Page, R. D. M. (1994). Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol.* **43:** 58-77.

MAPS BETWEEN TREES AND CLADISTIC ANALYSIS OF HISTORICAL ASSOCIATIONS AMONG GENES, ORGANISMS, AND AREAS

Roderic D. M. Page

Biogeography and Conservation Laboratory

The Natural History Museum, Cromwell Road

London SW7 5BD, United Kingdom

Address for correspondence:
Roderic D. M. Page
Department of Botany
The Natural History Museum
Cromwell Road
London SW7 5BD
United Kingdom
Tel: +44 71 938 9168
Fax: +44 71 938 9260
e-mail: R.Page@nhm.ic.ac.uk

23/03/98

Abstract.—The concept of a reconciled tree arose independently in molecular systematics, parasitology, and biogeography as a means of describing "historical associations." Examples of historical associations include genes and organisms, host and parasitic organisms, and organisms and areas. A reconciled tree combines the tree for a host and its associate into a single summary of the historical association between the two entities under the assumption that no horizontal transmission of associates has occurred. This paper defines reconciled trees, describes an algorithm for their computation, and develops measures to quantify the degree of fit between host and associate trees. Examples are given of applying the method to gene trees and species trees, host-parasite cospeciation, and biogeography. The problem of incorporating horizontal transmission of associates (e.g., dispersal or host switching) is also addressed by introducing the concept of maximizing the amount of codivergence (shared history) between the associates. [Biogeography; cladistics; congruence; cospeciation; gene trees; phylogeny; reconciled trees; species trees; tree mapping.]

> What about the nonideal data, which like Lincoln's common men God must have loved, for He made so many of them. This is the real problem in biogeography, where our efforts must come together in the long run. [Nelson, 1984:293]

Recently workers in molecular systematics, parasitology, and biogeography have recognised that their disciplines may all be attempting to solve a common problem, namely reconstructing the history of an association between an "associate" (such as a gene, a parasite, or an organism) and its "host" (such as an organism, a parasite's host, or an area) (e.g., Baum, 1992; Doyle, 1992; Page, 1993a). That there is a general problem raises the possibility of a general solution, a methodology that can be used to examine associations at all levels among genes, organisms, and areas. My purpose in this paper is to describe one candidate for a general method. This method emerged independently (and in various stages of development) in molecular

systematics (Goodman et al., 1979), parasitology (Mitter and Brooks, 1983; Humphries et al., 1986), and biogeography (Nelson and Platnick, 1981), and was motivated by the realisation that some incongruence between host and associate cladograms might be more apparent than real.

Fundamental to any study of historical association is the relationship between the genealogies of members of the association. In one sense this is analogous to the problem in systematics of inferring the history of character change on a tree. Given a character whose states varying among a group of taxa, together with some hypothesis of the relative weight or cost of changes between those character states (such as a character state tree or a step matrix) then the goal is to find the assignment(s) of character states to internal nodes of the tree that minimizes the summed cost of character state changes (see Maddison and Maddison, 1992). Indeed, this analogy is the basis of Brooks' (1981) method for inferring host phylogeny from parasite phylogeny ("Brooks Parsimony Analysis", BPA), which has subsequently been applied to biogeography (e.g., Wiley, 1988a, 1988b). BPA treats the parasites as character states and the parasite phylogeny as a character state tree. The most parsimonious reconstruction of this character state tree on the host tree describes the history of the host-parasite association (Brooks and McLennan, 1991). However, in treating parasites as character states rather than as lineages, BPA can produce anomalous reconstructions requiring considerable post hoc interpretation (Wiley, 1988b). Because of these problems (discussed further below) I do not consider BPA in its present form to be an adequate method for studying historical associations (see also Page, 1990a).

This paper is an attempt to develop a method for recovering the history of associations among genes, organisms, and areas. Underlying the method is the concept of a map between two trees, introduced by Goodman et al. (1979) to help interpret incongruence between trees for vertebrate globin genes and trees for the vertebrates based on morphological data. This paper is an attempt to develop this concept further, in particular to elaborate on the use of "reconciled trees" (Page, 1990a; 1993b) to visualise the map. The problem of incorporating horizontal transmission of associates

(e.g., dispersal or host switching) is also addressed by introducing the concept of maximizing the amount of codivergence (shared history) between the associates.

ANALYSING HISTORICAL ASSOCIATIONS

Genes, Organisms, and Areas

In this paper I shall use the term "host" to mean any entity that in some sense "harbors" another entity, which I shall call the "associate." Examples of hosts (and their associates) are organisms (genes), host organisms (parasitic organisms), and areas (organisms). The parallel divergence of a host and its associate will be called "codivergence." If the associate differentiates in situ independently of the host then each instance of differentiation is a "duplication." If an associate leaves its host or extends its range to include other hosts this is "horizontal transfer." The absence of an associate where the reconciled tree predicts it should occur is a "loss."

The three "host"–"associate" assemblages involve quite different entities. I do not claim that all processes found in one system have their exact analogues in the other two systems, nor do I claim that the perhaps clumsy terminology of "host" and "associate" need apply equally well to all three. Rather, I suggest that there is sufficient commonality to justify attempting to develop a single method that encompasses all three cases. Certainly, the notion of a parallel between areas and taxa, and hosts and parasites is not new (e.g., Hennig, 1966; Brooks, 1981), although the proposal that the relationship between gene trees and organismal trees might be a third instance of the same problem is a more recent development (e.g., Baum, 1992; Doyle, 1992; Page, 1993a).

The processes corresponding to "codivergence," "dispersal," "duplication," and "loss" in each system are as follows:

Genes and organisms.—Codivergence corresponds to the acquisition of sequence difference following cladogenesis. A duplication may be literally a gene duplication giving rise to two paralogous genes, or may be the intraspecific differentiation of a gene giving rise to more than one allele. Horizontal gene transfer, for example by introgressive hybridization, constitutes "dispersal."Loss events include gene deletion and allele extinction.

Parasites and hosts.— Codivergence between host and parasite corresponds to strict cospeciation (Lyall, 1986). Duplications result from independent speciation of the parasites. Horizontal transfer from one host to another is host switching (=secondary infestation). Losses correspond to extinction of parasites.

Organisms and areas.—Codivergence between areas and organisms corresponds to classical vicariance. Duplications correspond to speciation of organisms independent of the vicariance of areas (for example a clade differentiating in response to some ecological differentiation within an area, and the descendants attaining broad sympatry prior to geological differentiation of that area). Horizontal transfer is dispersal (Platnick, 1976). Losses correspond to extinction of taxa.

Analytical Goals

Taking host-parasite assemblages as the paradigm example of an historical association, the first goal of an analysis is to establish which events in the phylogeny of the parasite correspond to which events in the host phylogeny. This requires a map between the two trees. Figure 1 shows a simple map between two congruent host and parasite trees. Each node in the associate trees has been mapped onto the most recent node in the host tree whose descendants host all the descendants of the associate node. For example, the descendants of node 6 are 2, 3, and 4, which are hosted by b, c, and d. The most recent common ancestor of these three hosts is node f, and so node 6 maps onto node f (Fig. 1b).

Constructing a map allows us to identify which cladogenetic events in the two clades are cospeciation events, and which are not. In the trivial example in Figure 1 all three cladogenetic events (nodes e, f, and g) are cospeciation events. This information can be used in various ways, perhaps the simplest being as the basis for quantifying the amount of cospeciation between host and parasite. Such an index can be computed for any pair of host and parasite trees, and hence could form the basis of a statistical test of a hypothesis of cospeciation (Page, 1990a, 1990b). A measure of fit between host and parasite trees could also be used as an optimality criterion for inferring host phylogeny from the phylogeny of one or more of its parasites when we have no information about the former (Brooks, 1981).

Beyond generating simple summary statistics the map between host and parasite trees provides the framework for comparative studies of evolution in the two clades. Studies of "coadaptation" (Brooks and McLennan, 1991) between hosts and parasites require that homologous evolutionary events in the two clades can be identified. This problem does not arise in comparative studies of the evolution of different attributes with the same clade (e.g., Harvey and Pagel, 1991; Harvey and Purvis, 1991) because the attributes are evolving on the same tree. Meaningful comparative studies of character evolution in host-parasite systems depend on the existence of a map specifying the relationship between the two trees. Without some form of map there is no basis for comparison. This requirement motivated my critique of previous (Hafner and Nadler, 1990; Page, 1990b) comparisons of genetic divergence between chewing lice and their pocket gopher hosts (Page, 1991). The map between the louse and gopher phylogenies specified which nodes in the louse phylogeny where putative cospeciation events and therefore could be compared with the corresponding nodes in the host tree.

Maps Between Trees

The example given in Figure 1 serves to illustrate a map between two trees, but is otherwise trivial. A more complicated example is shown in Figure 2. Here the parasite and host trees are incongruent because parasites 2 and 3 are monophyletic but their hosts b and c are paraphyletic. Yet, the two trees show some agreement in that parasites 2, 3, and 4 are monophyletic as are their hosts. How might we interpret the history of this host-parasite assemblage?

One approach is to ask under what circumstances the two trees could actually be congruent, that is, can the observed host and parasite cladograms be explained solely by "association by descent" (Mitter and Brooks, 1983)? If indeed cladogenesis in the parasite clade is a function of cladogenesis in the host clade then we would have to postulate that the observed parasite cladogram is a subtree of a larger tree (Fig. 2d). This larger tree is the reconciled tree (Page, 1990a). It can be thought of as representing the complete cladogram for the parasites, of which the observed parasite tree is a subtree. It might be that our sample of parasites is sparse, or that several parasites have gone extinct leaving only the observed relict pattern. In this example parasites 2 and 3 could be relicts of a larger clade comprising three parasites, one of which (on host d) is either present but uncollected or it is extinct. Likewise, parasite 4 is the relict of a clade of three parasites, two of which are now unknown or extinct. If we substitute a gene family for the parasite in Figure 2 and replace the host with the organisms from which the genes were sampled then we could interpret the reconciled tree to mean that genes 2 and 3 are paralogous with respect to gene 4, and that both sets of genes are the descendants of a gene duplication represented by node 6 (Goodman et al., 1979).

The map (Fig. 2b) between the two trees is constructed in exactly the same way as before; each node in the parasite tree has been mapped onto the most recent node in the host tree whose descendants host all the descendants of the associate node. In this example, two nodes in the parasite tree map (6 and 7) onto the same node (f) in

the host tree, hence the map between the two trees is not one-to-one. If a node in the associate tree and its ancestor map onto the same node in the host tree then, if we are to insist that the two trees are in fact congruent (i.e., the association has been strictly by descent) then we must postulate a duplication giving rise to two sympatric associate lineages that both track the same hosts. In Fitch's (1970) terminology these two lineages are paralogous.

Visualising the Map

The map between the two trees in Figure 2 can be represented in at least two ways. The parasite tree can be superimposed on the host tree in such as way that the nodes in the parasite tree are paired with their corresponding nodes in the host tree (Fig. 2c). Alternatively, the map can be depicted using a reconciled tree (Fig. 2d). In this paper I shall focus on reconciled trees and the statistics that can be derived from them. Superimposed trees will be explored elsewhere.

MAPS AND RECONCILED TREES

The map between host and associate trees defines a reconciled tree. A reconciled tree represents the most parsimonious interpretation of the data subject to the constraint of no horizontal transmission. If one is going to defend an hypothesis of strict cospeciation then the reconciled tree makes explicit the "cost" of that hypothesis. In this section I shall develop formal definitions of the map and its associated reconciled tree, before discussing various measures of fit between host and associate trees derived from the reconciled tree.

Maps

A map between the associate tree A and the host tree H is a function that maps each node in A onto a node in H. Each node in the associate tree is assigned a distribution, D, corresponding to the set of hosts occupied by that associate. For a terminal node D is the observed host distribution, for an internal node D is the union of the distribution of all the descendants of that node. Each node in the host tree has an associated cluster (=component, Nelson 1979) comprising the set of terminal descendants of that node. The map between A and H is constructed by finding for each node in the associate tree the smallest cluster in the host tree that contains the set representing the distribution of the associate. In more formal terms, the map f(A, H) between associate tree A and host tree H is given by the function $\text{lub}_H(D_i)$ for all nodes in A, where D_i is the distribution of the ith node in A, and lub_H is the least upper bound of the set D_i in H.

A Digression on Dollo Parsimony

Before describing the construction of the reconciled tree it is worth noting that the method described here for mapping one tree onto another is intimately related to the Dollo parsimony method for optimizing characters on a cladogram (Farris, 1977; Felsenstein 1979). Consider the cladogram T = (((A,B),(C,D)),E) and a binary character with the distribution 10010. We can also represent this character distribution as the set {A, D} of taxa with the derived character state (1). Under Dollo parsimony, the derived state can only arise once, although it can be lost many times. A most parsimonious reconstruction of a binary Dollo character on a tree requires finding the most recent ancestral node of all taxa with the derived state, in this case ABCD. This node is precisely the least upper bound in T of the set of terminal taxa with the derived state. Mapping two trees onto each other is equivalent to representing the distribution

of each node of the associate tree as a binary character, optimizing those characters onto the host tree using Dollo parsimony, and then noting below which nodes the 0 → 1 character state change occurs.

A Paradox in Dollo Parsimony

Felsenstein (1979:59) noted a paradox in the Dollo method—for some multistate characters we cannot assign any single character state to each ancestral node and still invoke a single origin of each derived character state. As an example, the character in Figure 3a with the character state tree ((2,3)1)0 cannot be optimized onto the tree shown in Figure 3b, without requiring more than one origin of either state 2 or 3.

Felsenstein (1979) notes that a solution to this paradox is to allow polymorphic ancestral taxa. By allowing the ancestral taxa to have both character states 2 and 3, we can still satisfy the requirement that states 2 and 3 arose just once. The terminal taxa A, B, C, and D have each lost either state 2 or 3, becoming monomorphic. In Wagner parsimony reconstructions there may be many possible character state assignments to an ancestral node but each reconstruction requires only a single character state to be assigned to any node (Swofford and Maddison, 1987). In the Dollo example given above, the reconstruction requires more than one state to be assigned to some ancestral nodes in any most parsimonious reconstruction.

When mapping two trees the equivalent problem occurs when two associate nodes map onto the same host node. Likewise, a solution is to postulate polymorphism, in this case the presence of two lineages of associates on the same host.

RECONCILED TREES

The map between host and associate trees defines a reconciled tree. In this section I shall develop a formal definition of a reconciled tree, before discussing various measures of fit between host and associate trees.

M-Trees

Reconciled trees differ from the kinds of trees usually employed in systematics, so it is useful to define a new kind of tree to accommodate them. Let the set $S = \{x_1, x_2, ..., x_n\}$. A multiset, M_S, derived from S is a set $\{x_1^{v_1}, x_2^{v_2}, ..., x_n^{v_n}\}$ where v_i is the number of times the ith element of S occurs in M. For example if $M = \{a^1, b^2, c^2, d^2\}$ then $M = \{a, b, b, c, c, d, d\}$. The set S is the base set of M_S, and is written S_M. For another, related application of multisets in cladistics see Minaka (1990). An m-tree T is a set of subsets of M satisfying the conditions:

1. $M \in T$, $\varnothing \notin T$.
2. $\{i\} \in T$ for all $i \in M$.
3. $A \cap B \in \{\varnothing, A, B\}$ for every $A, B \in T$.
4. Let X_i be the cluster of node i. For any two immediate descendants, a and b, of node c, $S_{X_a} \cap S_{X_b} = \varnothing$, or $S_{X_a} = S_{X_b} = S_{X_c}$.

Conditions 1-3 are inherited from the definition of a n-tree (e.g., Margush and McMorris, 1981). A n-tree (Fig. 4a) can be regarded as a special case of an m-tree in which $v_i = 1$ for all $\{i\} \in M$. Condition 4 ensures that the base sets of the clusters of each node in an m-tree are either disjoint or identical. Figure 4b shows an example of an m-tree on the multiset $M = \{a, b, b, c, c, d, d\}$.

Reconciled Trees

Let T_H be an n-tree on the set of hosts H, and T_A be the set of subsets of H implied by the associate tree (note that T_A need not be an n-tree). If we use the short-hand notation {b, c, d} ≡ bcd, then in the example shown in Figure 2, T_H = {abcd, bcd, cd} and T_A = {abcd, bcd, bc}. The reconciled tree, T_R, for T_H and T_A is the smallest m-tree that (1) contains all and only the clusters of T_H, and (2) contains T_A as a subtree (Fig. 5). For this example, T_R = {abcd, bcd, bcd, cd, cd}, which can also be written as {abcd, bcd^2, cd^2}. Although an m-tree need not be binary (i.e., fully resolved), in this paper I shall restrict my attention solely to binary reconciled trees. An algorithm for computing a reconciled tree is given in the Appendix 1.

Consensus Supertrees

Reconciled trees should not be confused with the superficially similar consensus supertrees (Gordon, 1986). Consensus supertrees are a means of combining two trees for two overlapping sets of taxa into a single tree for the combined set of taxa. The two trees being combined are both subtrees of the larger tree. One could use a consensus supertree, for example, to combine a tree for a group of taxa based on larval data with a tree based on adult morphology, where some taxa lacked larval data, and others lacked adult data. In contrast, a reconciled tree is an estimate of the larger tree that only one tree (the associate tree) is a subtree of, and this larger tree is constrained to have all and only the clusters of the host tree. Rather than summarising two estimates of the same tree, a reconciled tree describes how one tree "fits into" another.

MEASURES OF FIT

In some cases the observed tree for the associates will be identical to the reconciled tree (for example, if we have collected all the descendants of a gene duplication). However, in many cases the two trees will differ. It would be desirable to have some quantitative measure of the discrepancy between the two trees. The following measures were proposed by Goodman et al. (1979) and Nelson and Platnick (1981).

Duplications

Perhaps the simplest measure of fit is the number of duplications in the reconciled tree. Each duplication represents a divergence event that took place in the associate lineage "independently" of the hosts, for example a gene duplication, or sympatric speciation of parasites. The greater the number of duplications the smaller the number of codivergence events. For example, duplications in a parasite lineage reduces the number of hypotheses of cospeciation, and in a gene phylogeny duplications reduce the number of nodes that are potentially informative about organismal phylogeny. This point is elaborated on below.

Goodman et al. (1979:138) distinguished between two kinds of gene duplications (GDs):

> …hypothetical GDs—that is, GDs for which there is no other evidence than the nonconformity of the species phylogeny with the hypothetical gene phylogeny under consideration—and real GDs—that is, GDs that can be inferred independently, from the existence of related non-allelic loci in individuals of one or more species; e.g. the separate gene loci

> encoding beta and delta hemoglobin chains in humans and other hominoid species.[emphasis in original]

In cladistic biogeography (cospeciation) studies, the equivalent evidence for a "real" duplication is overlap in the geographic (host) ranges of the descendants of the same node. Page (1988:269) termed such nodes "redundant."

Leaves Added

Nelson and Platnick (1981:417) measured the degree of fit between two trees as the number of nodes added to the host tree to create the reconciled tree which they termed "items of error." Page (1990a) generalized this to the difference between the number of nodes in the associate and reconciled trees, $|T_R| - |T_A|$, which is always even since $|T_R|$ and $|T_A|$ are always odd. In the example shown in Figure 5 there are 13 - 7 = 6 items of error. The number of leaves added to the reconciled tree is half the items of error. While this measure is easy to compute it doesn't have a straightforward biological interpretation.

Minimum Number of Losses

Simply counting the number of leaves added to reconcile two trees may overestimate the number of actual events involved (Page, 1988:260). The absence of associates from a clade of hosts may be due to a single loss (or extinction event). Hence the third measure of fit is the minimum number of losses required to explain the distribution of associates. The reconciled tree shown in Figure 5 requires three losses.

An Example

There are three lactate dehydrogenase isozymes in vertebrates two of which (LDH-A and LDH-B) are found in all vertebrates, whereas the third (LDH-C) is known only in actinopterygian fishes, columbid birds, and mammals. Figure 6 shows the gene tree for 22 lactate dehydrogenase sequences published by Quattro et al. (1993), a tree for the organisms from which the sequences were obtained, and the reconciled tree. The reconciled tree depicts the ancestral gene duplication (labeled 1 in Fig. 6) giving rise to the two loci Ldh-A and Ldh-B prior to the divergence of the vertebrates, and the two duplications giving rise to the mammalian Ldh-C and actinopterygian Ldh-C loci (labeled 2 and 3, respectively). That 13 lineages lack LDH (hollow branches in Fig. 6) reflects the limited number of sequences available, rather than actual absence of Ldh loci. That the origin of mammalian LDH-C predates the avian-mammal divergence suggests that columbid bird LDH-C may be orthologous with mammalian LDH-C (Quattro et al., 1993:245).

The organismal tree in Figure 6 is one that minimizes the number of leaves added in order to construct the reconciled tree. This tree contradicts current views of vertebrate phylogeny (e.g., the teleost is grouped with lamprey rather than as the sister group of tetrapods). Quattro et al. (1993:244) note that "if this [vertebrate tree] topology is correct, multiple origins and losses of what we currently refer to as LDH-A within the vertebrates would have to be postulated," although they caution that placement of teleost LDH-A is not strongly supported by their data. The cost of alternative LDH and vertebrate phylogenies in terms of gene duplications and losses can be readily evaluated by computing a reconciled tree for each combination of gene and organismal tree.

INTERPRETING RECONCILED TREES

Widespread Associates

Several authors have discussed the interpretation of widespread associates (i.e., associates found on two or more hosts) (e.g., Platnick and Nelson, 1978; Zandee and Roos, 1987; Kluge, 1988; Wiley, 1988a, 1988b; Page, 1989a; 1990a), primarily focusing on widespread taxa in biogeography. This discussion has been marred by a conceptual confusion between an associate and its distribution (Page, 1989a). A recent example of this confusion is Brooks' (1990:18-20; see also Brooks and McLennan, 1991:215-217) discussion of a hypothetical biogeographic example purporting to show "relationships supported by the area cladogram that are inconsistent with the original estimates of phylogeny" (Brooks and McLennan, 1991:215).

Brooks' (1990) example (Figure 7b) shows the distribution of four taxa and their hypothetical ancestors mapped onto an area cladogram. The distribution of taxon 1 is mapped onto the area cladogram below the distributions of taxa 2, 3, and 4, which Brooks (1990:19) interprets as requiring taxon 1 to be ancestral to taxa 2-4 which conflicts with the phylogeny (Fig. 7a). However, what is being mapped is not a taxon but its distribution. The area cladogram in Figure 7b suggests that taxon 1 occurred in all four areas before those areas differentiated, and that since taxon 1 predates that differentiation, the ancestor of the clade (2,(3,4)) must also have been present in all four areas. Hence Figure 7 implies the sympatry of these two clades, not that taxon 1 is ancestral to taxa 2-4.

Terminal Widespread Associates

In one sense the focus on widespread terminal taxa has been misplaced (cf. Nelson and Ladiges, 1992). Nelson and Platnick's (1981) argument that a combination of geographic proximity and failure to speciate can lead to taxa having phylogenies seemingly incongruent with the history of the areas in which they occur applies equally to ancestral taxa. It is quite possible that area cladograms for two different taxa that comprise only endemics may be incongruent because of a previous history of persistence of widespread taxa followed by subsequent speciation.

That said, however, widespread terminal species do present a problem. If a species does not differentiate then its distribution may include quite unrelated areas (Platnick and Nelson, 1978; Page, 1989a, 1990a). Three general approaches to handling widespread associates suggest themselves. Firstly, we could simply map the widespread associates just as we would any other node. This is essentially the procedure adopted by Zandee and Roos (1987) and Wiley (1988a; 1988b). Alternatively we could choose to map just the endemic terminal taxa and the internal nodes, omitting the widespread terminal taxa (but ensuring that the range of the ancestor of a widespread terminal taxon includes the range of that taxon). This is close in spirit to Nelson and Platnick's (1981) Assumption 1. Lastly, we might map just part of the terminal node's distribution. This is effectively Nelson and Platnick's Assumption 2.

Rejecting Hypotheses of Association

Because it is possible to reconcile any two trees no matter how incongruent one might ask whether it is possible to reject the hypothesis of association upon which the method is predicated. One approach (Page, 1990a, 1990b) to testing an hypothesis of association is to compare the observed fit between the associate and host trees with the fit expected if the associate tree was drawn at random from the set of possible

trees. If the fit is no better than we could expect due to chance alone then the hypothesis of association can be rejected. Of course, it may be that a more modest hypothesis allowing a mixture of some association and some horizontal transmission is more reasonable. Rejection of the hypothesis of association for all the associates together does not exclude the possibility that a subset of the associates have codiverged with their hosts (Page, 1990b).

Because the reconciled tree and its associated map make predictions about the relative ages of divergence events in the host and associate lineages if we have estimates of these times of divergence then we can test these predictions. While this approach is limited by the difficulty of estimating divergence times it has been applied to host-parasite (Page, 1990b) and biogeographic (Page, 1993a) data.

HORIZONTAL TRANSMISSION

Perhaps the most glaring limitation of the method described here is that reconciling two trees explicitly assumes that all associations arise by descent, that is, by vertical transmission. This a priori assumption excludes horizontal transmission. The possibility of horizontal transmission not only raises the problem of assigning relative weights to vertical versus horizontal events (Sober, 1988) but plays havoc with the underlying assumption that the association can be accurately modelled by a tree.

The difficulties in attempting to combine both vertical and horizontal transmission in a single method can be illustrated using Brooks' Parsimony Analysis (Brooks and McLennan, 1991) which uses Wagner parsimony to map parasite cladograms onto host cladograms. Figure 8a shows a cladogram for five parasites 1-5 found on four hosts A-D. Let us further suppose that the occurrence of parasite 5 in host A is due to dispersal from host D. By coding the parasite cladogram as a suite of

binary characters (Table 1) and then optimizing those characters onto the host tree using DELTRAN optimization (see Wiley, 1988b:278), we obtain the result in Figure 8b. Parasite taxa 6 and 7 occur twice on the tree due to the presence of parasite 5 in host A. These two instances of "homoplasy" are due to the dispersal of a single terminal taxon. Obviously it is illogical to require, for example, the common ancestor of parasites 3-5 (i.e., parasite 7) to have dispersed along with parasite 5. This is an artefact of the method (Page, 1990a). The coding of the parasite cladogram in Table 1 assumes that the range of each ancestral taxon includes the range of all of its descendants. However, dispersal of a taxon onto new hosts can result in the range of the descendants being greater than the range of the ancestors. As a consequence, the codes for parasites 6-8 all include host A (Table 1). Not only does this require two instances of homoplasy, but also results in the inference that parasite 8 infested the common ancestor of all four hosts, rather than the ancestor of hosts B-D, as implied by Fig. 8a.

This problem is not confined to parasitology and biogeography. Doyle (1992) has recently argued that since nucleotide sequence data constitute evidence of gene phylogeny (and only indirectly organismal phylogeny), those data may be best treated as a single, highly resolved character-state tree when inferring organismal phylogeny, rather than using the "raw" sequence data (see also Baum, 1992). This provocative view raises the problem of handling gene trees. Doyle's (1992) "Assumption N" method uses BPA to code the gene tree creating a suite of binary characters that can be combined with other kinds of characters (as well as other gene trees). If horizontal transmission of genes has occurred then the problems encountered in the parasitological example discussed above will also arise.

The tree-mapping method described in this paper also assumes that the range of each ancestral taxon includes the range of the descendants. However, in constructing the reconciled tree, horizontal transmission is ruled out (one could achieve similar results using Dollo parsimony instead of Wagner parsimony, see above). While this

avoids the illogical results that can be obtained with BPA, it does so at the expense of a priori eliminating one possible explanation of the observed patterns — dispersal.

Maximizing Hypotheses of Codivergence

Is there a solution to the problem of incorporating dispersal? Obviously we can introduce as many dispersals as needed to render congruent otherwise incongruent host and associate trees, just as we can always postulate sufficient duplications to achieve the same ends without postulating dispersal (although, as outlined above, we can use a randomization test to decide whether the number of duplications required is excessive). One way out of this difficulty would be to develop a measure that incorporated both explanations of incongruence, and which allowed us to quantitatively assess the merits of a particular hypothesis.

One candidate for such a measure is to define the "best" reconstruction of the history of an association as that which maximizes the number of codivergent events. Humphries et al. (1986:61) wrote of this approach "[a]dmittedly this gives an a priori bias towards coevolution, but [it] has the merit of choosing a result in terms of explanatory power." That is, it seeks to explain the greatest possible amount of history as shared history (i.e., due to a common cause), a goal one could consider analogous to Hennig's (1966) auxiliary principle in systematics (e.g., Brooks and McLennan, 1991:223).

Interestingly, although a reconciled tree is constructed under the assumption of association by descent it need not maximize the number of hypotheses of codivergence, despite the fact that it does not allow dispersal events. Because each duplicated node in the reconciled tree is not a codivergent event but an independent event in the associate, only those nodes in the reconciled tree that are not duplications correspond to codivergent events. I shall term these nodes codivergent nodes.

How then, do we identify associates that have dispersed? Obvious candidates are those associates that cause the greatest degree of incongruence between host and associate trees. These associates can be readily identified by deleting each associate in turn and computing a reconciled tree for the remaining associates. By deleting a associate before constructing the reconciled tree we also remove the range of that associate from the range of its ancestors, thus avoiding the problem illustrated in Figure 8.

Pocket Gophers and Their Lice

The principle of maximizing the number of codivergence events can be illustrated using Hafner and Nadler's (1988, 1990) pocket gopher-chewing lice data. Hafner and Nadler (1988) obtained cladograms for eight pocket gophers and their 10 parasitic lice. Page (1990b) published a reconciled tree for these taxa that required four duplications and at least 10 independent extinctions of the lice. Since a duplicated node reflects a speciation event in the lice that happened prior to the differentiation of the corresponding host there are only 9 - 4 = 5 nodes in the louse cladogram that are postulated to reflect cospeciation events.

If we relax the constraint of strict association by descent we can increase the number of codivergent nodes for the gophers and lice. Hafner and Nadler (1988) suggested that the lice Geomydoecus actuosi and G. thomomyus had dispersed from the gopher Geomys bursarius onto gophers of the genus Thomomys. If we omit those two lice and reconcile the resulting lice tree with the gopher tree we obtain a much smaller reconciled tree that requires only a single duplication and three extinctions (Fig. 9a), leaving six nodes in the lice tree that are postulated to be due to cospeciation, an increase by one node over the result when no dispersal is allowed (the seventh node is a duplication). We could also remove G. setzeri, obtaining a reconciled tree with no duplications, but this reduces the number of codivergent nodes back down

to five, hence postulating two dispersals increases the amount of codivergence, postulating three dispersals does not.

This example illustrates that if we seek to maximize the number of hypotheses of cospeciation we need not be required to explain every association by descent — indeed allowing for horizontal transmission can increase the number of cospeciation events. Moreover, the criterion allows us to decide when to stop invoking dispersals; once the extent of codivergence has been maximized nothing is gained by postulating more dispersals.

Note that there may be more than one maximal set of codivergent nodes. For the gophers and lice we can also obtain six codivergent nodes by deleting the lice Geomydoecus ewingi and G. setzeri (Fig. 9b). A codivergent node in an associate tree and its corresponding node in the host tree are postulated to be homologous, that is, they reflect the same evolutionary event. It is upon these nodes that comparisons of rates of evolution, and the relative timing of speciation in hosts and parasites should be based (see Page, 1991: 190). Different sets of codivergent nodes might have quite different implications for studies of the rate and timing parameters of the host-parasite assemblage.

Deleting associates is one, admittedly crude, solution to incorporating dispersals. Although it allows an easy measure of the number of dispersals (= the number of associates deleted) it has the drawback of no longer describing the history of the associates that dispersed. For example, a parasite might disperse early in its history then subsequently track its host. In this instance deleting the descendants of that parasite would underestimate the amount of codivergence that actually occurred. Clearly there is scope for developing more sophisticated treatments of dispersal. In particular, sequential deletion of associates that cause the greatest amount of incongruence is not guaranteed to find any or all of the sets of dispersed associates that maximize the amount of codivergence.

Orthologous Gene Trees

Although gene trees and species trees may be incongruent due to the presence of paralogy (the original motivation for the concept of reconciled trees), phylogenies of orthologous genes can also be incongruent with species phylogenies due to sorting of ancestral polymorphism. This problem has received considerable theoretical attention (e.g., Pamilo and Nei, 1988; Takahata, 1989; Wu, 1991). When dealing with orthologous genes "duplications" are not literally gene duplications and "losses" are due to allele extinction or lineage sorting. However, the reconciled tree still contains useful information in interpreting orthologous gene trees.

Inferring organismal phylogeny from gene trees for loci with extensive transspecific polymorphism (such as the major histocompatibility complex, [Klein and Klein, 1991]) can be difficult. Takahata (1989:957) notes that

> . . . a phylogenetically informative event in a gene tree constructed from nucleotide differences consists of interspecific coalescences of genes in each of which two genes sampled from different populations are descended from a common ancestor.

It should be pointed out that not all interspecific coalescences are phylogenetically informative; a given coalescence, c(A,B), between genes from taxa A and B will only be informative if there has not been a coalsecence between any genes in A and B that are themselves descendants of c(A,B).

Given a gene tree and an organismal tree the phylogenetically informative coalescences are exactly the codivergent nodes on a reconciled tree. In the absence of other information (such as time of divergence among the alleles) we could use the number of informative nodes as a criterion for choosing between competing species trees based on a gene tree.

FINDING HOST TREES WHEN THEY ARE UNKNOWN

So far this paper has presupposed that we have a known host tree. What if we do not have a host tree, as is often the case in biogeography where our aim is to find the host tree (in this case an area cladogram)? Given that we can reconcile an associate tree with any host tree and compute one or measures of fit, we can proceed as we would for character data — search for the host tree (or trees) with the best fit to the associate tree (Page, 1990a). For small numbers of hosts we can enumerate all possible host cladograms (either explicitly or implicitly using a branch and bound search); for larger trees we must rely on heuristics.

Rosen's (1978) Data Revisited

Rosen's (1978) data on the fishes Heterandria and Xiphophorus have often been used to illustrate cladistic biogeographical methods. As the algorithms I used earlier (Page, 1988; 1989a) to analyse Rosen's data differ from those presented here (see Appendix 2), I shall now illustrate how the method described in this paper can be applied to Rosen's data, and show that it can reproduce the result obtained in those previous studies and by Platnick (1981).

Taking Rosen's (1978) taxon cladograms and distributions and searching for the area cladogram that minimizes the number of leaves that must be added to reconcile the two fish cladograms with an area cladogram, we obtain the area cladogram in Figure 10a. This cladogram requires a total of 32 leaves added, 10 for Heterandria and 22 for Xiphophorus. (In exactly the same way as the most parsimonious taxon cladogram for a given data set is the tree minimizing the sum of character state changes for each character on that tree, the most parsimonious area

cladogram for a set of taxa is the tree that minimizes the sum of leaves that must be added [for example] to reconcile each taxon cladogram with that tree.)

These 32 leaves added correspond to at least 18 independent extinctions, which seems an excessive burden in order to maintain that the two fish are geographically congruent. However, further analysis shows that the area cladogram for the two fishes combined is not optimal for each fish taken separately. Computing the optimal area cladograms for each taxon separately, we find one for Heterandria (leaves added = 1; Fig. 11a) and 15 for Xiphophorus (leaves added = 3). The 15 trees for Xiphophorus are identical except for the placement of area 7 which lacks any Xiphophorus and hence is free to float anywhere on the cladogram (Fig. 11b shows the Adams consensus of these 15 trees). The two trees agree only on the relationships of areas 1, 2, 45, 8, and 10. This is the same result obtained by Rosen (1978). Ignoring area 7 for which Xiphophorus is uninformative, we find the two cladograms differ on the relationships of areas 3, 6, and 9. As Platnick (1981) noted, these areas are all part of the range of widespread taxa in one or the other clade (but not both), raising the possibility that the conflict between the area cladograms for the two fish genera is due to geographically adjacent but cladistically unrelated areas sharing the same taxon.

If we delete from the range of each widespread taxon those areas about whose relationships the genera disagree (i.e., 3, 6, and 9), then the relationships of those areas will be determined by the relationships of the taxa endemic to those areas. Analysing this modified data set we obtain the three area cladograms (Fig. 10b) reported by Platnick (1981) and Page (1989a), which are also optimal for each fish taken separately. Hence, if we allow that the geographical proximity between areas 2 and 3 (sharing H. bimaculata), 45 and 6 (sharing X. alvarezi), and 9 and 10 (sharing X. "PMH") may have resulted in unrelated areas sharing the same taxon, Heterandria and Xiphophorus have congruent area cladograms.

DISCUSSION

The claim of generality made for the methodology described in this paper comes in part from its independent origins in molecular systematics (Goodman et al., 1979) and biogeography (Nelson and Platnick, 1981) and from recent treatments (Baum, 1992; Doyle, 1992) of gene trees using methods developed in biogeography and parasitology (Brooks, 1981). The methodology is in its infancy and has been applied in few empirical studies (e.g., Page, 1990a, 1990b; Paterson et al., 1993).

Its applicability to any given system depends on there being some history of the associate tracking its host. Genes track organisms with a high degree of fidelity. The gopher-lice assemblage studied by Hafner and Nadler (1988) and Page (1990b) is the paradigm example of a parasite closely tracking its host, even so, the available data suggest that at least some lice have still switched hosts, so that a combination of vertical and horizontal transmission must be postulated. In one sense the reconciled tree constitutes one bound on the hypothesis of historical association — it represents the most parsimonious interpretation of the data subject to the constraint of no horizontal transmission. If one is going to defend an hypothesis of strict cospeciation then the reconciled tree makes explicit the "cost" of that hypothesis.

The suggestion of incorporating horizontal transmission by deleting associates to maximize hypotheses of codivergence is somewhat ad hoc, nor is it guaranteed to produce an optimal result. But it does offer the prospect of being able to accommodate both vertical and horizontal processes. A method that combines both processes in a single one-step procedure would obviously be highly desirable. Brooks' Parsimony Analysis (Brooks and McLennan, 1991) valiantly attempts this but fails (Page, 1990a; Carpenter, 1992). Improved methods await development.

ACKNOWLEDGMENTS

For their comments on the manuscript I thank Jim Carpenter, Chris Humphries, and Nobuhiro Minaka. The detailed criticisms of the two anonymous referees, an associate editor, and Mike Miyamoto were particularly helpful in improving the manuscript. This research was funded by a Research Fellowship from The Natural History Museum, London.

REFERENCES

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

Brooks, D. R. 1981. Hennig's parasitological method: A proposed solution. Syst. Zool. 30:229-249.

Brooks, D. R. 1990. Parsimony analysis in historical biogeography and coevolution: Methodological and theoretical update. Syst. Zool. 39:14-30.

Brooks, D. R., and D. A. McLennan. 1991. Phylogeny, ecology, and behavior. Univ. Chicago Press, Chicago.

Carpenter, J. M. 1992. Incidit in scyllam qui vult vitare charybdim [Review of Brooks and McLennan, 1991]. Cladistics 8:100-102.

Doyle, J. J. 1992. Gene trees and species trees: Molecular systematics as one-character taxonomy. Syst. Bot. 17:144-163.

Farris, J. S. 1977. Phylogenetic analysis under Dollo's law. Syst. Zool. 26:77-88.

Felsenstein, J. 1979. Alternative methods of phylogenetic inference and their interrelationship. Syst. Zool. 28:49-62.

Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. Syst. Zool. 19:99-113.

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

Gordon, A. D. 1986. Consensus supertrees: The synthesis of rooted trees containing overlapping sets of leaves. J. Classif. 3:335-348.

Hafner, M. S., and S. A. Nadler. 1988. Phylogenetic trees support the coevolution of parasites and their hosts. Nature 332:258-259.

Hafner, M. S., and S. A. Nadler. 1990. Cospeciation in host-parasite assemblages: Comparative analysis of rates of evolution and timing of cospeciation. Syst. Zool. 39:192-204.

Harvey, P. H., and M. D. Pagel. 1991. The comparative method in evolutionary biology. Oxford University Press, Oxford, England.

Harvey, P. H., and A. Purvis. 1991. Comparative methods for explaining adaptations. Nature 351:619-624.

Hennig, W. 1966. Phylogenetic systematics. University of Illinois Press, Urbana, Illinois.

Humphries, C. J., J. M. Cox, and E. S. Nielsen. 1986. Nothofagus and its parasites: A cladistic approach to coevolution. Pages 55-76 in Coevolution and systematics (Stone, A. R., and D. L. Hawksworth, eds.). Clarendon Press, Oxford, England.

Klein, J., and D. Klein. 1991. Molecular evolution of the major histocompatibility complex (NATO ASI Series H:59). Springer-Verlag, Berlin.

Kluge, A. G. 1988. Parsimony in vicariance biogeography: A quantitative method and a greater Antillean example. Syst. Zool. 37:315-328.

Lyall, C. H. C. 1986. Coevolutionary relationships of lice and their hosts: A test of Farenholtz's Rule. Pages 77-91 in Coevolution and systematics (Stone, A. R., and D. L. Hawksworth, eds.). Clarendon Press, Oxford.

Maddison, W. P., and D. R. Maddison. 1992. MacClade: Analysis of phylogeny and character evolution. Version 3.0. Sinauer Associates, Sunderland, Massachusetts.

Margush, T., and F. R. McMorris. 1981. Consensus n-trees. Bull. Math. Biol. 43:239-244.

Minaka, N. 1990. Cladograms and reticulated graphs: A proposal for graphic representation of cladistic structures. Bull. Biogeogr. Soc. Jpn. 45:1-10.

Mitter, C., and D. R. Brooks. 1983. Phylogenetic aspects of coevolution. Pages 65-98 in Coevolution (Futuyma, D. J., and M. Slatkin, eds.). Sinauer Associates, Sunderland, Massachusetts.

Nelson, G. 1979. Cladistic analysis and synthesis: Principles and definitions, with a historical note on Adanson's Famille des Plantes (1763-1764). Syst. Zool. 28:1-21.

Nelson, G. 1984. Cladistic biogeography. Pages 273-293 in Cladistics: Perspectives on the reconstruction of evolutionary history (Duncan, T., and T. F. Stuessy, eds.). Columbia Univ. Press, New York.

Nelson, G., and P. Y. Ladiges. 1992. Three-area statements: Standard assumptions for biogeographic analysis. Syst. Zool. 40:470-485.

Nelson, G., and N. I. Platnick. 1981. Systematics and biogeography: Cladistics and vicariance. Columbia Univ. Press, New York.

Page, R. D. M. 1988. Quantitative cladistic biogeography: Constructing and comparing area cladograms. Syst. Zool. 37:254-270.

Page, R. D. M. 1989a. Comments on component-compatibility in historical biogeography. Cladistics 5:167-182.

Page, R. D. M. 1989b. COMPONENT user's manual, release 1.5. University of Auckland, Auckland.

Page, R. D. M. 1990a. Component analysis: A valiant failure? Cladistics 6:119-136.

Page, R. D. M. 1990b. Temporal congruence and cladistic analysis of biogeography and cospeciation. Syst. Zool. 39:205-226.

Page, R. D. M. 1991. Clocks, clades, and cospeciation: Comparing rates of evolution and timing of cospeciation events in host-parasite assemblages. Syst. Zool. 40:188-198.

Page, R. D. M. 1993a. Genes, organisms, and areas: the problem of multiple lineages. Syst. Biol. 42:77-84.

Page, R. D. M. 1993b. Parasites, phylogeny, and cospeciation. Int. J. Parasitol. 23 (in press).

Pamilo, P., and M. Nei. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568-583.

Paterson, A. M., R. D. Gray, and G. P. Wallis. 1993. Parasites, petrels and penguins: Does louse presence reflect seabird phylogeny? Int. J. Parasitol. 23 (in press).

Penny, D., M. D. Hendy, and M. A. Steel. 1992. Progress with methods for constructing evolutionary trees. Trends Ecol. Evol. 7:73-79.

Platnick, N. I. 1976. Concepts of dispersal in historical biogeography. Syst. Zool. 25:294-295.

Platnick, N. I. 1981. Widespread taxa and biogeographic congruence. Pages 223-227 in Advances in cladistics: Proceedings of the first meeting of the Willi Hennig Society (Funk, V. A., and D. R. Brooks, eds.). New York Botanical Garden, New York.

Platnick, N. I., and G. Nelson. 1978. A method of analysis for historical biogeography. Syst. Zool. 27:1-16.

Quattro, J. M., H. A. Woods, and D. A. Powers. 1993. Sequence analysis of teleost retina-specific lactate dehydrogenase C: Evolutionary implications for the vertebrate lactate dehydrogenase gene family. Proc. Natl. Acad. Sci., USA 90:242-246.

Rosen, D. E. 1978. Vicariant patterns and historical explanation in biogeography. Syst. Zool. 27:159-188.

Sober, E. 1988. The conceptual relationship of cladistic phylogenetics and vicariance biogeography. Syst. Zool. 37:245-253.

Swofford, D. L., and W. R. Maddison. 1987. Reconstructing ancestral character states under Wagner parsimony. Math. Biosci. 87:199-229.

Swofford, D. L., and G. J. Olsen. 1990. Phylogeny reconstruction. Pages 411-501 in Molecular systematics (Hillis, D. M., and C. Moritz, eds.). Sinauer Associates, Sunderland.

Takahata, N. 1989. Gene genealogy in three related populations: Consistency probability between gene and population trees. Genetics 122:957-966.

Wiley, E. O. 1988a. Parsimony analysis and vicariance biogeography. Syst. Zool. 37:271-290.

Wiley, E. O. 1988b. Vicariance biogeography. Annu. Rev. Ecol. Syst. 19:513-542.

Wu, C.-I. 1991. Inferences of species phylogeny in relation to segregation of ancient polymorphisms. Genetics 127:429-435.

Zandee, M., and M. C. Roos. 1987. Component-compatibility in historical biogeography. Cladistics 3:305-332.

Received 12 December 1992; accepted

APPENDIX 1

Algorithm for constructing a reconciled tree

This appendix describes the basic algorithms for constructing a map between two trees and for building the corresponding reconciled tree. Before describing the algorithms we need to define some operations on binary rooted trees. LEFT (T, n) and RIGHT (T, n) return the left and right descendants, respectively, of node n in tree T, if present; otherwise it returns **nil**. ISLEFTDESC (T, m, n) returns **true** if node m is a left descendant of node n in tree T, otherwise it returns **false**. COPYOFSUBTREE (T, n) returns a copy of the subtree in T that is rooted at node n. ADDBELOW (T, s, n) adds the subtree s below node n in tree T and returns the ancestor of n and s. IMAGEOF (T, n) returns the node whose cluster = $\text{lub}_T(X_n)$. DUPLICATIONAT (A, H, n) returns true if node n in tree H corresponds to a duplication in the associate tree A, otherwise it returns false. FINDNODEABOVE (T, n, X) returns the node in tree T that is a descendant of node n and has the cluster X. The notation $|X|$ means the cardinality of the set X (i.e., the number of elements in X). RANGE (n) returns the set of hosts occupied by node n or all its descendants.

Additional operations to store and retrieve nodes are also needed: CLEARSTACK (S) clears a push down stack, PUSH (n, S) inserts node n as the top entry of stack S, POP (S) deletes the top entry from the stack S, TOP(S) returns the top entry of stack (S), and EMPTY(S) returns **true** if there are no items in stack S. Two arrays are used to store the map: image and duplication. For the k nodes in the associate tree, image stores the corresponding node in the host tree, and duplication stores a Boolean variable indicating whether a duplication is required.

Map Between Associate and Host Trees

Procedure MAPTREES (A, H) computes the map between the associate (A) and host (H) trees and counts the number of duplications, leaves added, and losses required to reconcile the two trees.

```
procedure MAPTREES (A, H)
begin
    /* Initialise the map */
    for i = 1 to k do begin
        image[i] ← nil
        duplication[i] ← false
        end
    duplications ← 0
    leaves ← |H_root|
    losses ← 0
    MAP (A_root)
    leaves ← leaves – |A_root|
end
```

MAPTREES calls the recursive procedure MAP which visits each node in the associate tree in postorder (i.e., the descendants of each node are visited before the node itself is visited).

```
procedure MAP (a):
begin
    if a ≠ nil then begin
        MAP (LEFT(a)
        MAP (RIGHT(a))
```

```
/* Find corresponding node in host tree */
image[a]  ← IMAGEOF (H, RANGE(a))
if a is not a leaf then begin
        /* get immediate left and right descendants of a */
        l ← LEFT (a)
        r ← RIGHT (a)
        /* test for duplication due to redundancy */
        duplication ← RANGE(l) ∩ RANGE(r) ≠ ∅
        /* test for duplication due to immediate descendant mapping
        onto same node */
        duplication ← duplication or (image[a] = image[l])
                or (image[r] = image[a])
        /* update duplications, leaves added, and losses */
        if duplication then begin
                /* count duplications */
                duplications ← duplications + 1
                /* duplicating the subtree rooted at image[a] adds
                |image[a]| leaves to the reconciled tree. */
                leaves ← leaves + |image[a]|
                /* count losses in left subtree */
                count ← 0
                LOSSES (LEFT(image[a], RANGE[l])
                losses ← losses + count
                /* count losses in right subtree */
                count ← 0
                LOSSES (RIGHT(image[a], RANGE[R]))
                losses ← losses + count
                end
        else
```

```
                    if (a = A_root) then begin
                        /* if root does not require a duplication then
                        count losses for tree above root but below any
                        node corresponding to a duplication */
                        count ← 0
                        LOSSES (a, RANGE[a])
                        losses ← losses + count
                        end
                end
            end
end
```

To compute the number of losses MAP calls the procedure LOSSES (n, X) which traverses the subtree in tree H rooted node n and counts the number of alive → extinct transitions. The procedure does not traverse the host tree above any node that corresponds to a duplication in the associate tree.

```
procedure LOSSES (n, X, count)
begin
    if  n ≠ nil then begin
        if DUPLICATIONAT (A, H, n) then
            /* Don't traverse host tree above a node that corresponds to a
            duplication in A  */
            n is alive
        else begin
            /* Visit rest of tree */
            LOSSES (LEFT(n), X, count)
            LOSSES (RIGHT(n), X, count)
            if n is a leaf  then begin
```

```
                    /* Is associate present? */
                    i = X_n ∩ Y
                    if i ≠ ∅ then
                            n is alive
                    else    n is extinct
                    end
            else begin
                    /* internal node */
                    l ← LEFT (n)
                    r ← RIGHT (n)
                    if LEFT(n) is alive or RIGHT(n) is alive then begin
                            n is alive
                            /* If one or other (but not both) of the
                            descendants has no alive associate then there has
                            been a loss */
                            if Left(n) is alive xor RIGHT(n) is alive then
                                    count ← count + 1
                            end
                    else n is extinct
                    end
            end
    end
```

Algorithm RECONCILE

Algorithm RECONCILE (A, H, R) takes as input the associate and hosts trees, A and H, and returns the reconciled tree R. This procedure assumes that the map **between** the two trees has already been computed (see MAPTREES above). It uses

two stacks, SA and SR, to keep track of the duplications in the associate and reconciled trees.

```
procedure RECONCILE (A, H, R)
begin
        CLEARSTACK (SA)
        CLEARSTACK (SR)
        /* Initially the reconciled tree has same topology as the host tree */
        R ← COPYOFSUBTREE (H, H_root)
        EXTINCT (R_root, RANGE (A_root))
        TRAVERSE (A_root)
end
```

RECONCILE calls the recursive procedure TRAVERSE which constructs the reconciled tree. The procedure traverses the associate tree in preorder (i.e., each node is visited prior to any of its descendants). If the node being visited requires a duplication, then the subtree in the host tree that is rooted at the corresponding node is copied and added to the reconciled tree.

```
procedure TRAVERSE (a):
begin
        if a ≠ nil then begin
                if duplication[a] then begin
                        /* Get corresponding cluster in host tree */
                        Y ← lub_H(RANGE(a))
                        /* Locate start of search in reconciled tree */
                        if EMPTYSTACK (SR) then
                                s ← a
                        else begin
```

```
                    s ← TOP (SR)
                    if ISLEFTDESC (A, a, s) then
                        s ← LEFT (T, s)
                    else    s ← RIGHT (T, s)
                    end
                \* Find location in reconciled tree where subtree will be added,
                copy the subtree, then add to the reconciled tree */
                q ← FINDNODEABOVE (R, s, Y)
                r ← ADDBELOW (R, COPYOFSUBTREE (R, q), q);
                /* Mark extant associates */
                EXTINCT (LEFT(r), RANGE (LEFT(a)))
                EXTINCT (RIGHT(r), RANGE (RIGHT(a)))
                /* Put this duplication on the stack */
                PUSH (SA, a)
                PUSH (SR, r);
                end
            /* Visit the rest of the tree */
            TRAVERSE (LEFT(a))
            TRAVERSE (RIGHT(a))
            if duplication [a] then begin
                \* We've looked at everything above this duplication, so pop it
                from the stack */
                POP (SA)
                POP(SR)
                end
            end
        end
```

Both RECONCILE and TRAVERSE call the recursive procedure EXTINCT (n, Y) which marks the leaves of the subtree of the reconciled tree rooted at n as either "alive" or "extinct," depending on whether the cluster of each leaf is an element of the set of hosts Y (i.e., whether the leaf has an extant associate). If a node is alive then that node is assigned the label L_{node} of the associate, otherwise $L_{node} \leftarrow 0$. If either of an internal node's immediate left and right descendants are alive then that node is also alive.

```
procedure EXTINCT (n, Y)
begin
    if n ≠ nil then begin
        if n is a leaf  then begin
            /* Is associate present? */
            i = X_n ∩ Y
            if i ≠ ∅ then begin
                n is alive
                L_n ← i
                end
            else begin
                n is extinct
                L_n ← 0
                end
            end
        /* Visit rest of tree */
        EXTINCT (LEFT(n), Y)
        EXTINCT (RIGHT(n), Y)
        if n is not a leaf then
            /* If either left or right descendant (or both) has an associate
            then node is "alive" */
```

```
            if LEFT(n) is alive or RIGHT(n) is alive then
                n is alive
            else n is extinct
        end
end
```

APPENDIX 2

Previous Algorithms for Finding Area Cladograms

The approach described in this paper for finding area cladograms differs from the one I proposed earlier (Page, 1988) and subsequently implemented in version 1.5 of the program COMPONENT (Page, 1989b). The goal of the methods described in Page (1988) was to find the set of trees that conformed to one or more constraints, such as constraining areas sharing the same widespread taxon to be either "monophyletic" or "paraphyletic" on any area cladogram. This approach suffers from the limitation that it simply follows an algorithm rather than optimizes an optimality criterion (cf. Swofford and Olsen, 1990; Penny et al., 1992), hence it does not provide any optimality statistics for a given host tree. In addition the algorithm for Assumption 2 (Page, 1988:269-270) suffered from the effects of the combinatorial explosion in the number of possible starting configurations of widespread taxa and redundant distributions, as well as placements for areas that are part of the range of widespread taxa. The lack of an optimality criterion also made it impossible to find the host tree(s) that was optimal for two or more associate trees, unless the set of inferred host trees for each associate contain trees in common (as is the case for Rosen's [1978] data [see Page, 1989a]).

The method presented in this paper (which is a refinement of that presented in Page [1990a] and implemented in COMPONENT 1.5 as the FIT command) is an optimality method and hence does not suffer from these limitations. For any given host tree the fit between that tree and a set of associate trees is the sum of the fit of each individual associate tree onto the host tree. Finding the globally optimal host tree requires finding the tree (or trees) that minimize this sum of fits.

Software

The algorithms described in this paper are implemented in version 2.0 of the computer program COMPONENT which runs under Microsoft Windows version 3.0 and later. COMPONENT also features a wide range of tree comparison, consensus, and randomization methods. For details on obtaining this software please contact the author.

Table 1. Binary coding of the parasite cladogram in Fig. 8.

	Codes								
Hosts	1	2	3	4	5	6	7	8	9
A	1	0	0	0	1	1	1	1	1
B	0	1	0	0	0	0	0	1	1
C	0	0	1	0	0	0	1	1	1
D	0	0	0	1	0	1	1	1	1

Figure captions

Figure 1. A simple example of a map between two congruent host and associate trees (a). Hosts a-d harbor associates 1-4. For each node in the associate tree the map (b) specifies the corresponding node in the host tree. The map can be represented visually by superimposing the associate tree on the host tree (c).

Figure 2. An example of incongruent host and parasite trees (a) and a map between them (b). This map can be represented pictorially by superimposing the parasite tree on the host tree (c) or as a reconciled tree (d).

Figure 3. Felsenstein's (1979) Dollo parsimony paradox. Given this character state tree (a) and a most parsimonious reconstruction of that character's evolution on tree (b), two internal nodes of the tree require both states 2 and 3 to be assigned, as assigning just state 2 or just state 3 to either node would require multiple origins of the one of the character states, violating the Dollo criterion.

Figure 4. A n-tree on the set {a, b, c, d} and a m-tree on the multiset {a, b, b, c, c, d, d}.

Figure 5. Two views of the reconciled tree for the host and parasite trees shown in Fig. 2a showing that the parasite tree is a subtree of the reconciled tree, and that all the clusters in the reconciled tree are clusters of the host tree. In (a) each solid black node in the tree is labeled with the corresponding node in the parasite tree, and the hollow nodes represent parasites that are either extinct of uncollected. The thick branches trace out the original parasite tree. In (b) the nodes are labeled instead with the corresponding nodes in the host tree, and the clusters are indicated. Note that the reconciled tree consists of the host tree with an extra copy of the subtree (b,(c,d))

grafted below node f. The node labeled f* corresponds to the duplication event that gave rise to the two lineages of parasites.

Figure 6. A cladogram for 22 lactate dehydrogenase (LDH) sequences of bacteria, plant, and vertebrates, a cladogram for those organisms, and the corresponding reconciled tree. The three gene duplications required are numbered 1-3. By following the solid branches in the reconciled tree the reader can trace out the gene tree. Because of the paucity of LDH sequences the hollow branches on the reconciled tree reflect lack of sampling rather than lack of LDH. (Gene tree after Quattro et al., 1993: fig. 2.)

Figure 7. (a) Phylogeny for four taxa 1-4 occurring in four areas A-D coded for Brooks Parsimony Analysis (BPA), and (b) an area cladogram for those four areas with the ranges of each taxon mapped using BPA. Solid bars indicate distributions showing no homoplasy, the hollow bar indicates the extinction of taxon 1 in area D (after Brooks, 1990: figs. 12 and 13; see text).

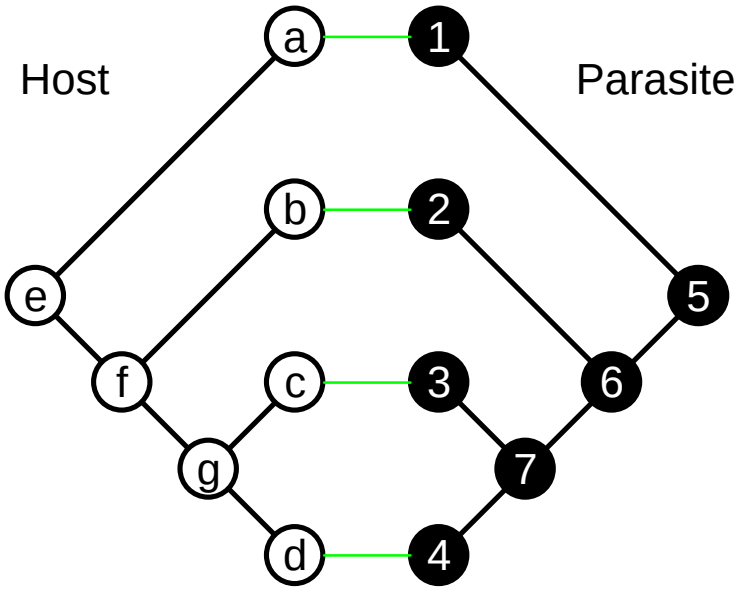
Figure 8. (a) Cladograms for five parasites (1-5) and their four hosts (A-D) coded for Brooks Parsimony Analysis (see also Table 1) and (b) the interpretation of the host-parasite association given the host cladogram. Solid bars represent single origins for the codes in Table 1, shaded bars represent parallel origins. Note that the occurrence of parasite 5 on host A can be explained by a single dispersal (arrow), yet requires two cases of homoplasy (codes 6 and 7).

Figure 9. Two reconciled trees for pocket gophers and their parasitic chewing lice after removal of two lice taxa postulated to have dispersed. In tree (a) Geomydoecus actuosi and G. thomomyus have been removed, in tree (b) G. ewingi and G. setzeri have been removed. Both these reconciled trees have six codivergent (= cospeciation) events, which is the maximum number possible for the gopher and lice

trees, and is one more than when no dispersal of the lice is allowed. Duplications are indicated by (○), solid branches indicate presence of parasite, hollow branches indicate inferred absence of parasite, shaded branches represent hosts which lack any lice. Tree (a) requires one duplication and three losses, tree (b) requires one duplication and one loss (the lack of lice on G. bursarius and O. underwoodi is treated as missing data).

Figure 10. (a) The area cladogram that has the minimum number of leaves added (32) when reconciled with Rosen's (1978) cladograms for Heterandria and Xiphophorus together. (b) The strict consensus of the three optimal area cladograms (leaves added = 0) for the same two fishes after deleting areas that are part of the range of a widespread taxon in one clade but not the other from the range of those widespread taxa (Nelson and Platnick's [1981] assumption 2; see text).

Figure 11. (a) The single optimal area cladogram for Heterandria alone, and (b) the Adams consensus of the 15 optimal area cladograms for Xiphophorus alone(see text).

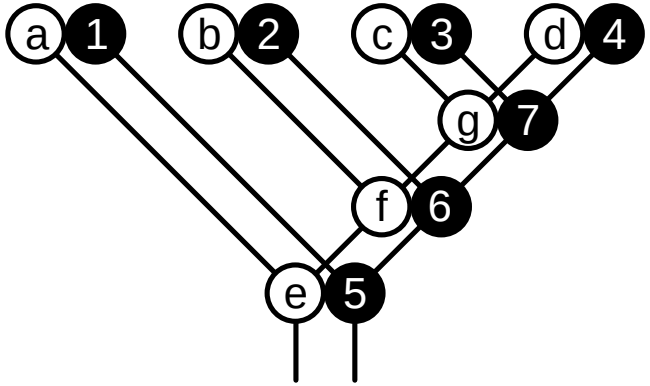


(a)

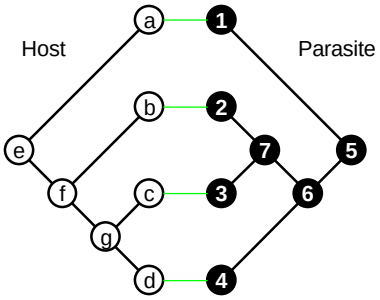
Map

1 → a
2 → b
3 → c
4 → d
5 → e
6 → f
7 → g

(b)

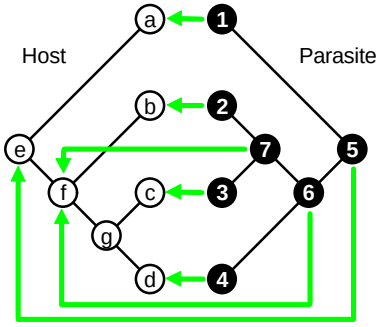


(c)



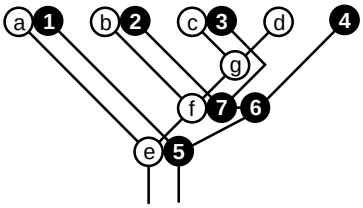
(a)

Map



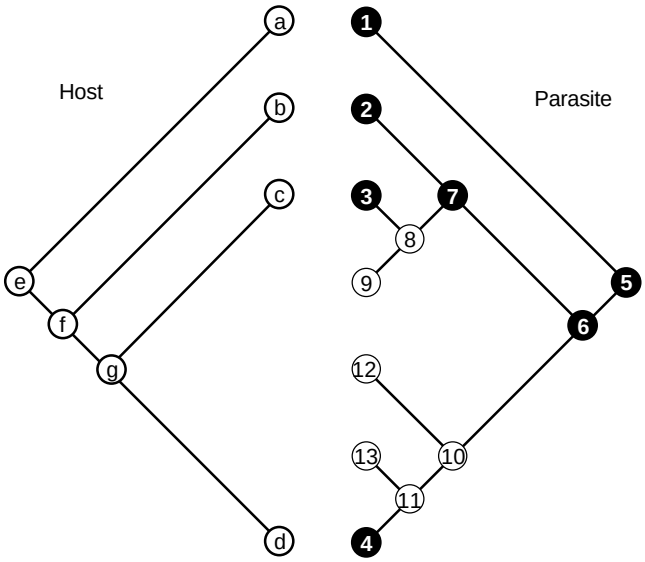
(b)

Superimposed trees

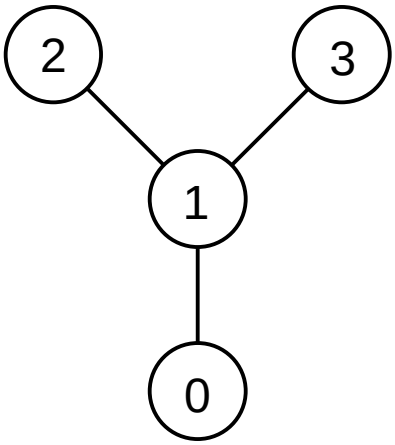


(c)

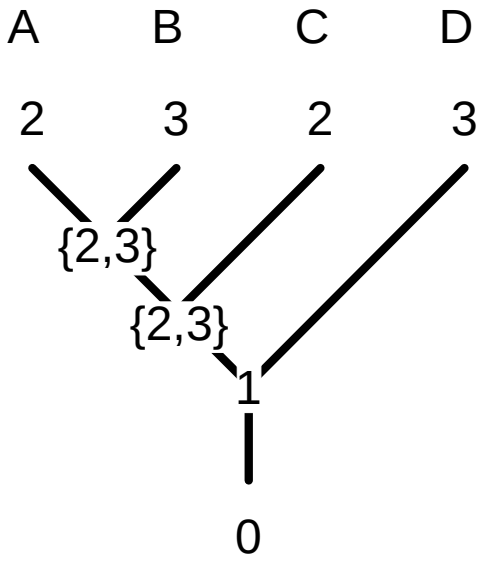
Reconciliation



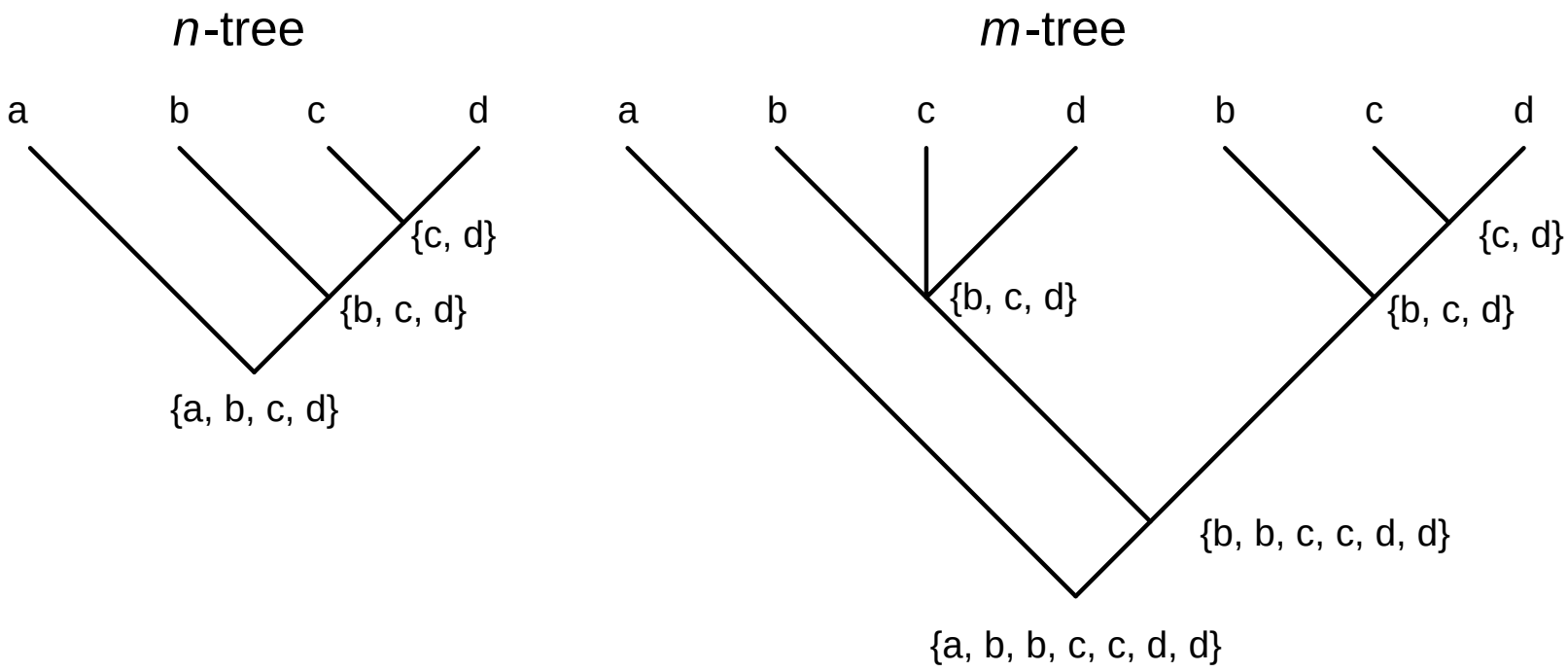
(d)



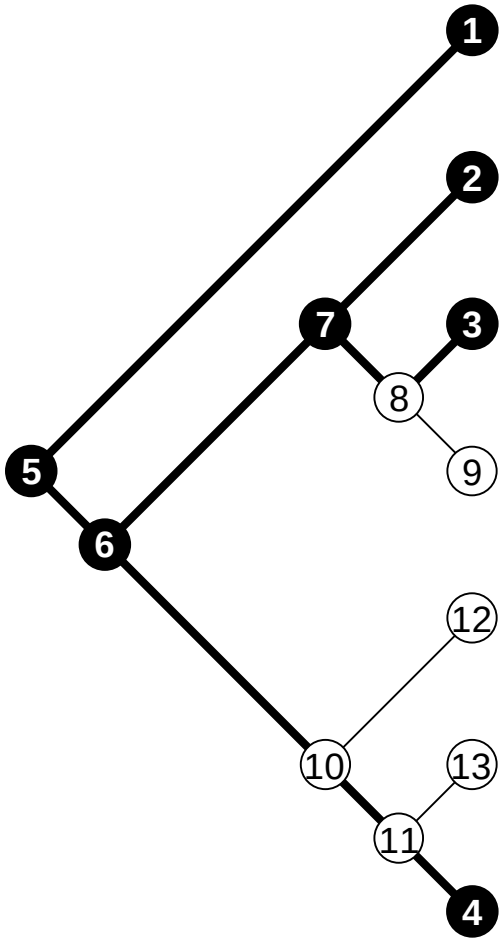
(a)



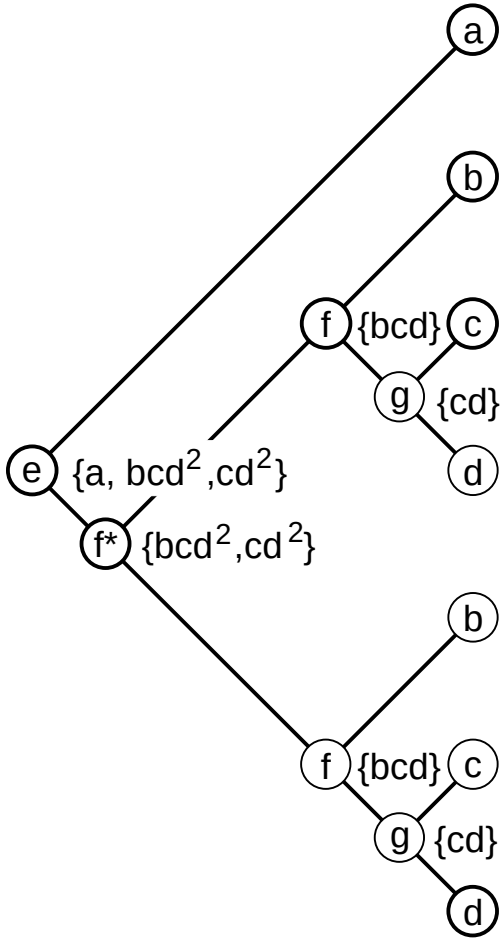
(b)


n-tree
a
b
c
d
{c, d}
{b, c, d}
{a, b, c, d}
m-tree
a
b
c
d
b
c
d
{c, d}
{b, c, d}
{b, c, d}
{b, b, c, c, d, d}
{a, b, b, c, c, d, d}

(a) (b)

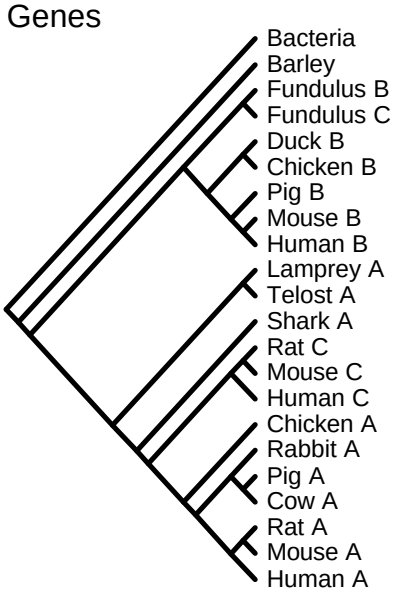

1
2
7
3
8
9
5
6
12
10
13
11
4

(a)

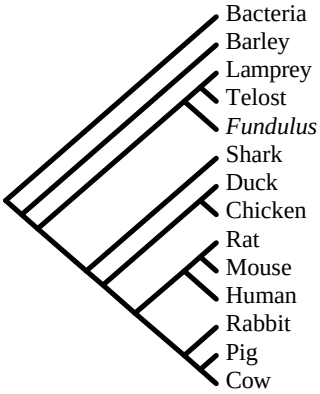

a
b
f {bcd} c
g {cd}
d
e {a, bcd2,cd2}
f* {bcd2,cd2}
b
f {bcd} c
g {cd}
d

(b)

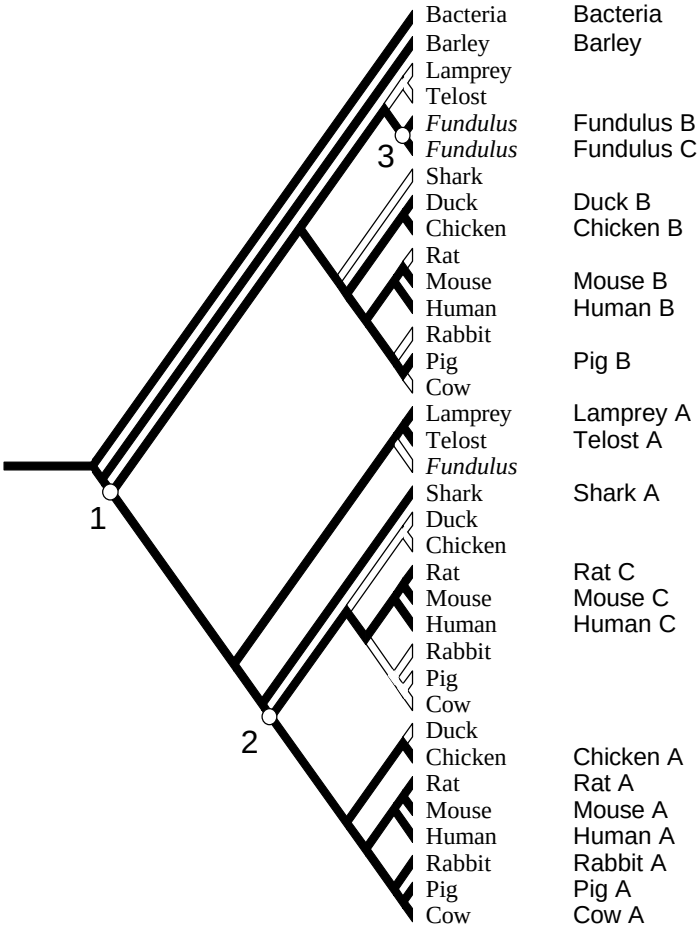
Genes

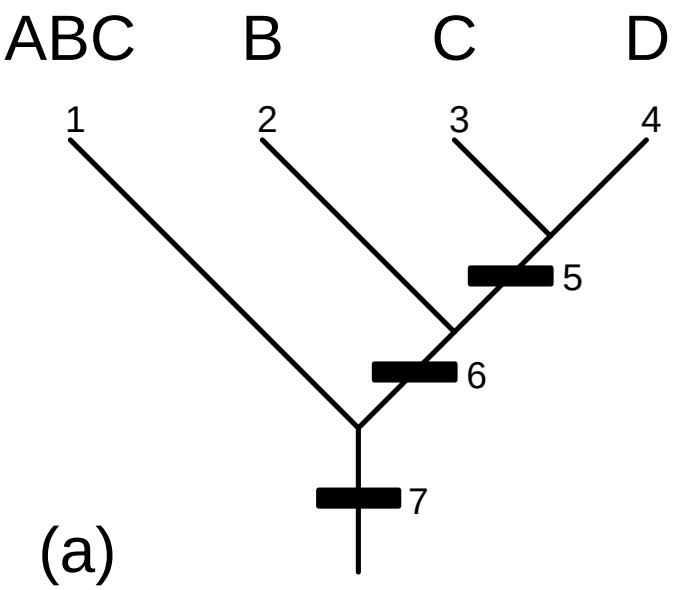


Organisms

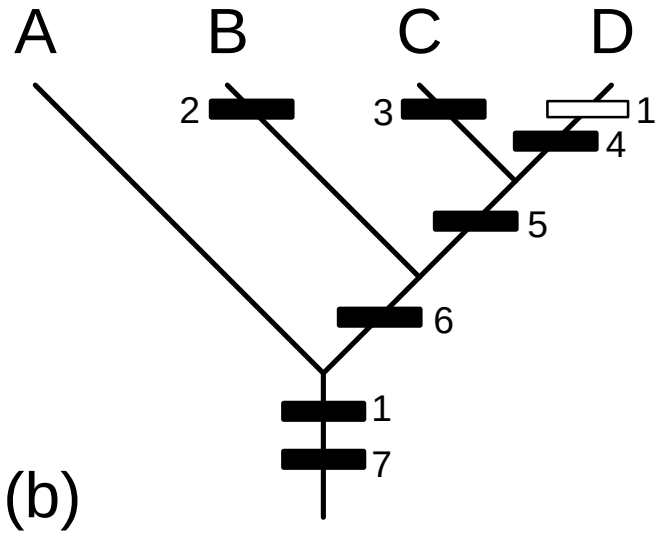


Reconciled tree

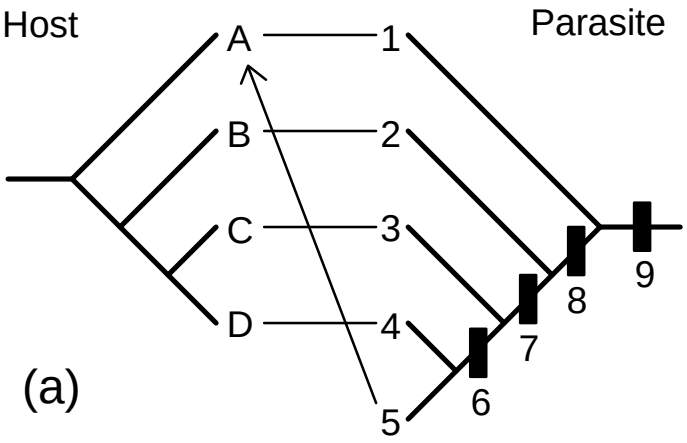


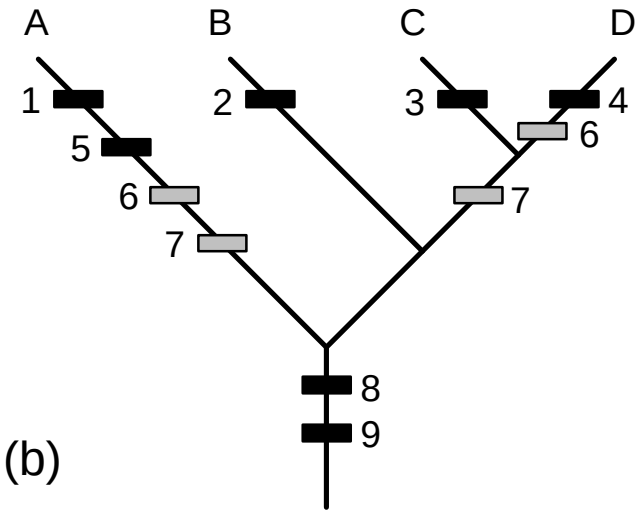


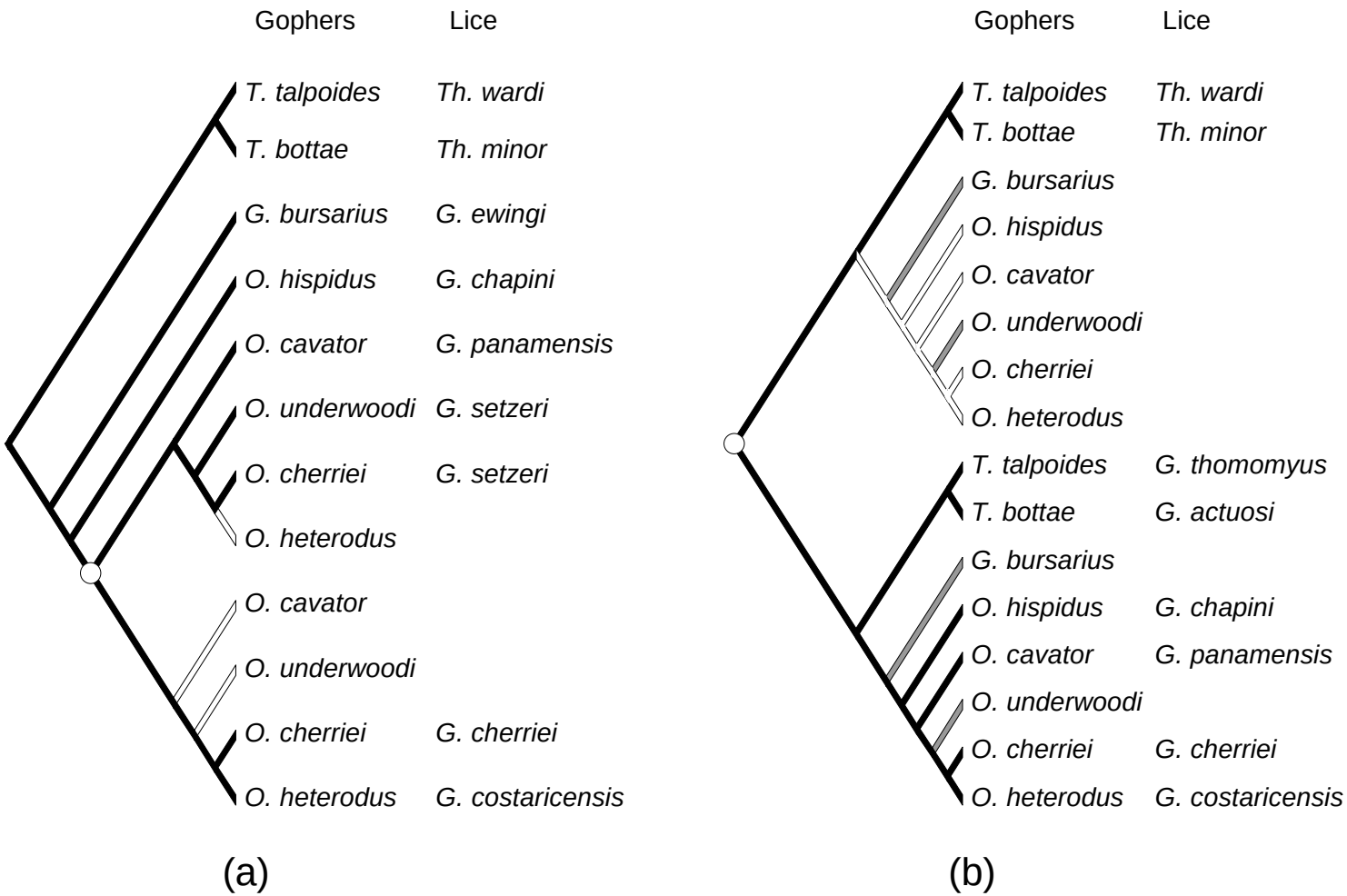
(a)

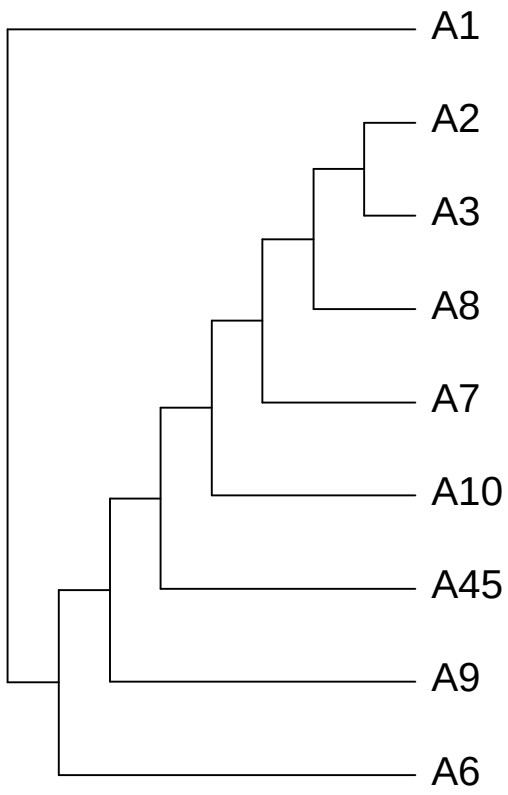


(b)

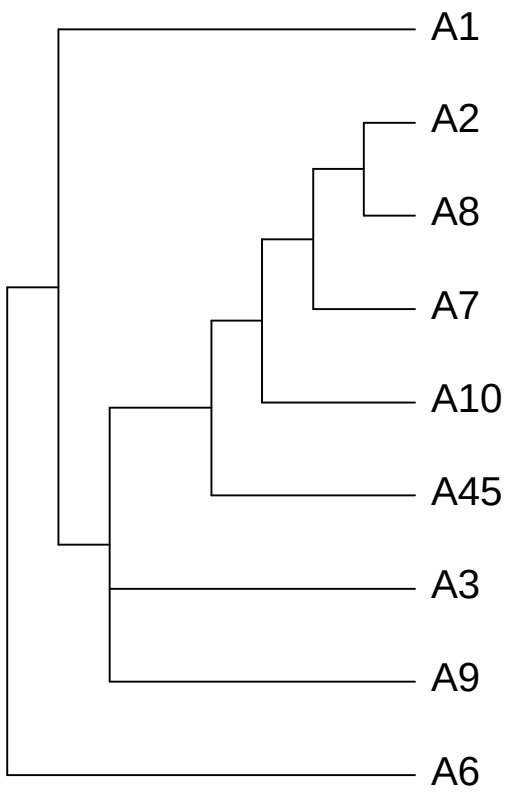

Host
Parasite
A
B
C
D
1
2
3
4
5
6
7
8
9
(a)


A
B
C
D
1
2
3
4
5
6
7
6
7
8
9
(b)

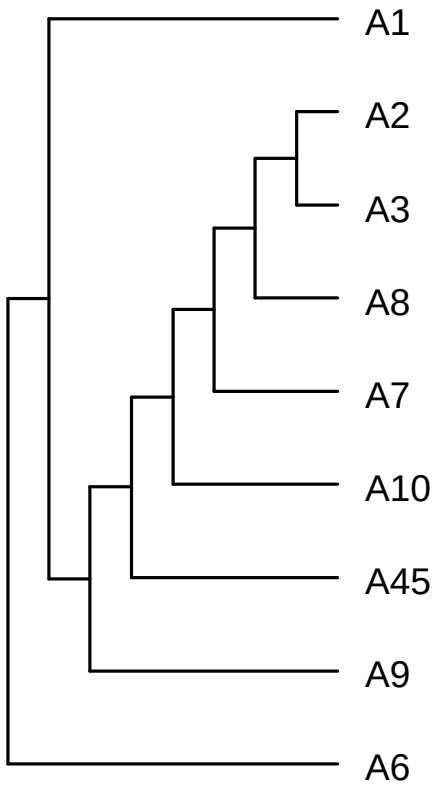

Gophers
Lice
T. talpoides
Th. wardi
T. bottae
Th. minor
G. bursarius
G. ewingi
O. hispidus
G. chapini
O. cavator
G. panamensis
O. underwoodi
G. setzeri
O. cherriei
G. setzeri
O. heterodus
O. cavator
O. underwoodi
O. cherriei
G. cherriei
O. heterodus
G. costaricensis
(a)
Gophers
Lice
T. talpoides
Th. wardi
T. bottae
Th. minor
G. bursarius
O. hispidus
O. cavator
O. underwoodi
O. cherriei
O. heterodus
T. talpoides
G. thomomyus
T. bottae
G. actuosi
G. bursarius
O. hispidus
G. chapini
O. cavator
G. panamensis
O. underwoodi
O. cherriei
G. cherriei
O. heterodus
G. costaricensis
(b)



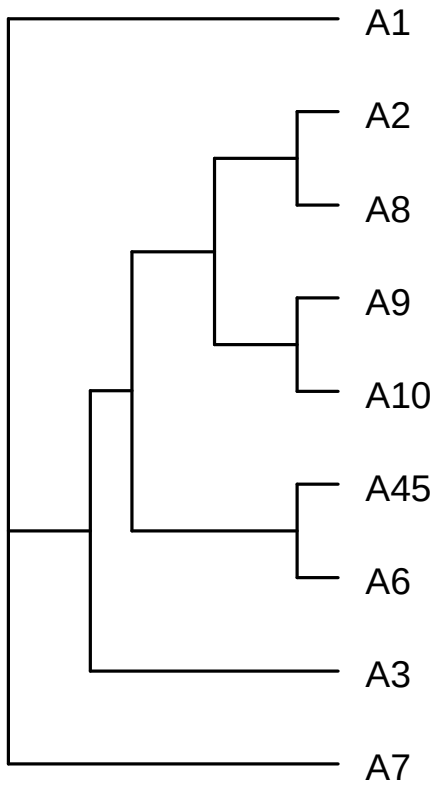
(a)



(b)



(a)



(b)