Chapter 7
Maps between trees

This chapter describes tree mapping routines in COMPONENT. These routines allow you to map gene trees onto species trees, parasite phylogenies onto host phylogenies, and taxon cladograms onto area cladograms.

Overview

This section introduces some basics of tree mapping. A full discussion will be presented elsewhere (Page, submitted; Page [1990a] describes an earlier implementation of tree mapping).

History

Goodman et al. (1979) and Nelson and Platnick (1981) independently developed a method for mapping one tree onto another that explains any incongruence between the two trees by invoking the presence of unrecognised multiple lineages in one of the trees. Goodman et al. developed the method to reconcile trees for mammals obtained from protein sequence data with trees derived from morphological data. They suggested that the incongruence could be due to some of their protein sequences being paralogous rather than orthologous (Fitch, 1970), hence their protein trees were confounding the history of genes with the history of organisms.

Nelson and Platnick (1981: 410-467) suggested ways the effects of poor taxonomic sampling and extinction, among other factors, could lead to incongruence between area cladograms for different taxa. They suggested that in the presence of two or more sympatric lineages of taxa, poor sampling could obscure the underlying area relationships for the same reason that sampling from a collection of paralogous genes may give a confused picture of species relationships (see Page, 1993a). This same idea of multiple lineages was invoked by Humphries et al. (1986) in their study of cospeciation between *Nothofagus* and its parasites.

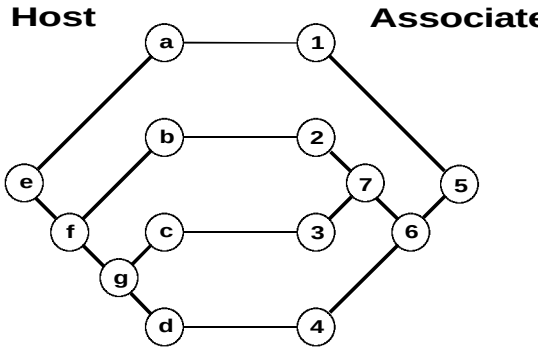
Basic concepts

Tree mapping is designed to help analyse historical associations between "hosts" and "associates." Examples of such associations are:

Hosts	Associates
organisms	genes
host organisms	parasitic organisms
areas	organisms

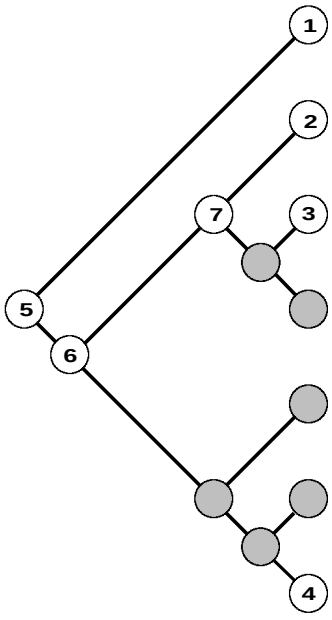
The key concept in tree mapping is reconciling a tree for the associate with a tree for the hosts, under the assumption that the host-associate relationship is due to "association by descent" (Mitter and Brooks, 1983; Brooks and McLennan, 1991). This procedure provides a measure of fit between host and associate trees that can be tested statistically, and also generates hypotheses about the relative ages of divergence events in the two lineages.

Figure 7.1 Incongruent host and associate trees



Given incongruent host and associate trees (e.g., Figure 7.1) we can reconcile the two trees so that the observed associate relationships and host associations can be explained solely by "association by descent." To accomplish this we would have to postulate that the observed associate cladogram is a subtree of a larger tree:

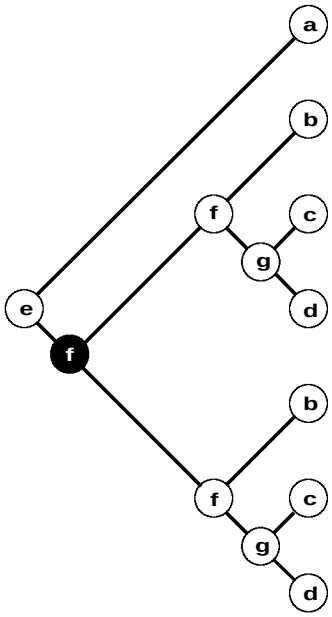
Figure 7.2 The reconciled tree for the host and associate trees in Figure 7.1



This larger tree is called a *reconciled* tree. You can think of it as representing the complete cladogram for the associate, of which the actual associate tree is considered to be a subsample or relict. It might be that our sample of associates is poor, or that several have gone extinct. For example, if the associates were DNA sequences then we could interpret the reconciled tree to mean that sequences 2 and 3 are paralogous with respect to sequence 4, and both sets of sequences are the descendants of a gene duplication represented by node 6 above.

Although it is probably easier to interpret reconciled trees by considering just the associates, it is important to realise that all the nodes in the reconciled tree are all clusters of the host tree. Figure 7.3 shows the same reconciled tree as above but with the nodes labelled with the corresponding node in the host tree.

Figure 7.3 The same reconciled tree shown in Figure 7.2 but with the nodes labelled with the corresponding node in the host tree



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 that the two subtrees are rooted at (●) corresponds to a *duplication* event that gave rise to the two lineages of associates. COMPONENT distinguishes between two kinds of duplication: those that are required because the host and associate trees are incongruent (as in the above example) and those that are required because the descendants of a given associate have overlapping host ranges. A trivial example of the latter would be if two sister parasite taxa had the same host.

Maps between trees

Reconciling two trees involves making a map between the associate and host tree (Page, 1990a). The map requires that each node in the associate tree is assigned a distribution, *D*. For the terminal nodes this is the observed distribution. For internal nodes it is the union of the distribution of all the descendants of that node. The map is then 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.

More formally, given an *n*-tree T (see Chapter 0) on the set N = {1, ..., *n* }, the least upper bound in T of any subset M of N is the smallest cluster $X \in T$ such that $M \subseteq X$, and is denoted lub_T (M). The map $f(T_A, T_H)$ between the associate tree T_A and the host tree T_H is given by the function $lub_{T_H}(D_i)$ for all $D_i \in T_A$, where D_i is the distribution of the ith node in the associate tree. For the example above, node 7 in the associate tree has the distribution {2, 3}. The lub in the host tree of this set is {2, 3, 4}, so that node 7 in the associate tree maps onto node f in the host tree. The complete map between the two trees is:

Table 7.1 Map between the two trees in Figure 7.1

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

Pictorially this map can be expressed as:

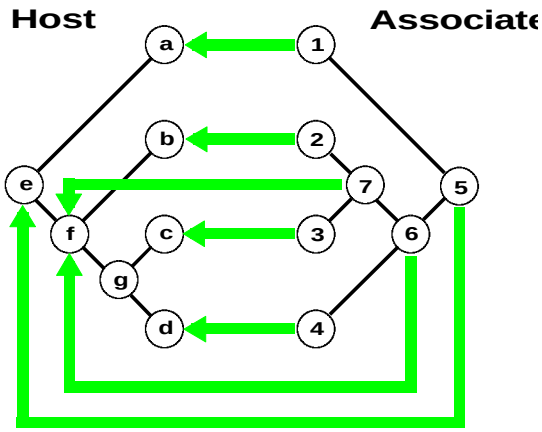


Figure 7.4 The map between the host and associate trees in Figure 7.1 (see also Table 7.1)

If the map is one-to-one, that is each node in the associate tree maps onto a different node in the host tree, then the associate tree is either identical with the host tree, or a consistent subtree of the host tree. However, if more than one associate node maps onto the same node in the host tree then we have a *duplication* (Goodman, et al., 1979). In the map shown in Figure 7.4 above both nodes 6 and 7 map onto node f so we have a duplication at node f.

Measures of fit

COMPONENT can compute three measures of fit between host and associate trees:

- number of duplications
- number of leaves added
- number of independent losses

Duplications

The number of duplications is simply the number of times a duplication of a lineage has to be postulated to reconcile the host and associate trees.

Leaves added

Nelson and Platnick (1981:417) measured the degree of fit between two trees as the difference in the number of nodes in the associate and reconciled tree, which they termed the "items of error." In the example given in Figures 7.1-7.4 above, the reconciled tree contains 13 nodes, the associate tree has 7, so there are six items of error. COMPONENT computes the number of leaves added as half the items of error.

Losses

Simply counting the number of leaves added to reconcile two trees may overestimate the number of actual events involved (Page, 1988: 260). For example, a single loss may account for the absence of associates from a large clade of taxa. Goodman et al. (1979) used the number of gene losses as their measure of fit between gene and species trees.

Note that in the example in Figures 7.1-7.4, the number of losses equals the number of leaves added.

Input format

Unlike all other COMPONENT routines, tree mapping requires the input file to have at least one DISTRIBUTION block (see Chapter 3). This block is used to specify the distribution and relationships of the associate. The file must also contain a TAXA block listing the names of the associates. The host tree(s), if known, are placed in the TREES block.

Viewing trees in different blocks

A COMPONENT Tree window displays the trees from just one block at a time. To view another block within the same file use the drop down list box to select the desired block:

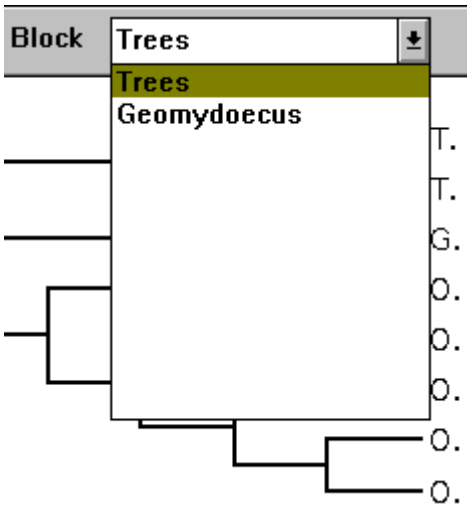


Figure 7.5
The block drop down list box, in this example listing the TREES block and a DISTRIBUTION labelled "Geomydoecus"

The TREES block is always listed as "Trees." If your DISTRIBUTION block(s) contain a TITLE command then the list box will display that title, otherwise it will display "Block *n*" where *n* is the order of the block in the input file.

Example of a taxon-area input file

This example illustrates Rosen's (1979) cladogram for the fish genus *Heterandria* which occurs in Central America. There is no area cladogram, hence the file lacks a TREES block.

Figure 7.6 Example data file specifying the distribution and relationships of a single clade found in nine areas

```
#NEXUS

[! Rosen's data for Heterandria ]

BEGIN TAXA;
[ The nine areas of endemism harbouring Heterandria ]
        DIMENSIONS NTAX=9;
        TAXLABELS
                A1 A2 A3 A45 A6 A7 A8 A9 A10;
ENDBLOCK;

BEGIN DISTRIBUTION;
[ The distribution and relationships of Heterandria ]
        TITLE = 'Heterandria';
        NTAX=8;
                        [ fish ]      [ area ]
        RANGE
                        attenuata      : A6,
                        jonesi         : A1,
                        litoperas      : A9,
                        obliqua        : A45,
                        anzuetoi       : A10,
                        cataractae     : A7,
                        dirempta       : A8,
                        bimaculata     : A2 A3
                        ;
        [ The taxon cladogram for Heterandria ]
        TREE Rosen=(attenuata,(jonesi,(litoperas,
                (obliqua,(anzuetoi,(cataractae,
                (dirempta,bimaculata)))))));
ENDBLOCK;
```

Example of host-parasite input file

This example file contains Hafner and Nadler's (1988) trees for pocket gophers and their parasitic chewing lice. The TAXA block names the eight gopher hosts, the DISTRIBUTION block lists the ten parasitic lice, their distribution on the gophers, and their interrelationships, and the TREES block specifies the host cladogram.

Figure 7.7 Example data file for 10 taxa parasitic on eight host species where both parasite and host relationships are known

```
#NEXUS

[!
Hafner and Nadler's pocket gopher - chewing lice data.
]

BEGIN TAXA;
[ The eight pocket gopher hosts belonging to Thomomys,
  Geomys, and Orthogeomys ]
       DIMENSIONS NTAX=8;
       TAXLABELS
              T._talpoides
              T._bottae
              G._bursarius
              O._hispidus
              O._cavator
              O._underwoodi
              O._cherriei
              O._heterodus;
ENDBLOCK;

BEGIN DISTRIBUTION;
[ The 10 Thomomydoecus and Geomydoecus lice parasitic on
  the eight pocket gophers ]
       TITLE  = 'Geomydoecus';
       NTAX=10;
       RANGE
              [ Lice ]              [ Gophers ]
              Th._wardi       : T._talpoides,
              Th._minor       : T._bottae,
              G._thomomyus    : T._talpoides,
              G._actuosi      : T._bottae,
              G._ewingi       : G._bursarius,
              G._chapini      : O._hispidus,
              G._panamensis   : O._cavator,
              [This next louse occurs on two gophers]
              G._setzeri      : O._underwoodi O._cherriei,
              G._cherriei     : O._cherriei,
              G._costaricensis: O._heterodus
              ;
       [ UPGMA dendrogram for lice ]
       TREE lice = ((1,2),((3,(4,5)),(6,((7,8),(9,10)))));
ENDBLOCK;

BEGIN TREES;
       [! UPGMA dendrogram for gophers ]
       TREE gophers = ((1,2),(3,(4,(5,(6,(7,8))))));
ENDBLOCK;
```

Computing reconciled trees

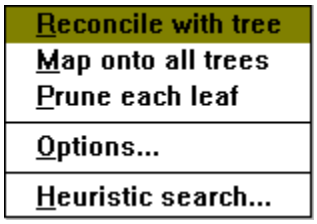
This section describes how you can use COMPONENT to compute a reconciled tree and how to interpret the output. Computing a reconciled tree requires that your input file has at least one DISTRIBUTION block and a TREES block. Hence, if you are doing a biogeographic analysis you must have one or more area cladograms, for a cospeciation analysis you need one or more host cladograms, and for a gene tree/species tree analysis you need one or more species trees. If you do not have any "host" trees you can ask COMPONENT to compute them (see the section "Searching for optimal host trees" below).

Reconciling two trees

- Ensure that the Tree window is displaying the associate tree (e.g., the tree for the parasites if you are reconciling host and parasite cladograms). This tree is in the DISTRIBUTION block.

- Choose the **Map trees** command from the **Trees** menu. A sub menu will be displayed:

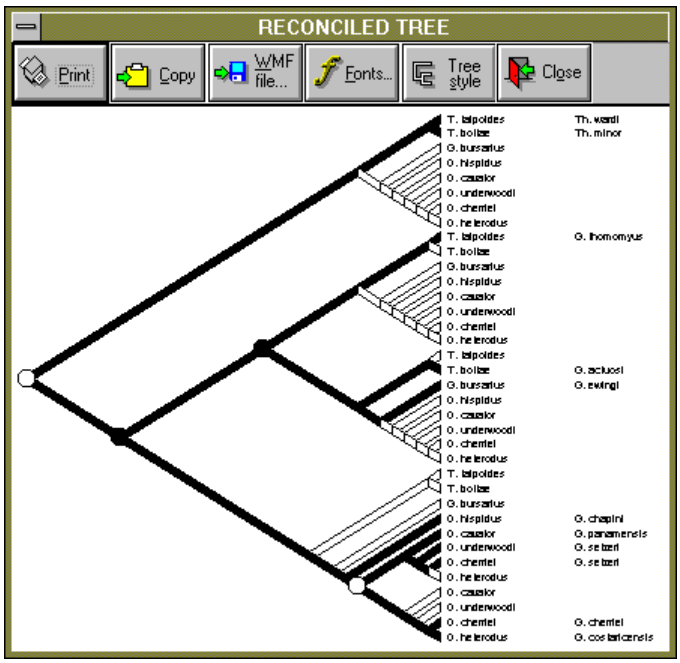
Figure 7.8
*The **Map trees** sub menu*



- Choose the **Reconcile with tree** command.

COMPONENT will compute the reconciled tree for the currently displayed associate tree and current tree in the TREES block (this is the last tree displayed when you were viewing the TREES block) and display it in a dialog box:

Figure 7.9
The Reconciled tree dialog box

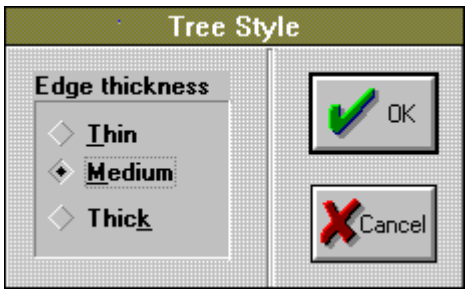


Printing the reconciled tree

You can print the reconciled tree displayed in the Reconciled Tree dialog box by clicking on the **Print** button.

COMPONENT lets you alter the thickness of the lines used to draw the tree and the font used for the labels. To alter the line thickness click on the **Tree style** button. The following dialog box appears:

Figure 7.10 The Tree Style dialog box



To change the font click on the **Font** button. The following dialog appears listing the fonts supported by your printer.

☞ *At present COMPONENT's ability to print reconciled trees is not particularly sophisticated and is* not *WYSIWYG; selecting new line thicknesses and fonts will not alter the appearance of the tree in the dialog box, but it will change how the tree looks when printed, copied to the clipboard, or saved as a graphics file. You may want to use a graphics program to smarten up the picture (see below).*

Copying a picture to the Windows Clipboard

You can copy a *metafile* picture of the reconciled tree to the Windows Clipboard by clicking on the **Copy** button. This picture can be pasted into other applications, such as a word processor.

Saving a picture to a graphics file

You can save a picture of the reconciled tree to a Windows Metafile (WMF) format file. This format is recognised by most, if not all Windows graphics and word processing programs.

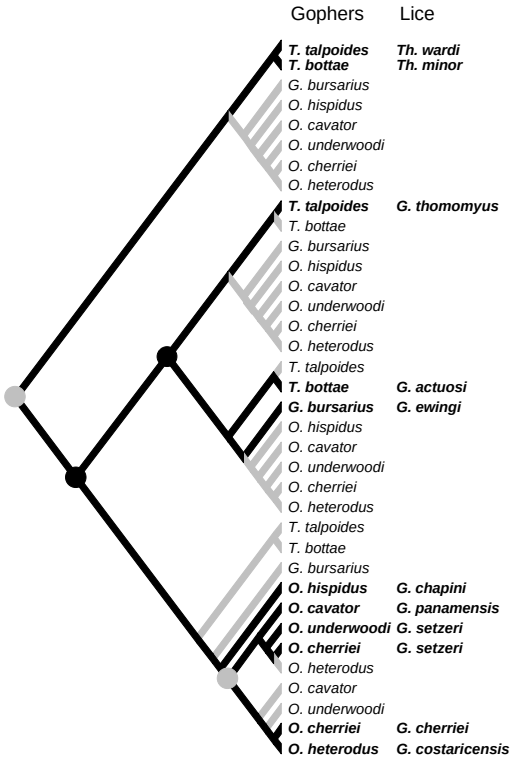


Figure 7.11 The reconciled tree shown in Figure 7.9 after the picture has been saved as a WMF file then edited with a graphics program.

Output

The reconciled tree is also shown in the display buffer, together with a key to the symbols used to draw the tree:

The numbers assigned to the internal nodes of the reconciled tree correspond to the internal nodes of the host tree. The display buffer also contains the measures of fit between the two trees, and the map:

*Figure 7.12 Example output from the **Reconcile trees** command showing tree statistics and map*

```
Reconciled tree statistics
   Duplications = 1
   Total leaves = 7
   Leaves added = 3
   Losses = 3

Map between associate and host tree

   Associate                  Host                  Duplication?

    5                  ->       5                   NO
   associate 1         ->      host a               -
    6                  ->       6                   YES
    7                  ->       6                   NO
   associate 2         ->      host b               -
   associate 3         ->      host c               -
   associate 4         ->      host d               -
```

Reconciling with more than one host tree

The **Reconcile with tree** command only lets you reconcile two trees. There may be times when you might wish to map the associate tree onto many host trees, for example, to compute the distribution of the fit statistics between the associate tree and random host trees (Page, 1990a, 1990b). This distribution allows you to assess whether the fit between host and associate trees is greater than you would expect due to chance alone.

COMPONENT will map the associate tree currently displayed in the Tree window onto all active trees in the TREES block. Once completed the program displays histograms of both the number leaves added and number of losses that each host tree required.

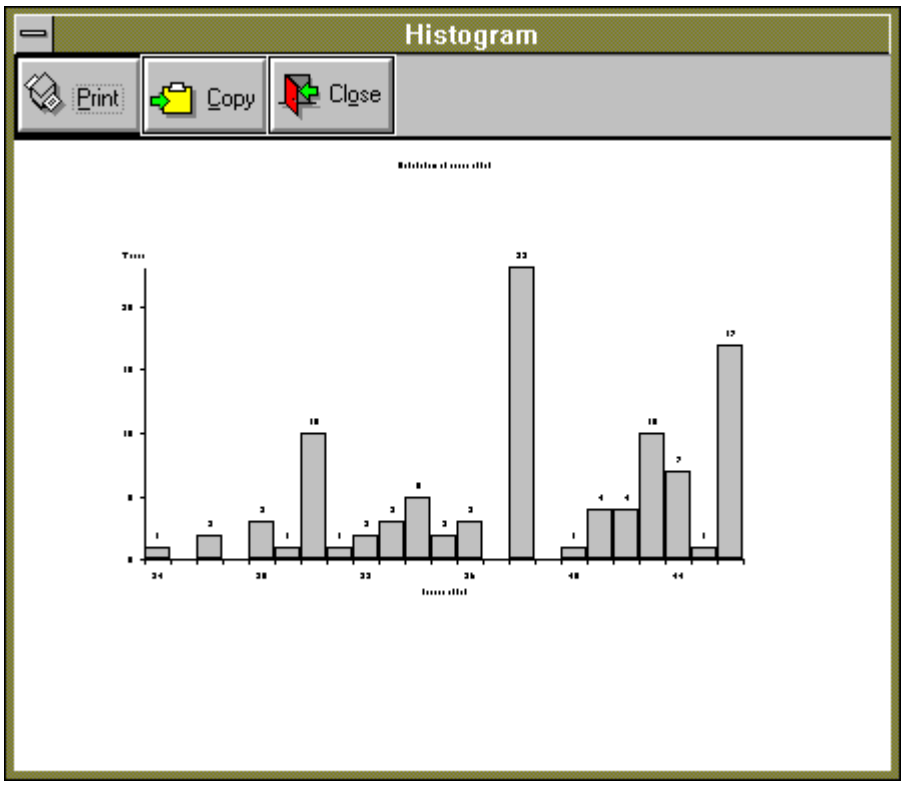


Figure 7.13
A Histogram dialog box showing the distribution of number of leaves added for 100 random host trees

The actual values themselves are written in the display buffer.

Reconciling pruned associate trees

If an associate has dispersed from one host to another then it may cause a poor fit between the two trees, causing the amount of associate extinction and/or collection failure to be overestimated (see the worked example in Chapter 8, and Page, 1990b; submitted). You can use the **Trees | Prune or graft leaves** command to delete associates that might have dispersed. If removing an associate greatly improves the fit between the two trees then that can be considered evidence for dispersal.

You can automate this procedure by using the **Prune each leaf** command. COMPONENT will remove each leaf from the associate tree, reconcile the pruned tree with the host tree, restore that leaf, then remove the next leaf and so on. The display buffer will list each leaf deleted and the reconciled tree statistics:

*Figure 7.14 Example output from the **Prune each leaf** command*

Statistics for reconciled tree after pruning each leaf

Leaf	Label	Dups.	Added	Losses	Leaves
1	Th._wardi	4	27	11	36
2	Th._minor	4	27	11	36
3	G._thomomyus	3	19	8	28
4	G._actuosi	3	19	8	28
5	**G._ewingi**	**2**	**9**	**4**	**18**
6	G._chapini	4	23	10	32
7	G._panamensis	4	22	9	31
8	G._setzeri	3	19	7	28
9	G._cherriei	4	27	11	36
10	G._costaricensis	4	22	9	31

In the example above we can see that by pruning the parasite "G._ewingi" we obtain the closest fit to the host tree. Hence one could postulate that this taxon has dispersed.

Searching for optimal host trees

In many cases the "host" tree may be unknown, indeed, the goal of the study might be to estimate the host tree. For example, a biogeographer might want to estimate the area cladogram for one or more taxa. COMPONENT implements a simple heuristic algorithm to search for the host tree(s) that minimises any of the three fit statistics described above.

- Choose the **Map tree** command from the **Trees** menu. A sub menu will appear: choose the **Heuristic search** command.

- COMPONENT will display a dialog box listing the various options available:

Figure 7.15 The Heuristic Search Options dialog box

The **Branch swapping options** group lists the two methods available for rearranging the initial tree in search of more parsimonious trees: nearest neighbor interchanges and subtree pruning and regrafting (see Swofford and Olsen, 1990). If

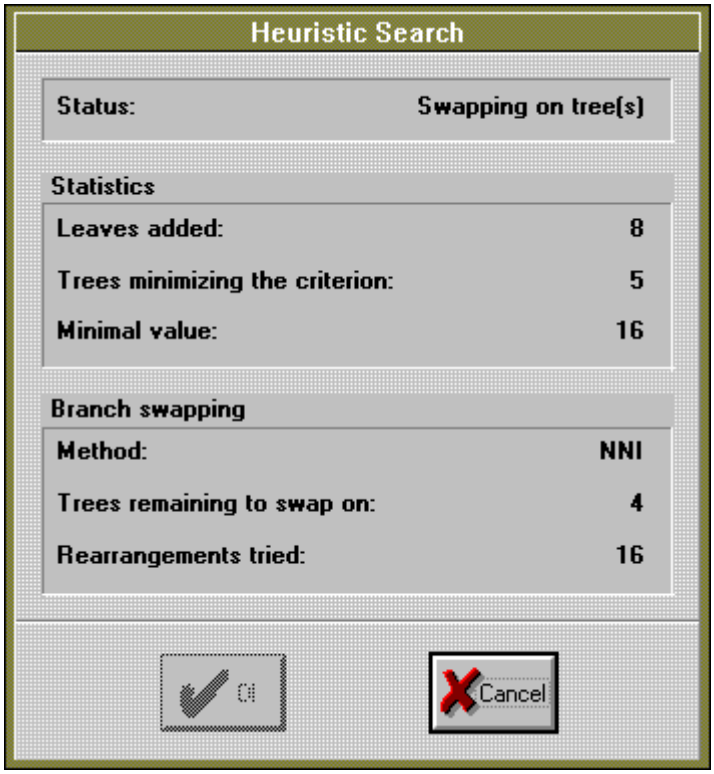
you choose **No swapping** then the program finds one tree in a single pass. This tree may often be a poor estimate of the minimal tree. If you choose the **Nearest neighbor interchanges** or **Subtree pruning-regrafting** options the program will rearrange the initial tree in search of a better solution. All equally optimal trees (up to the limit of 1000) will be retained by the program.

The **Criterion to be minimised** group list the three possible criteria that you can choose to minimise: duplications, leaves added, and losses.

If you have more than one DISTRIBUTION block in the input file you have the option of including all the blocks or just the current block. If you include all the blocks the program will search for the host tree that has the overall minimal value of the chosen fit statistic.

While COMPONENT is computing the trees it displays a status dialog box showing the progress of the search:

Figure 7.16
The Heuristic Search progress dialog box



If you interrupt the search (by clicking on **Cancel**) COMPONENT will ask you if you want to keep the trees found so far.

COMPONENT replaces any trees in the TREES block (if present) with the trees found during the search. You can then access and analyse those trees just as you would any trees in a TREES block.

Options

There are two options that affect how COMPONENT reconciles two trees.

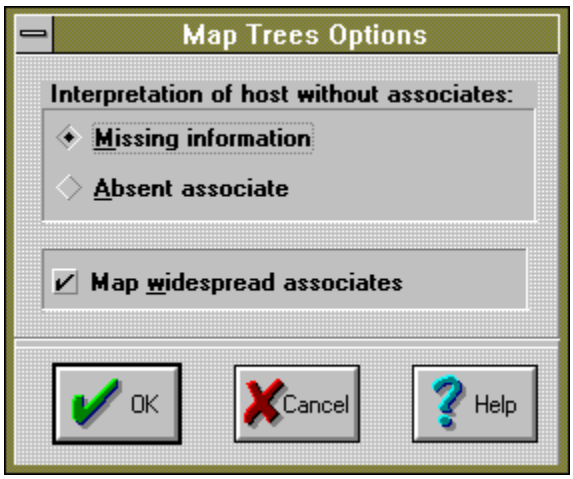
Hosts without associates

It may be that one or more of the "hosts" lacks associates. For example, you may be attempting to construct an area cladogram for 10 areas but some of the taxa are not known in all of the areas. In this context a host has no associate if within a given DISTRIBUTION block none of the associates listed in the TAXA block occurs on that host.

COMPONENT offers two interpretations of missing associates: either the absence is treated as a genuine absence and hence is counted as a loss, or the absence is treated as a "missing value" and is not included in the tree fit statistics. The latter is the default.

To alter the setting of this option choose the **Options** command from the **Map trees** sub menu. The following dialog box will appear enabling you to toggle between the two different interpretations of hosts with no associates:

Figure 7.17
The Map Trees Options dialog box



Mapping widespread associates

By default COMPONENT maps all associates onto the host tree(s). This is equivalent to the treatment of widespread taxa under "Assumption 0" (Zandee and Roos, 1987; Page, 1990a). As two or more areas may share the same widespread taxon due to geographical proximity rather than close relationship, widespread taxa can be misleading. COMPONENT offers you the option of not mapping widespread taxa. This is effectively "Assumption 1" (Nelson and Platnick, 1981; Page, 1990a). Nelson and Platnick's (1981) "Assumption 2" is not implemented directly in COMPONENT 2.0. However, you can treat widespread taxa in a similar manner to Assumption 2 by deleting all but one area from the range of each widespread taxon. For an example see Page (submitted) and Chapter 8.

The setting of this option can be altered using the **Options** command from the **Map trees** sub menu.

Algorithms

The algorithms used will be described in detail elsewhere (Page, submitted; Page [1990a] describes an earlier version of the algorithms). However, one limitation of the algorithms is described below.

Nonbinary trees

The algorithms for reconciling two trees are defined only for binary (i.e., fully resolved trees). If an associate or host tree is not binary then COMPONENT arbitrarily resolves any polytomies. If you are reconciling two trees with the **Reconcile two trees** command then the arbitrary nodes introduced by this procedure will be marked with the "&" symbol.

Since not all resolutions of a polytomy may be equally parsimonious explanations of the original character data I recommend that you input only binary trees (you can input multiple associate trees in the DISTRIBUTION block). If you have many equally parsimonious trees which fall into distinct clusters of relatively similar trees (for example, if the set of trees contains several islands; Maddison, 1991) it may be preferable to input representative trees from these different clusters, rather than input a single, poorly resolved consensus tree.