome have the same history, each linked locus should potentially capture that history. Given the pattern of lineage sorting described and ignoring conventional sources of homoplasy, a direct combination of all loci across all chromosomes is likely to result in a tree with the topology of the tree based on the large chromosome, the single linkage group that does not agree with the remaining chromosomes or with the species tree. This example is a restatement of the example in Figure 2 using units larger than the genetic locus. Treating the nucleotide as the unit character would cause independence problems but so would dealing separately with individual loci, regardless of whether a consensus or recoding approach were taken.

Unfortunately, detailed knowledge of nuclear gene linkage relationships is generally not available to systematists. Even were such information available, however, recombination within chromosomes would severely limit the utility of any method that assumed strict correlation between physical and historical linkage groups. Only in organellar genomes does it seem possible to make assumptions about linkage of different genes or about the likelihood of their historical independence from nuclear genes (e.g., Moore, 1995). No solutions are proposed here, but it is clear that any proposals about data combination should consider the fundamental issue of historical lineage association of linked genes.

The Sufficiency of Gene Trees

The case has been made (e.g., Doyle, 1995) that gene trees are insufficient for some purposes, such as delimiting species, because they do not adequately track the relationships of the organismal units within which they are “included.” There are other circumstances, however, when species relationships may be in some sense irrelevant and gene trees are quite adequate; for example, in studies of organellar genome evolution, knowledge of organismal phylogenies may not be necessary. Miyamoto et al. (1994) made effective use of this situation in testing the utility of different portions of the mitochondrial genome for phylogeny reconstruction. Similarly, the question of whether losses of sequences from plant chloroplast genomes represent independent molecular events (e.g., Doyle et al., 1995) can be addressed adequately with chloroplast gene trees whether or not those same trees are faithful as species trees.

Biogeographic studies use taxon trees to reconstruct area relationships (e.g., Page, 1993). For such studies, the problem of discordance between gene trees and taxon trees may be irrelevant. Avise et al. (1987) coined the term *intraspecific phylogeography*
for the study of area relationships from gene trees. If discoverable geographic relationships among alleles or haplotypes exist, a gene tree will be sufficient for inference of area relationships. It should make no difference from which taxa the alleles or haplotypes are sampled, and disagreements between gene trees and species trees should not be important. Of course, different samples (loci) from the same taxa are not likely to be historically independent, being "included" in the same set of genomes. Thus, one is still confronted with the need to construct area hypotheses based on multiple unrelated taxa and not simply based on multiple loci from the same set of organisms.

Gene Trees in Morphology

Species histories prune and shape gene trees, and therefore gene trees can be used to make inferences about species histories. This use of "included" gene histories spans hierarchical levels. In contrast, genes and morphological characters belong to different organizational rather than hierarchical levels: genotype and phenotype. These organizational levels are connected by the products of gene expression, proteins and structural RNAs, from which organisms are constructed. To this extent, genes are "included" as a lower level within morphological characters, and gene tree topologies are connected to morphological character distributions by spatial and temporal patterns of gene expression and regulation.

Although systematists generally agree that molecules and morphology are complementary sources of phylogenetic information, molecular data now provide the preponderance of characters for phylogeny reconstruction. Nonmolecular characters should not be forgotten, as suggested by Lanyon's (1988:572) observation that stochastically evolving molecular characters may be inferior to selectively maintained morphological features in studies of ancient radiations: "morphological characters may well retain synapomorphies under circumstances under which molecular characters would not." The complexity of many morphological features, compared with the small number of states of nucleotides or amino acids, is an additional advantage of morphology in homology assessment, but that very advantage is compromised by the ambiguities associated with scoring morphological characters (e.g., Stevens, 1991). Morphological complexity has roots in the network of gene interactions that underlie all but the most simple phenotypic characters. Is there a way to combine the advantageous aspects of molecular and morphological data? Hillis and Moritz (1990:515) suggested that "molecular biology may begin to provide more information on the molecular basis of development, so that a true synthesis of molecular and morphological data can occur." If that synthesis is to occur, it will require a shift of emphasis in "molecular" systematics, which currently treats all genes as essentially equivalent grab-bags of nucleotides and ignores their functional roles.

There has been a tremendous explosion of knowledge about the molecular underpinnings of morphological features and ontogeny in both plants and animals, and the evolution of developmental systems is under active investigation (e.g., Akam et al., 1994). Forging links between systematics and molecular developmental genetics requires, among other things, identifying common areas in which the techniques of one field can be applied to the other. I have argued (Doyle, 1994a) that providing objective (i.e., phylogenetic) determinations of paralogy and orthology is a potential major contribution of systematics to molecular developmental genetics. Many workers studying homeoboxes make explicit use of these phylogenetic concepts (e.g., Schubert et al., 1993), and laboratories studying floral development have recently taken note of the information available in phylogenetic studies (Purugganan et al., 1995; Tandre et al., 1995). In return, molecular developmental genetics offers systematists detailed information about the connection between genes and morphologies. Such information can be useful in a variety of ways for phylogenetic issues. For
example, Whiting and Wheeler (1994) drew on the detailed understanding of dipteran development to formulate a hypothesis of the origin of halteres in Strepsiptera, combining this information with their own studies of insect phylogeny using 18S ribosomal genes. For phylogeny reconstruction from morphological characters, molecular developmental genetic data are of clear relevance for the issue of character independence, most simply in cases of pleiotropy. Mutations in several of the genes involved in floral development are expressed at more than one stage of development, and some have dramatic effects on more than one organ. Mutations in the SUPERMAN gene of *Arabidopsis*, for example, affect the number of stamens as well as the texture of the seed coat (Gaiser et al., 1995).

Of interest here is the general connection between molecules and morphology in phylogeny reconstruction, particularly the implications of the association between morphological character patterns and the trees of the genes “included” in them. Expectations concerning levels of homoplasy are central to debates about the relative utilities of different classes of characters, whether the classes are both morphological or both molecular or whether morphology and molecular data are being compared. Specifically, it is worth exploring how homoplasy at one level is connected to homoplasy at the other.

The connection is not simple nor is it expected to be, given the complexity of the overall connection between genotype and phenotype (Roth, 1988, 1991). After all, genes do not so much control morphology as participate in it (Nijhout, 1990), and the participation is a complex one, better described as a network than as a linear pathway. But the fact that the connection between genotype and phenotype is not simple does not mean that the disconnection in homology between the two levels cannot be illustrated by simple examples (e.g., Doyle, 1996). Phenotypes are produced as an expression of genotype, and morphology is only one of several manifestations of gene expression. Mobility of a native protein in an electric field is also a phenotypic character, albeit a much more simple one than morphology. Mobility differences among proteins encoded by orthologous loci (allozymes) are due to differences in net charge under particular electrophoretic conditions. Charge in turn is determined by the primary amino acid sequences and thus ultimately by the nucleotide sequences of the genes encoding the proteins. The character “electrophoretic mobility” could show homoplasy as a result of parallel changes in net charge. This homoplasy could be due to a parallel nonsynonymous nucleotide substitution, in which case there would be a one-to-one correspondence in homoplasy at the nucleotide, amino acid, and phenotypic (mobility) levels. But the same false homology for mobility could be caused by different nonsynonymous substitutions at the same codon, resulting in identical but nonhomologous amino acids, or even by substitutions involving completely different amino acids. In these cases (illustrated by Doyle, 1996), homoplasy occurs in the phenotype (mobility) but not at the level of the gene. Conversely, because relatively few nonsynonymous nucleotide substitutions will result in net charge changes, many homoplastic replacement substitutions that would be detected at either the nucleotide or amino acid levels will not appear as homoplasies for the phenotypic character “mobility.” Of course, the most dramatic—and familiar—disconnection between organizational levels occurs between the nucleotide and amino acid levels because of the expectation that synonymous changes will generally far outnumber nonsynonymous substitutions.

These considerations presumably apply for any gene whose tree one might wish to compare with the phylogenetic distribution of a phenotypic character it underlies. With morphological characters, the connection is still more tenuous, however. Gene trees are usually constructed from coding regions, yet sequences critical for the spatial and temporal expression of the gene are generally located outside this region, and even subtle changes in expres-
sion patterns can produce profound effects on the phenotype. All of these factors represent disconnections between single genes and morphological characters. Beyond this level lie the further sources of disconnection caused by interactions among several genes (for examples, see Doyle, 1996). Weston (1987) suggested that biosynthetic pathways can serve as useful models for the more complex network of ontogenetic interactions (cf. Alberch, 1985). An example of the disconnection between gene tree homoplasy and homoplasy involving micromolecular characters is disruption of a pathway at two different steps in two unrelated taxa, leading to a parallel loss of a particular compound. There need be no homoplasy at all in the trees for any of the pathway's genes, yet homoplasy will be present for the chemical character.

A consequence of the disconnection between homoplasy at the two organizational levels is that it defies facile characterization in the “molecules versus morphology” debate. One simplistic view that may subconsciously pervade perceptions of the connection between genotype and phenotype in phylogeny reconstruction is suggested by the following statement (Doyle, 1990:6):

> morphological characters . . . are likely to be largely determined by genes undergoing the same complex evolutionary processes as those from which we attempt to construct our molecular phylogenies, and in addition are likely to show further layers of complication from . . . a host of other processes.

This statement implies a belief that morphological characters must somehow sum the homoplasy at the two organizational levels that influence them. If four genes affect a particular morphological feature, each gene is expected to show some homoplasy, which may lead to uncertainty in its gene tree and, consequently, in the inference of taxic phylogeny made from it. In this naive view, the morphological character would be much more (perhaps four times more) unreliable than any of the genes that influence its expression. The above examples show that this will rarely, if ever, be the case.

Knowledge of the genotypic basis of phenotypic characters should be helpful in providing new criteria for homology assessment. Genic similarity extends the usual concept of structural similarity: are the same genes involved in a character hypothesized to be homologous? Moreover, are these genes expressed in similar ways, spatially and temporally, in different taxa? It might not be apparent that application of these criteria involves “included” gene trees. Certainly, it may be trivial to determine that two genes are not the same (non-homologous) if, for example, they show virtually no nucleotide similarity above random levels. But homology is certainly (if not exclusively) a phylogenetic concept (e.g., de Pinna, 1991), and relationships among genes are often complex, in many cases caused by successive waves of duplication and divergence that are best depicted and understood by hierarchical relationships, i.e., gene trees. Ultimately, what is meant by “the same” gene is “identical alleles at orthologous loci,” although in studies at deeper levels it may be sufficient to focus on orthology relationships alone.

To illustrate the application of gene trees to assessment of homology in a phenotypic character, consider a unique morphological structure, found in relatively few taxa of some larger group. A phylogenetic analysis of these taxa suggests that the structure is not homologous in all of the taxa but rather has arisen more than once in the group. Obvious tests of this hypothesis would include anatomical and ultrastructural studies. One could also ask how the features that make this structure unique at the morphological level develop at the molecular level. Are these unique features the result of the action of orthologous genes expressed in similar ways in the structures of all taxa? The nodule of leguminous plants provides an example of such a character. Symbiotic fixation of atmospheric nitrogen by leguminous plants and soil bacteria (nodulation) is most parsimoniously hypothesized to have arisen independently in at least two, and perhaps three, groups of Leguminosae (Doyle, 1994b; Doyle et al., 1997). The nodules that house the bacteria are considered novel plant organs, and their homologies with other organs are a
matter of debate (Hirsch and LaRue, in press). Within nodules, bacteria in their nitrogen-fixing stage (bacteroids) are surrounded by a complex membrane of host origin, forming what has been called a novel organelle, the symbiosome (reviewed by Mylona et al., 1995). Nodules show considerable morphological, developmental, and biochemical diversity in the Leguminosae but, based on positional and structural criteria, appear to be homologous throughout the family. The apparent homology of nodules in diverse legumes combined with the complexity of the nodulation syndrome make it perhaps more plausible that nodulation arose only once in the family but was lost independently in many lineages.

Additional criteria for assessing nodule homologies would therefore be useful as tests of the competing hypotheses of single versus multiple origins, and such criteria could potentially be provided by genes encoding nodulins, proteins that perform a variety of enzymatic and structural roles in the nodule (Mylona et al., 1995). Nodulins could be products of truly novel genes, produced by exon shuffling or by recent gene duplication from genes with other roles in the plant; alternatively, they could be recruited for function in the nodule while retaining their normal functions elsewhere in the plant. In principle, the homology tests proposed are simple: construct gene trees for gene families that include nodule-expressed genes and determine the orthology relationships of nodulin genes across taxa (Fig. 3; Doyle, 1994b). Finding similar roles performed by paralogous genes would constitute a falsification of the hypothesis of nodule homology because it would suggest independent gene origin or recruitment. However, finding orthologous genes performing similar roles in all taxa would be consistent either with nodule homology (a single gene origin or recruitment) or nonhomology (e.g., independent recruitment of orthologous genes).

In practice, the situation is much more complex. For one thing, some gene families evolve concertedy, obscuring orthology relationships and making the test impossible to perform; one of the most abundant nodulins, leghemoglobin, is encoded by such a family (e.g., Doyle, 1994b). Even in gene families where orthology relationships can be hypothesized, patterns of gene expression may be very complex. In the cytosolic glutamine synthetase family, for example, some species include constitutively expressed genes with low level nodule expression, as well as paralogous genes that are dramatically up-regulated in the nodule but also are expressed in some other tissues (e.g., Temple et al., 1995; Doyle and Doyle, 1997). Shifts in nodulin gene expression seem likely to have occurred, given the structural and biochem-

![Figure 3](https://image-store.com/figure3.png)

**FIGURE 3.** Using gene trees and expression patterns of genes "included" in a morphological character (nodules) to assess homologies of that character. (a) Species tree for seven taxa; taxa 1, 6, and 7 produce nodules, and the most-parsimonious optimization of the nodule character is a parallel gain (solid bars). (b, c) Gene trees for two loci, one of whose paralogous genes encodes a product expressed in the nodule. In (b), paralogous genes appear to have been recruited for a role in nodulation in taxon 1 and in the ancestor of taxa 6 and 7. This situation is consistent with a separate recruitment of genes in the course of a parallel origin of nodulation in these taxa. In (c), orthologous genes are expressed in the nodules of these taxa. This situation is consistent with either a single origin of nodulation (with loss of the trait in taxa 2–5) or parallel recruitments during separate origins.
ical diversity that exists even among nodules of taxa sharing recent common ancestors. Such shifts could result in misleading conclusions: paralogous genes could be expressed identically in homologous nodules, or orthologous genes could encode nodulins in nonhomologous nodules.

Because the tests are not foolproof, results from any given nodulin gene family may not be reliable, and therefore information from several different gene families will be required to produce a robust hypothesis of nodule homology. This is, of course, precisely the same situation that exists in other studies that rely on "included" trees. The area cladogram for any particular taxonomic group may be misleading for a variety of reasons; only after several such cladograms are available are conclusions considered well supported. The gene tree for any particular gene may not track organismal histories accurately; several independent gene trees are required for a robust hypothesis of species relationships.

**COMBINING DATA FROM "INCLUDED" TREES: SOME FINAL COMMENTS**

The typical solution to all of these problems is to obtain data from additional genes or taxa whose "included" trees are being reconstructed. How many genes or taxa and how to utilize that information are open and controversial questions. The gene/species tree issue can be treated as a direct parallel to biogeography: the goal in phylogenetic reconstruction is to infer species relationships from "included" gene trees, whereas in biogeographic studies the goal is to infer area relationships from trees of "included" taxa. Both types of studies are conducted under the assumption that a direct inference of relationships at one level can be made from those at the other, an assumption that in both cases is expected to be invalid under a number of circumstances. For molecular systematics, problems of historical independence (bearing on the very definition of the gene) be-devil the issue of how to combine data at the level of closely related species, where the greatest potential for incongruence appears to lie. The definition of species is critical in such studies because delimitation of species to a considerable degree sets the threshold for expected incongruence among gene trees. Nontopological alternatives exist at various stages in the analysis of species using molecular data; these alternatives do not seem to be available for biogeography.

The issue of combining data takes quite a different form when the final class of "included" trees discussed here is considered. I have argued that topologies of "included" gene trees may have little relationship to gene expression and hence to phenotypic characters. In the discussion of testing homologies of morphological characters with gene trees, the question to be answered requires a simple "yes" or "no": structures are either homologous or they are not. Correspondingly, the information provided by each "included" gene tree is also binary: are orthologous genes performing a similar role in the taxa under consideration? Combining data to address the overall homology question takes the form of a simple tally of "yes" and "no" answers, one for each locus considered. Of course, homologous morphological characters exist as transformation series. Once characters are shown to be homologous, the topologies of "included" trees of an orthologous gene could perhaps be useful in ordering such a series. The disconnection between gene trees and phenotype suggests that this ordering could often be difficult, but in at least some cases it may be possible (e.g., Rogers et al., 1997); how trees from the various genes involved in a morphological character could be combined for this purpose is not immediately obvious.

If there is a single conclusion from all of this, it is that the business of inferring relationships at one hierarchical or organizational level using trees from another level is a tricky one because of the disconnections that exist between the levels. Nevertheless, as the history of systematics over the past decade has shown, gene trees represent a vast source of information.
that can be exploited to address a host of phylogenetic questions. The disconnections between genes and species are reason for caution, not despair. Similarly, despite the disconnections between genes and morphological characters, there is hope that the marriage of molecular developmental genetics and systematics can be a productive one.

ACKNOWLEDGMENTS

I thank Mark Siddall and Richard O'Grady for the invitation to participate in the 1995 SSB symposium that led to this article. I am grateful for many productive discussions with my colleagues in the Bailey Hortorium, especially Jerry Davis. I thank those who read drafts of the manuscript when it was first written, long ago, and also Dick Olmstead, Dave Cannatella, and two anonymous reviewers for comments during its final stages. My research has been supported by NSF; most recently by grant DEB 94-20215.

REFERENCES


